Hierarchical Methods for Large Population Speaker Identification using Telephone Speech

Lerato Lerato

Submitted to the faculty of Engineering and Built Environment, University of Cape Town, in fulfilment of the requirements for the Master of Science in Electrical Engineering.

Cape Town, December 2003

To Lebohang Lerato

Declaration

I declare that this dissertation is my own work. It is being submitted for the Degree of Master of Science in Electrical 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 of Author .....

Cape Town
15 December 2003

Acknowledgements

Many thanks to Dr. Dije Mashao for the most consistent supervision one can ever imagine. Marena a be le lelapa la hae ka ho sa feleng. Credit also goes to the Dept. of Electrical Engineering, CoE and NRF for the financial support. Motsoali le bo-Malome kea le leboha. Ho bo motsoala le bana beso ho joalo feela. **All my friends** who really supported me throughout difficult times deserve a BIG "thank you". Teboho "Morenaka", MaliaK "Baba", Lehana "Majase", Tsebang "Morantate", Ngaka Mohlabi, Lameka (*my main man*) *et al.* ha e 'ne e sise e tsoele le mohasula. 'Me' Cina kea leboha ka ts'ehetso e ke keng ea lekanngoa. 'Me' *Motšelisi "MC"* , ke mo leboha ho menahane ka ts'ehetso e matla e entsoeng ka pelo e ntle. Helen for being such a loyal and warm hearted Guardian. I thank UH boys (boholo ke Mafokothi) for all the moral support for "liMastas". I also thank Mahipi (linika) from Lesotho for relieving the stress that I encountered every now and then. The most important person and a dear colleague of mine, ntate Limpho "*Frequency Domain* $F(\omega)$ Makherenkhoa" Mothae, made my two years of stay in CRG the funniest in many ways, banna! Empa tsohle li phethoa ke Tlhahla-Macholo, Ralefaloe, Rammoloki, Atla-li-Maroba, Sekhele, Ramaseli, Ramohau, Mohokong, 'Mopi, Pilo, Molimo e leng eena moreri, moqeti le mophethahatsi oa lintho tsohle tse ntle lefats'eng. Ke mathetho. Ea khaola ea ea!

Abstract

Speaker recognition is the area of interest in this study. This area comprises speaker identification (SiD) and speaker verification. The SiD system identifies the person from his or her voice, mostly with assumption that the person belongs to a known group of speakers. Speaker verification is the same as SiD except that the speaker claims who he or she is before recognition process executes. This study focuses on speaker identification.

Several problems such as acoustic noise, channel noise, speaker variability, large population of known group of speakers within the system and many others limit good SiD performance. The SiD system extracts speaker specific features from digitised speech signal for accurate identification. These feature sets are clustered to form the speaker template known as a speaker model. As the number of speakers enrolling into the system gets larger, more models accumulate and the interspeaker confusion results. This study proposes the hierarchical methods which aim to split the large population of enrolled speakers into smaller groups of model databases for minimising interspeaker confusion. The group detector algorithm is used for this purpose. This method is called the group detection hierarchical method. The second procedure is done by connecting new feature extractor which exhibits uncorrelated errors with the baseline feature extractor. The new feature extractor complements the errors made by the baseline, using a decision threshold which triggers the execution of the complementary module. This procedure is called the N-best hierarchical method because the identification decision is made from the top N scoring speaker models from using the complementary module.

The results from the group detection method indicate the improvement of the SiD performance. The main problem with it is lack of robustness of group detectors if grouping sentences differ. The N-best hierarchical system however, has improved

the baseline SiD performance and has also reflected robustness as the population of speakers in the model databases vary.

Contents

List of Figures

List of Tables

List of Abbreviations

ANN	-	Artificial Neural Networks
ARVM	-	AutoRegression Vector Model
DFT	-	Discrete Fourier Transform
EIH	-	Ensemble Interval Histogram
FFT	-	Fast Fourier Transform
GMM	-	Gaussian Mixture Model
GSM	-	Global System for Mobile Communications
GSM codec	-	GSM **co**mpression/**dec**ompression
IDFT	-	Inverse Discrete Fourier Transform
IFFT	-	Inverse Fast Fourier Transform
LPCC	-	Linear Predictive Cepstral Coefficients
LPF	-	Low Pass Filter
MFCC	-	Mel-Frequency Cepstral Coefficients
PFS	-	Parameterised Feature Sets
PLP	-	Perceptual Linear Prediction
RASTA	-	RelAtive SpecTrA
SiD	-	Speaker iDentification
SVM	-	Support Vector Machine

Chapter 1

Introduction

1.1 Background

Speech technology seeks to find ways of enabling electronic devices to acquire speech signals and processing them for any desired application. Speech processing is divided into three main areas[1], namely, synthesis, recognition and coding. Figure 1.1 illustrates different types of speech processing.

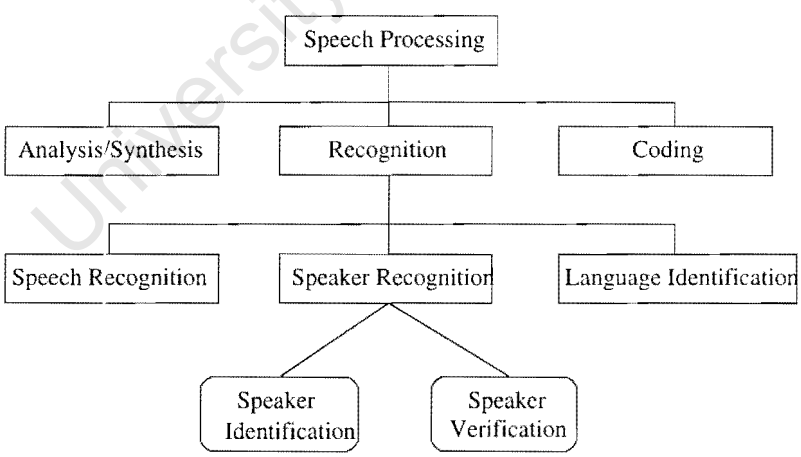

Figure 1.1: Speech processing categories.

Recognition in speech processing exists in three forms. These are language identification, speech- and speaker-recognition. Language identification[2, 3] involves modelling the phonotactic, prosodic and acoustic properties of languages. Accent

detection [4] and dialect identification [5, 6] utilise the same modelling techniques as language identification. On the other hand, speech recognition is the process of converting *speech* to text using a machine. However, speech recognition systems are adversely affected by the speaker-dependent variables, acoustical variables and the inconsistent manner in which a speaker pronounces words [7, 8].

By contrast, speaker recognition comprises both identification and verification [1, 9] of the speaker from a spoken word, phrase or sentence. However, this study concentrates on speaker identification. All these recognition tasks require speech input for analysis. The speech required for recognition analysis is commonly obtained from several commercial speech databases [10, 11, 12].

Speaker recognition is a pattern recognition problem [13] and therefore comprises both feature extraction and classification modules. The feature extraction part is known as the front-end, while the classifier is called the back-end. Figure 1.2 illustrates a generic pattern recognition system in the context of speech processing. Ideally, the features derived from a speech signal should reflect speaker-specific information for optimal performance.

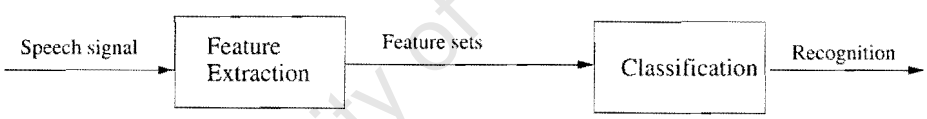

Figure 1.2: Pattern recognition for speech processing.

Speaker identification and verification are similar systems in that they both recognise a speaker from his or her voice [14]. The only difference is that a person claims his or her identity in speaker verification tasks, whereas no prior identity claims are made for speaker identification. Both systems can either recognise a person from a prompted phrase (text-dependent recognition) or from any unknown utterance (text-independent recognition). Speaker identification is a harder problem than speaker verification due to several factors as listed in table 1.1 [7, 9].

Campbell [1] suggests that other factors that affect the performance of a speaker recognition system include extreme variations in a speaker's emotional conditions, speech acquisition microphone placement, inconsistent room acoustics, channel mismatch, sickness and aging. Over the past four decades, several studies (cited in Chapter 2) were undertaken in order to develop robust speaker recognition systems

despite these constraints [14]. This study contributes towards speaker identification (SiD) research by solving some of the problems which hinder the robustness of SiD systems. Researchers tackle one or few problems at a time so that the gradual development towards robust SiD systems can be attained. The issues that motivate the execution of this study are dealt with in the next section.

1.2 Subject and Scope of this study

Several factors that hinder the performance of SiD systems have already been mentioned in the previous section. The size of a speaker population and the degradations introduced by noisy communication channels (e.g. a telephone channel) are some of the major problems in speaker identification [15, 16]. Although some researchers [17, 18, 19, 20] have observed the large population problem, very few investigated methods of solving it. There are various studies which have covered ways of compensating the noise and spectral shaping introduced by a communication channel [21, 22, 23, 24]. Nevertheless, this study aims to address the large population problem in order to achieve high performance.

The speech database used for all evaluations in this study is called NTIMIT which is a derivative of TIMIT [10, 25]. The TIMIT database was created by Texas Instruments (TI) in collaboration with Massachusetts Institute of Technology (MIT). The NTIMIT (N*oisy* TIMIT) database was created by transmitting TIMIT through tele-

Identification	Verification
Speaker may be reluctant	Speaker is cooperative
Voice disguise a problem	Mimicry a problem
Must test many patterns	Need compare to only one pattern
System response can be slow	System response must be fast
Vocabulary may be different	Vocabulary can be restricted to standard test phrase
Channels may be poor or different	Can frequently control channel characteristics
Signal-to-noise ratio may be poor	Can usually control signal-to-noise ratio

Table 1.1: Differences in specifications of SiD and SV [7].

phone channels. The total number of speakers in both TIMIT and NTIMIT database is 630. This population of 630 speakers is regarded as large population in speaker recognition evaluations.

An SiD process has two phases, namely, training and testing. During training, the extracted features are used to estimate the parameters that define a speaker model. Two feature extraction methods have been used in this study. Feature extraction is the way of extrapolating *speaker-specific* information from a speech signal. Using the feature sets, the speaker model parameters are estimated during the training of the classifier. The test speaker's voice patterns are compared to the speaker models. The speaker whose model best matches the test pattern is considered as the identified speaker. The voice patterns are represented as feature sets which occupy a multi-dimensional feature space. These features are assumed to have a multivariate Gaussian densities. The Gaussian densities associated with different speakers overlap in the feature space if there are a large number of speakers that enroll into the system. Figure 1.3 illustrates this phenomenon. These overlaps are caused by speakers with similar voices. It has been observed that the performance of an SiD system decreases with increasing enrollment population [9, 15].

Gaussian mixture models have been used to represent the enrolled speakers. The mean and the covariance are the parameters that represent the enrolled speakers' voice patterns. Part (a) of figure 1.3 shows that speaker A can be clearly identified because her test pattern is lying very close to her own *train* pattern. Part (b) however, indicates that the introduction of Speakers C and D limits the chance of correctly identifying speaker A. This ultimately leads to poor SiD performance.

In order to test if misclassified speakers sounded alike, an objective test was carried out by listening to their speech. The results hinted that the overlap of models was due to speakers who sounded the same. It is rather complex to directly solve these overlap problems since speaker-specific information cannot be accurately extracted from a speech signal in such a way that the speaker's unique physiological characteristics are isolated. Physical attributes such as vocal tract dimensions and tongue shape could be ideal for perfect speaker identification if it was easy to obtain them

from a speech signal.

This study proposes hierarchical methods for solving the large population problem without deteriorating the performance of the *existing SiD system* [26] (hereafter called the *baseline system*). Hierarchical methods in this study refer to the procedures which classify a smaller population of speakers given the large population of enrolled speakers (see figure 3.1). These methods should ideally minimise the inter-speaker confusions on the feature space. Several researchers implement the "hierarchical systems" according to hierarchical configurations which suit their investigations [27, 28]. The hierarchical methods proposed for this study are classified into two categories, namely, the speaker group detection hierarchical method and the N-best list hierarchical method.

1.2.1 Group detection hierarchical method

Figure 1.4 illustrates an ideal result if the group detection hierarchical method was to work perfectly. This result is illustrated as the *dotted* horizontal line labelled " Ideal hierarchical system performance" in figure 1.4. The experimental results may, however, hypothetically lie between the two graphs in figure 1.4.

The group detection method tries to solve the problem by grouping the speaker models according to several groups. For example, if the number of enrolled speakers is

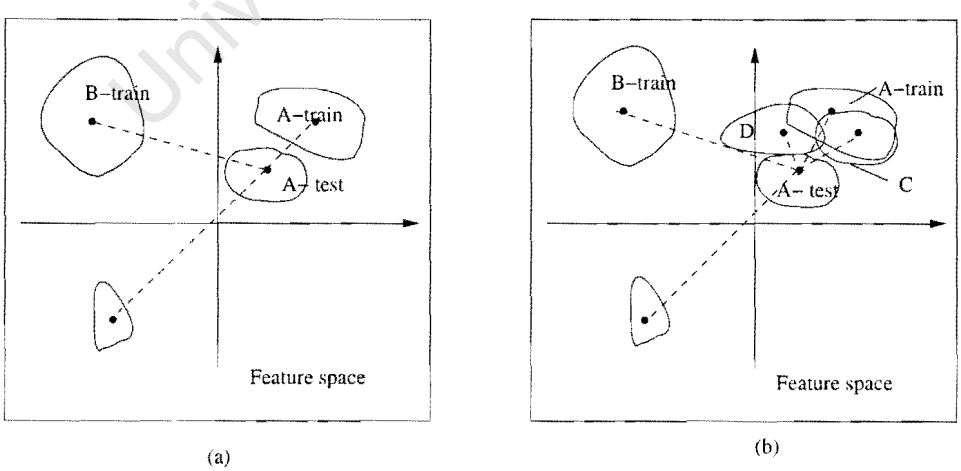

Figure 1.3: Overlapping speaker voice patterns in the feature space.

500, the group detector can equally place their models in groups A, B, C, D and E as illustrated in figure 1.4. This is done using a group detection algorithm that categorises speakers using speech information which is *different* from the baseline feature extraction method (uncorrelated features). Using figure 1.4 as an example, this fixed number of speakers is 100 for group A and the baseline identification is above 90%. It is hypothesised that if the test speakers whose features belong to either group A or B or C etc. are compared to the corresponding group models (e.g. B to B), then the hierarchical system would perform identification on 100 speakers if it were to be perfect. This implies that the system should correctly identify a speaker 90% of the time. The problem with the above mentioned hypothesis is that, the group detection algorithm might not necessarily split speakers equally. Furthermore, the same speaker might be placed into one group during training and another during testing. The latter case implies a definite identification error. This means that in reality the performance may not be as high as 90%. This approach raises a question of investigating how far a group detection algorithm can improve the baseline system. Most of the reported results, however, explore how far a perfect group detector can improve the baseline system. Abdulla and Kasabov [29] use the gender group detection method in speech recognition by splitting the speakers into groups of male and female speakers.

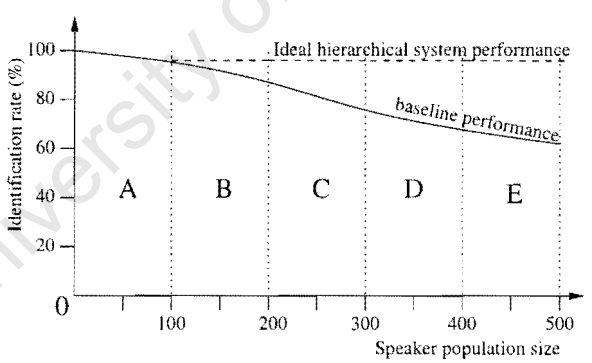

Figure 1.4: Hypothetical solution to large population SiD.

1.2.2 N-best hierarchical method

The N-best hierarchical method is the second proposed hierarchical algorithm in this study. Preliminary experiments and results of this method have been reported previously [30]. This method is termed "hierarchical" because identification is done on a smaller set of speaker models. This method has two stages. The first stage determines the likelihood scores on the high population of speakers, while the second stage classifies speakers using the top N speakers. The first stage is similar to the baseline system. The second stage uses the front-end to extract new features from the top N scoring speakers using the training data of the baseline process. The *second* front-end should have uncorrelated errors with the baseline feature extractor. The best of the N top models is regarded as one which belongs to the test speaker indicating correct identification. The second stage is meant to improve the core (first) stage of the identification process. The N-best list of speakers is selected according to a decision threshold. The studies on which this proposal is based employ different kinds of back-ends in order to build similar kind of a hybrid system [18, 19, 31, 32]. The threshold or decision level which decides when to resort to the N-best list stage is crucial during the implementation of this method. The main disadvantage of this method is that the first classification is done on large population of speakers and therefore the hypothesis illustrated in figure 1.4 is not entirely feasible. From previous work [19, 31] it is known that the hybrid system's performance improves the baseline identification. It can therefore be concluded that the aim of improving our baseline system's performance is possible.

1.3 Objectives of the Thesis

The objectives of this study are to:

- implement the baseline speaker identification system.
- evaluate the baseline performance using NTIMIT.
- design, implement and evaluate algorithms that are aimed at solving the large population size problem.

- report final performance that illustrates the solution to large population problem.
- draw necessary conclusions and recommend further research.

1.4 Contribution of this study

Literature which is related to the proposed hierarchical methods is reported in chapter 3. The proposed work introduces the group detection module to the conventional *front-end back-end* SiD system. This approach was not found in the available literature. The group detection module fails if not perfect and therefore the investigation is more of ground level hypothesis that needs further research. The series of experiments under a group detection hierarchical system in this study largely assume a perfect group detection algorithm so as to determine whether the idea of grouping speakers for identification holds. This part of the investigation also hints at the possible large margin of improvement if the group detection hierarchical system were to perfectly work without "forcing" the grouping routine. Similar work has been previously done. Abdulla and Kasabov [29] illustrate exactly the same system configuration in speech recognition. Our approach is in speaker identification and on large population compared to the five speakers used in [29]. Chung-Hsien Wu and Jau-Hung Chen [33] implemented the three level hierarchical system with the main aim of reducing the identification time. They use the population of 36 speakers. Their system also tries to reduce the over-crowding of the feature space. Pan et al [28] reduce training time by the use of on-line hierarchical system for the large population of 290 speakers. The highest score is determined from the two sub-databases using vector quantisation techniques. Our system differs from these two in that, the identification is only performed on one relevant group out of several groups of model databases.

The N-best methods of enhancing speaker identification scores is used by several researchers [19, 31, 32, 34]. Almost all of these studies improve the scores using different back-ends. The proposed study introduces ways of using feature extraction engines that produce uncorrelated errors. The same back-end is used for both feature extractors. This proposal introduces the retraining process of the top N-best scoring speakers from the first classification stage. It uses a different feature-extractor in-

stead of a different back-end as commonly found in literature. The features are then modelled again for final classification on the N-best list of models. Fine *et al* [31, 32] use support vector machines (SVMs) to enhance Gaussian mixture models (GMM) [35] scores using a certain decision threshold. A decision threshold which triggers the complementing feature extractor to process information is also utilised in this study. The configuration of this proposed system is described in chapter 4. The main results of this study evolve from the use of LPC cepstrum (LPCC) [36] as the second feature extractor for the retraining process. The baseline feature extraction utilises parameterised features sets (PFS) [37]. Gaussian mixture models are used in the back-end of the baseline system. Ultimately, the improvement of the baseline system is achieved. The proposed implementation contributes to the existing N-best list-based SiD investigations by using a new feature extraction process instead of another back-end. The second observation is that most N-best SiD systems utilise small speaker populations, (see table 5.12) while large population is the key issue for this work. The training and testing times do not differ much with the baseline execution.

1.5 Previous work

This section highlights some of the experiments that were carried out at the beginning of this research. The aim of these experiments was to investigate the possibilities of incorporating speaker identification using mobile phones. These small tasks were motivated by studies done by Grassi et al [38] who investigated the influence of GSM speech coding on speaker recognition. The full rate GSM compression and decompression (GSM codec) algorithm was used to code and decode TIMIT speech in order to generate a GSM transmitted speech database. The GSM codec compresses transmitted data for minimal bandwidth utilisation and decompresses it again on the receiving end. The codec used was the GSM 06.10 [39] which is publicly available. The author reported the detailed procedure of this experiment in [40]. A similar experiment was performed using real GSM speech from a local cellular network. The results are tabled and discussed in [41]. The author also reported results which demonstrated the inter-dialect and gender-confusion of speakers during identification [42]. The experiments on inter-dialect and gender confusion initiated

the hierarchical approach towards solving the large population problem [42]. From all these experiments it can be concluded that the GSM codec affects a speech signal by causing it to lose some speaker-specific information. These preliminary experiments have led to the proposal of further investigation into how well the simulated (i.e. GSM 6.10) and the real GSM data [43] can be identified compared to telephone speech (NTIMIT). However, this thesis concentrates on the performance of SiD on NTIMIT so that the results can be referred to future work that needs to be done on the identification of speakers using GSM speech.

1.6 Constraints

The nature of the NTIMIT database is the main constraint in this work. NTIMIT is the derivative of the TIMIT database [10, 25]. The TIMIT database was recorded using read sentences in one session. This database is ideal for text-independent speaker recognition systems but not usable for text-dependent ones. The text-dependent capability of NTIMIT database could have been useful for testing the proposed group detection algorithm which displayed inaccurate performance under text-independent conditions. The text-dependent nature of NTIMIT could have yielded better conclusions on how well the proposal works. The financial constraints and the time to get proper text-dependent capable databases made it difficult to perform text-dependent identification for this particular task.

Another limitation on speaker identification research is that most studies have not found results which are good enough for SiD commercial deployment [14]. The difficulty in reaching high performance illustrates the complexity of the SiD problem. It is therefore not conclusive that these new proposals will exhibit large improvement margins as shall be seen from the results in chapter 5. Our system evaluation includes the statistical significance test. The limitation of this kind of evaluation in speaker identification is that virtually all the cited publications in this thesis do not reveal any statistical significance test and therefore it makes it difficult to draw conclusions on whether the published SiD system performances are significant or not. This also limits the comparisons in terms of significance levels commonly used in this type of research. Gillick and Cox also state that these evaluation tests are almost always overlooked in speech processing research [44].

Related studies employ different speech databases for tests. These varying databases make it difficult to report valid comparisons. Consequently, the conclusions made from literature reviews are subject to a certain level of inaccuracies.

1.7 Plan of development

Chapter 2 is a detailed description of a speaker identification system. This chapter first describes the databases used in speaker identification research. Front-ends and back-ends which are popular in SiD evaluations are explained and their examples are provided. The problems associated with the poor performance of SiD are highlighted and some of the solutions provided by different researchers are cited. Finally some applications that can utilise SiD systems are given.

Chapter 3 is a literature review of some of the work related to the proposed hierarchical methods of improving speaker identification. The methods discussed in this chapter are not specifically on speaker identification but also on speaker verification and speech recognition.

Chapter 4 describes the way in which the baseline SiD system is configured together with some fundamental parameters associated with the design. The chapter proceeds to illustrate and the proposed hierarchical SiD system architectures. The difference between the existing system configurations mentioned in chapter 3 and the proposed systems is clarified in this chapter.

Chapter 5 begins by tabulating the results which justify the rest of the evaluation procedures with associated parameters. Subsequently, the group detection hierarchical system performance is reported and necessary discussions are made. The N-best hierarchical system performance is also illustrated by the results together with the relevant discussions. Finally a short analysis of the system performance based on the results is made.

Chapter 6 concludes this report based on the results that were obtained and evaluated in chapter 5. Future directions relevant to the the hierarchical SiD system approach are also included in this chapter.

An appendix is provided at the end of this document. Appendix A contains some

relevant data tables which were used to generate some of the important evaluation graphs in chapter 5.

Chapter 2

Speaker Identification System

Speaker identification is the process of identifying a talker out of a group of speakers using his voice signal. It comprises two processing stages, namely, feature extraction and classification. Feature extraction is the process of extracting speaker related characteristics from the speaker's utterance. Several feature extraction algorithms exist which capture speaker-discriminative information such as vocal tract length and pitch. During classification these features are first used to build the speakers' models as they enroll into the system. This process is known as training. Secondly a decision logic inside the classifier utilises the speech features or characteristics from the test speaker and compares them to the existing models in order to find the model which matches them most closely. Finally, the closest matching model is considered as the one that represents the unknown speaker. This final step is called testing.

Speaker identification is divided into two categories, namely, text-dependent and text-independent. Figure 2.1 is an illustrative categorisation of the SiD systems. Text-dependent SiD [45] requires fixed and known sentence(s) from a speaker. Text-independent SiD system [21] acquires any utterance that a speaker projects regardless of the meaning of the sentence or phrase. Both text-dependent and text-independent SiD systems can either be closed set or open set.

A closed set SiD system [1] assumes that the test speaker is among the *a priori* known speakers who have enrolled into the system. Let S be the total number of enrolled speakers. The identification decision for one talker is $1/S$ for a closed set SiD system. An open set SiD [19] constitutes the set of enrolled speakers with

no prior knowledge of the existence of the test speaker in the enrolled set. The identification success rate is then $1/(1+S)$ in order to allow the possibility that the speaker's model might not exist among the known set of S enrolled speakers. In general, the performance of the SiD system is determined by the ratio of the number correctly identified test speakers to total number of enrolled speakers. For example, if 10 speakers are tested, one at a time, against 10 models, the SiD success rate is 100% if all 10 are correctly identified. The SiD success rate of 80% occurs if only 8 speakers are correctly identified out of these 10 models.

Speech used by most researchers in this field comes from commercial speech databases [11, 12]. Some prominent databases used in research are described in section 2.1. The databases differ a lot depending on the nature of research. The way of reading databases depends on the sampling rate of speech and also the header information of the creators [10] of the database.

Several algorithms for extracting the features exist in speaker identification studies [46]. These algorithms are called feature extraction methods. The short-term speech segment is analysed using the feature extraction algorithm so that the waveform is assumed to be time-invariant. The information obtained from the feature extraction process is kept as a feature vector. This means that several feature vectors are computed for each speaker utterance since one utterance could have many speech segments which are also known as frames. Speaker related information such as vocal tract dimensions is contained within the feature vector. Section 2.2 highlights some of the generally used feature extraction methods.

Speaker information patterns resulting from feature vectors are modelled into a clus-

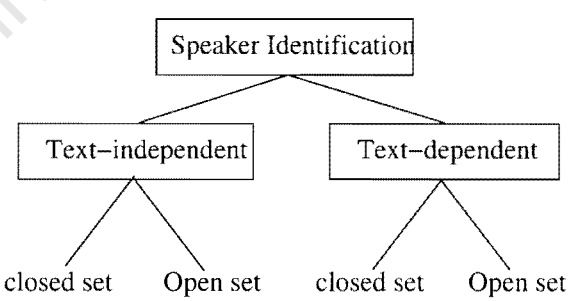

Figure 2.1: Categories of speaker identification system.

ter which is uniquely placed in the feature space to represent a particular speaker. The test speech feature patterns are also placed in the feature space so that they can be compared to the enrolled speakers' models. The decision logic uses distortions or distances [13] to find out how closely the test patterns compare with the existing patterns in the feature space. Classifiers are responsible for both the creation of the speaker model and identification of the speaker. Several classifiers that are used in SiD systems are mentioned in section 2.3. Figure 2.2 shows a generic speaker identification system.

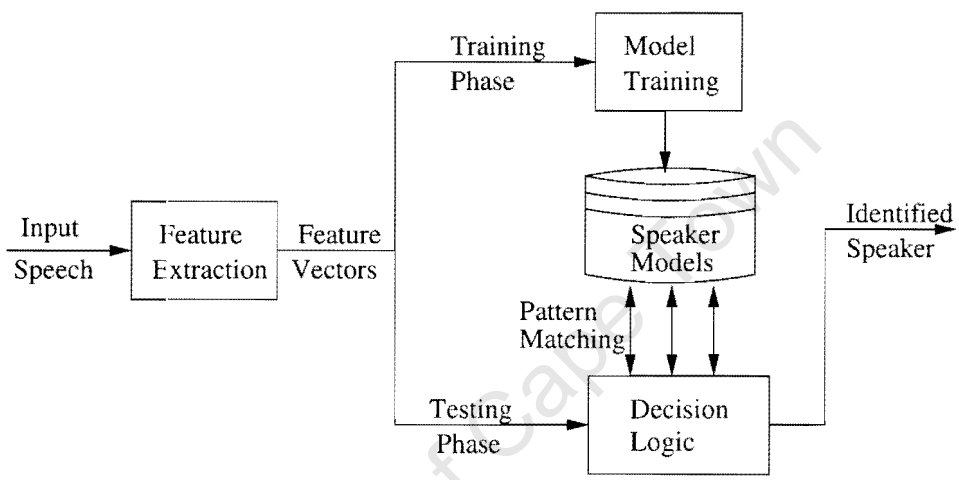

Figure 2.2: Generic speaker identification system.

The feature extraction part of SiD is called the front-end while the classification part is known as the back-end. The front-ends and back-ends are designed to optimally achieve high performance. This is not normally the case because the microphones used to acquire speech introduce some noise [47, 48]. This speech acquisition related problem is one of many other problems that hinder good speaker identification performance. Section 2.4 is a slightly detailed description of some factors responsible for low SiD system performance. Although this study deals with the solution towards separating the confusable speaker models in the features space, previous work which deals with other issues is cited in section 2.4. This chapter is concluded by highlighting some of the possible applications of speaker identification.

2.1 Speech databases

It is beneficial for most researchers to evaluate their SiD systems using similar speech data in order to make it possible to compare the performance of their architectures. Some databases were recorded in ideal environments in order to use them as baseline data for testing new SiD system designs. Other databases were recorded under varying conditions or even using different microphones in order to emulate real-life situations. The SWITCHBOARD corpus in table 2.1 is an example of a *close-to-reality* database. The other speech corpora are obtained by transmitting speech through communication channels. SIVA, SWITCHBOARD and POLYCOST [10] are examples of databases that were created by transmitting speech over telephone lines. In a real SiD task speakers might enroll into the system at a given time but only get tested a month or even a year later. This makes their voices slightly different from the enrollment utterance. In some databases, data is collected over a period of days, weeks, or even months. These spread recording times are called *intersession* intervals [10] as shown in table 2.1. Many databases and their specifications are listed in [12].

As stated previously, NTIMIT is the database of choice for this study. It was obtained by transmitting TIMIT sentences through multiple channels of a public switched telephone network (PSTN) and has a speaker population of 630. Of these, 438 are male and 192 are female [25]. Each speaker utters 10 sentences each of which lasts for roughly 3 seconds. The first two sentences are the same for all speakers while the remaining 8 differ from speaker to speaker. In addition, all speech files are sampled at 16 KHz. This makes NTIMIT a reasonably good candidate for text-independent SiD evaluations. Speaker identification studies normally utilise more sentences to train and less to test [49]. For instance, 8 sentences may be used for training and 2 for testing. Speakers in NTIMIT databases have labels and they come from eight different dialect regions of the United States. The next sections discuss how speech from a database is processed throughout the SiD process.

2.2 Front-ends

As stated previously, the frond-end is the signal processing module of an SiD system. It is in this block that the integrity of a signal is measured. This refers to the measurement of sampling rates and the speech file formats depending on the database used. The speech is then segmented into small frames. The final output of the front-end are the feature sets (or vectors) which should contain characteristics unique to a speaker. A block diagram of a front-end is depicted in Figure 2.3. There are several features which convey speaker-specific information from the speech signal. Examples of such features are based on pitch frequency, dialect or accent, vocal chords shapes and vocal tract length or shape [7, 9].

Various feature extraction methods have been implemented and optimised in order to improve identification rates of SID systems. Examples of such algorithms

Database	**Language**	**Acoustic Environment**	**acquisition**	**Type of Speech**	**Intersession**
TIMIT	English	Sound booth	close-talking microphone	read sentences	none
YOHO	English	office	high quality handset	Prompted digit phrases	days-month
SWITCHBOARD	English	home/office	PSTN+ variable handsets	Conversational	days-weeks
SIVA	Italian	home/office	PSTN+ variable handsets	Prompted words and digits, Short questions and read text	days-months
POLYCOST	English	home/office	ISDN+ variable handsets	Prompted digit strings, read sentences, free monologues	days-weeks

Table 2.1: Various databases taken from Campbell.

are linear prediction cepstral coefficients (LPCC) [8, 36, 50], mel-frequency cepstral coefficients (MFCC) [51], noise compensation filters such as RelAtive SpecTrA (RASTA), [52] etc. This study employs parameterised feature sets (PFS) which are essentially MFCCs with extra parameters. These features were introduced by Mashao in his PhD thesis [53]. PFS have proved to be competitive since the SiD system attains 100% recognition accuracy on clean speech (TIMIT) which is a similar result achieved by most SiD systems [16, 52]. The PFS-based SiD system also yields about 72% accuracy on NTIMIT speech [26] compared to the 60% identification rate obtained from a similar system [15, 49] which uses MFCC and exactly the same back-end. Some LPCC-based SiD systems which utilise the NTIMIT database show a slightly lower identification rate and are listed in table 5.11.

2.2.1 Signal pre-processing

Speech is an acoustic wave which is converted into an analogue signal using a telephone handset or a microphone [1] (see table 2.1). This analogue signal is sampled and segmented into presumably time-invariant short-term frames before it goes into the SiD system. A pre-emphasis filter is then applied to this speech segment [8, 9]. Pre-emphasis limits the signal's susceptibility to finite precision effects by amplifying the higher frequency spectrum. This pre-emphasised signal is then windowed at fixed time intervals which are normally 10 - 20ms [1, 46]. These intervals are considered to be time-invariant speech frames in order to perform short-term spectral analysis because it is computationally intensive to use the entire waveform. A Hamming window is normally used in most speaker identification studies [8, 46]. Windowing minimises the signal discontinuities at either end of the frame. The Hamming window function with length N samples (N=320 implies 20ms window for a sampling rate of 16 kHz) is shown in equation 2.1.

$$w(n) = 0.54 + 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \leq n \leq N-1 \tag{2.1}$$

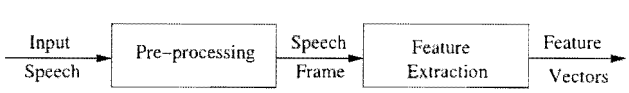

Figure 2.3: Generic front-end.

The windowed frames progress at the frame rate of usually 10ms which implies that frames overlap by about half the frame length if a 20ms frame is used. This overlap is meant to minimise aliasing. In order to avoid parameterising unvoiced speech or noise in a noisy speech signal, a voice activity detector (VAD) is sometimes added as a filter [9, 46, 15] prior to feature extraction process. Finally, the feature extraction process is carried out and a feature vector from a particular frame is generated. This means that the number of feature vectors is the same as that of VAD-accepted frames. Figure 2.4 below illustrates the pre-processing of speech.

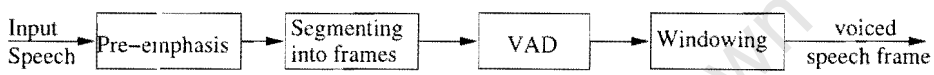

Figure 2.4: Preprocessing of the speech signal.

2.2.2 Feature extraction

Feature extraction process in SiD entails the computation of physiological attributes that individuate a speaker from a speech signal. This is because the speech-related uniqueness of a speaker comes from the vocal tract anatomy during voice production [8, 36]. Other characteristics of a speaker which are perceptually acquired when a person speaks are also utilised for feature extraction [21]. This is done because the human auditory system can enable one to identify pitch, accent or dialect, gender and age [7] of a speaker [21]. Ideally, a feature extractor should be able to extract speaker information regardless of environment, speech acquisition equipment and communication channel effects [14].

The earliest feature extractors were spectrally based ones [9, 54]. During the generation of these based feature sets, the windowed speech signal is transformed using the discrete Fourier transform (DFT) and then the log-magnitude of the resulting spectrum was computed. The log spectrum is further transformed into the *cepstral* domain using the inverse DFT (IDFT) as illustrated in Figure 2.5. However, the problem with the cepstral features is their lack of robustness to noise [7].

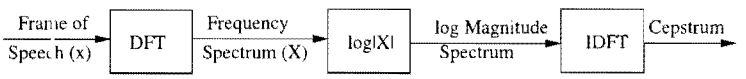

Figure 2.5: Cepstral features generation.

As research in speaker identification continues, more and better ways of extracting features emerge. The more prominent ones are vocal tract-related and mel-filterbank based feature extraction algorithms [1, 7]. The mel-frequency related cepstrum is more robust to noise compared to vocal tract features [49].

Popular feature extraction methods in speaker identification are mel-frequency cepstral coefficients (MFCC), [51, 54, 55, 56, 57], linear prediction cepstral coefficients (LPCC) [1, 38. 58, 59, 60], and their derivatives. Some studies show that it is possible to use auditory-based feature vectors *[8]*. The ensemble interval histogram (EIH) [8] and perceptual linear prediction (PLP) [61] are examples of auditory-based feature sets. There are several other feature sets which largely evolve from either LPCC, MFCC and auditory based techniques [7, 21]. Further details of the MFCC, PFS, LPCC and PLP are provided since the experiments performed for this investigation utilised them.

2.2.3 Mel-frequency cepstral coefficients (MFCC)

MFCC feature sets are commonly used in speaker recognition applications as discussed in the previous section. These MFCC features are derived from a mel-frequency filterbank [8]. Mel-frequency results from warping the frequency power spectrum of a windowed speech segment. The perceived difference in tone or pitch of a person reveals the gender, accent and even the individuality of that person.

Psychophysical studies have been carried out to find means of quantifying the way in which human ear perceives sound [8]. One of the findings of these investigations is that the human perception of the pitch of tones does not follow a linear scale. The pitch of the tone with some frequency f measured in Hertz was subjectively

measured on a scale called *mel*. The pitch of a 1000 Hz tone which is 40 dB above the perceptual hearing threshold [8] is the reference point of the mel-scale. This reference point is defined as 1000 mels. Figure 2.6 illustrates this subjective pitch as a function of frequency (f). It can be observed from the logarithmic graph that the *mel* is linear up to 1kHz. The mel-frequency is obtained as follows:

$$mel\, f = 2595 \log(1 + \frac{f}{700Hz}) \tag{2.2}$$

Equation 2.2 portrays the objective computational model which transforms the physically measured spectrum into mel-scale or a psychological subjective spectrum [8]. The simulation of this subjective spectrum is done by the use of a uniformly spaced filter bank using a nonlinear mel-scale as shown in figure 2.7. The triangular band-pass filters are spaced by 150 mels with a bandwidth of 300 mels. These schemes are popular in speaker identification research.

The warping of the power spectrum using mel-scale illustrated in figure 2.8 adapts the frequency spectrum to a resolution perceptible to the human ear [57]. The filter

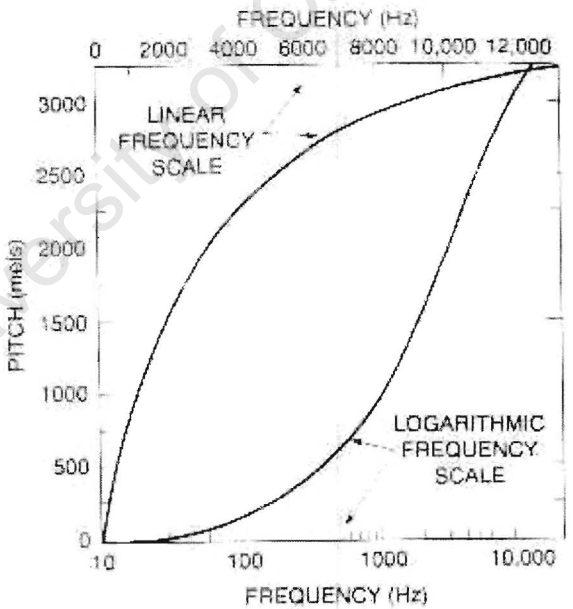

Figure 2.6: Subjectively perceived pitch of a tone as a function of frequency [8].

bank which consists of several overlapping triangular filters is applied. Finally, either discrete cosine transform (DCT) or inverse fast Fourier transform (IFFT) derives the MFCC feature vector related to a particular frame of speech.

2.2.4 Linear predictive cepstral coefficients (LPCC)

Linear prediction coefficients are based on the vocal tract characteristics of a speaker [36]. They therefore capture speaker information from the speech signal. The LPC coefficients [50] are obtained from the vocal tract transfer function $V(z)$, which evolves from the time-varying digital filter $H(z)$, whose steady-state form is expressed in equation 2.3 below:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 + \sum_{k=1}^{p} a_k z^{-k}} = \frac{G}{A(z)} \tag{2.3}$$

and $V(z) = \frac{1}{A(z)}$. p is the LPC order and a_k are predictor coefficients. G is a gain factor and u_n is the present input.

The linear prediction process starts by modelling the pre-processed speech signal,

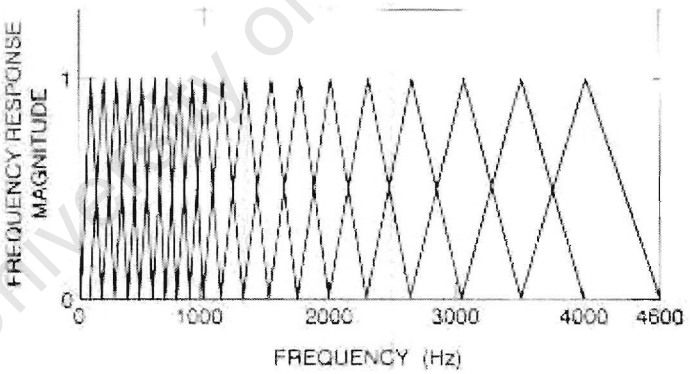

Figure 2.7: Mel-scale filter bank [8].

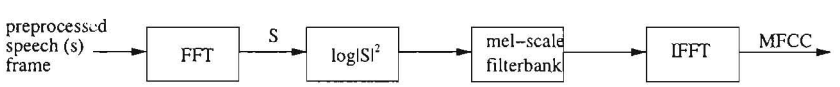

Figure 2.8: Mel-cepstral feature generation

s_n, as a linear combination of its past values s_{n-k} such that:

$$\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k} + G.u_n \tag{2.4}$$

Generally, u_n is unknown and therefore the predicted signal $\hat{s}_n$ is given by:

$$\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k} \tag{2.5}$$

Since the $G.u_n$ term from equation 2.4 is ignored, the prediction error e_n is expressed as a difference between the present signal and the predicted value as shown in equation 2.6.

$$e_n = s_n - \hat{s}_n \tag{2.6}$$

which yields:

$$e_n = s_n - \sum_{k=1}^{p} a_k s_{n-k} \tag{2.7}$$

The optimisation of LPC analysis is achieved by minimisation of the mean square prediction error (MSE) $E = \sum e_n^2$ [7, 50]. The autocorrelation function r_T expressed in equation 2.8 is used to minimise the MSE. T, also known as *lag*, is the number of previous samples of s_n used for prediction.

$$r_T = \sum_{n=0}^{N-1-T} s_n s_{n+T} \tag{2.8}$$

Expression 2.8 results in the matrix system of equation 2.9 which is solved using the Levinson-Durbin recursive algorithm [7, 36, 50].

$$\begin{bmatrix} r_0 & r_1 & \dots & r_{p-1} \\ r_1 & r_0 & \dots & r_{p-2} \\ \dots & \dots & \dots & \dots \\ r_{p-1} & r_{p-2} & \dots & r_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \dots \\ a_p \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \dots \\ r_p \end{bmatrix} \tag{2.9}$$

The Levinson-Durbin recursion in equation 2.10 solves for the predictor coefficients a_k of order p from the system of equations in equation 2.9. In this process, the predictor coefficients a_k for all orders less than p are obtained together with their corresponding mean square errors ($MSE_i = E_i/r_0$) ($i = 1, 2, ..., p$) . During recursion, the prediction order is increased and the corresponding error is determined until the value of the minimum corresponding error, E_i, is reached. This determines the termination of recursion. That is,

$$\begin{aligned}
E_o &= r_0 \\
k_i &= \frac{-\left[r_i + \Sigma_{j=1}^{i-1} a_j^{(i-1)} r_{i-j}\right]}{E_{i-1}} \quad 1 \leq j \leq p \\
a_i^{(i)} &= k_i \\
a_j^{(i)} &= a_j^{(i-1)} + k_i a_{i-j}^{(i-1)} \quad 1 \leq j \leq i-1 \\
E_i &= (1 - k_i^2) E_{i-1} \quad i = 1, 2, ..., p \\
a_j &= a_j^{(p)} \quad 1 \leq j \leq p.
\end{aligned} \tag{2.10}$$

The k_i in equation 2.10 are known as reflection coefficients [36]. In addition, the predictor coefficients a_k are then determined as a solution to equation 2.9. Finally, the LPC coefficients are calculated by way of the recursive LPC-to-cepstrum conversion routine using equation 2.11 below. This process is similar to taking the inverse Fourier transform of the predictor coefficients [9] and is described mathematically as follows:

$$c_n = a_n + \sum_{k=1}^{n-1} \frac{k}{b} c_k a_{n-k} \tag{2.11}$$

where c_n are the feature-sets that are further used by the back-end of the system.

2.2.5 Parameterised feature sets (PFS)

Parameterised feature sets (PFS) were proposed with the realisation that *the mel-scale* spectral compression used in most studies is constant. The signal parameterisation utilises two parameters that are varied on a two dimensional space for optimal performance of the SiD system [62]. Figure 2.9 illustrates how these features are generated.

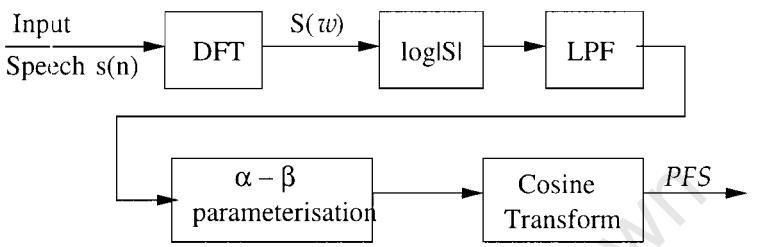

Figure 2.9: The PFS feature extraction process.

During the PFS feature extraction, the pre-processed speech frame is converted into the frequency domain using the discrete Fourier transform (DFT). The log magnitude spectrum is then calculated and subsequently filtered using a low pass filter (LPF) for the removal of high *quefrencies* (high frequencies of spectrum in the log magnitude domain). Let $X(m)$ be the log magnitude spectrum of length N and let the impulse response of the low pass filter be $h(m)$ of length N_f, where $N_f$is odd. $X(m)$ is warped according to equation 2.12 giving a warped sequence $X_f(m)$ of length $N + N_f$ [53].

$$\begin{aligned} X_f(m) &= X(N_f/2 - 3/2 - m) \quad m = 0, 1, 2, ..., (N_f - 1)/2 - 1 \\ X_f(m + N_f/2 - 1/2) &= X(m) \quad m = N_f/2 - 1/2, N_f/2 + 1/2, ..., N - 1 \\ X_f(m + N_f/2 + N - 1/2) &= X(N - m) \quad m = 0, 1, 2, ..., (N_f - 1)/2 - 1 \end{aligned} \tag{2.12}$$

The filtered log magnitude spectrum, $X_l(n)$ is then obtained by convolution as shown in equation 2.13.

$$X_l(m) = X_f(m) \star h(m) \tag{2.13}$$

Non-linear sampling is applied to the resulting spectrum, $X_l(m)$. This makes use of two parameters, α and β, defined in equation 2.14 as:

$$\sum_{i=1}^{\alpha} A\beta^{i-1} = \frac{N}{2} \tag{2.14}$$

In equation 2.14, A is a constant that is determined by α and β. It has been found to be the size of the first region of the parameterised spectrum. The parameter α represents the number of regions to be made on the spectrum while the β parameter determines the spectral compression if greater than 1. If $\beta = 1$, all the regions are uniformly sampled. Figure 2.10 illustrates the spectral compressions obtained when $\alpha = 8$ and varying the β parameter. The β parameter is also called the warping factor because it varies the spectrum as compared to mel-frequency warping illustrated in figure 2.6 which has a constant compression factor. The warped spectrum, $Z(r)$, is then formed from the sampled points as per equation 2.15.

$$Z(r_i + r) = X_l\left(\frac{\alpha A\beta^{i-1}}{R}r\right) \tag{2.15}$$

where $r = 1, 2, ..., \frac{R}{\alpha}$, $r_i = \frac{R}{\alpha}i \quad for \quad i = 1, 2, ..., \alpha$ and R is a constant greater than feature vector dimension.

The cosine transform [63] is finally applied to the warped spectral magnitude giving the PFS cepstrum (cepstral coefficients), c_n, according to equation 2.16. That is,

$$c_n = \sum_{r=0}^{R-1} Z(r)cos\left(\frac{n(r+0.5)}{R}\right), \quad n = 1, 2, ..., F \tag{2.16}$$

where F is the size of the feature vector. These parameterised feature sets have been used in several experiments [26, 63, 64] and the performances obtained are reported in chapter 5.

2.2.6 Perceptual linear prediction

The perceptual linear prediction (PLP) feature extraction technique was proposed by Hermansky for speech recognition applications [61]. This method estimates the au-

ditory spectrum using the critical-band spectral resolution, the equal-loudness curve and the intensity-loudness power law. This method tries to discard or preserve the spectral details of a speech signal according to their auditory prominence. Figure 2.11 demonstrates the PLP analysis.

The pre-processed speech frame is transformed into the frequency domain using the FFT. The power spectrum, $P(\omega)$, is then calculated according to equation 2.17 if $S(\omega)$ the frequency spectrum of speech.

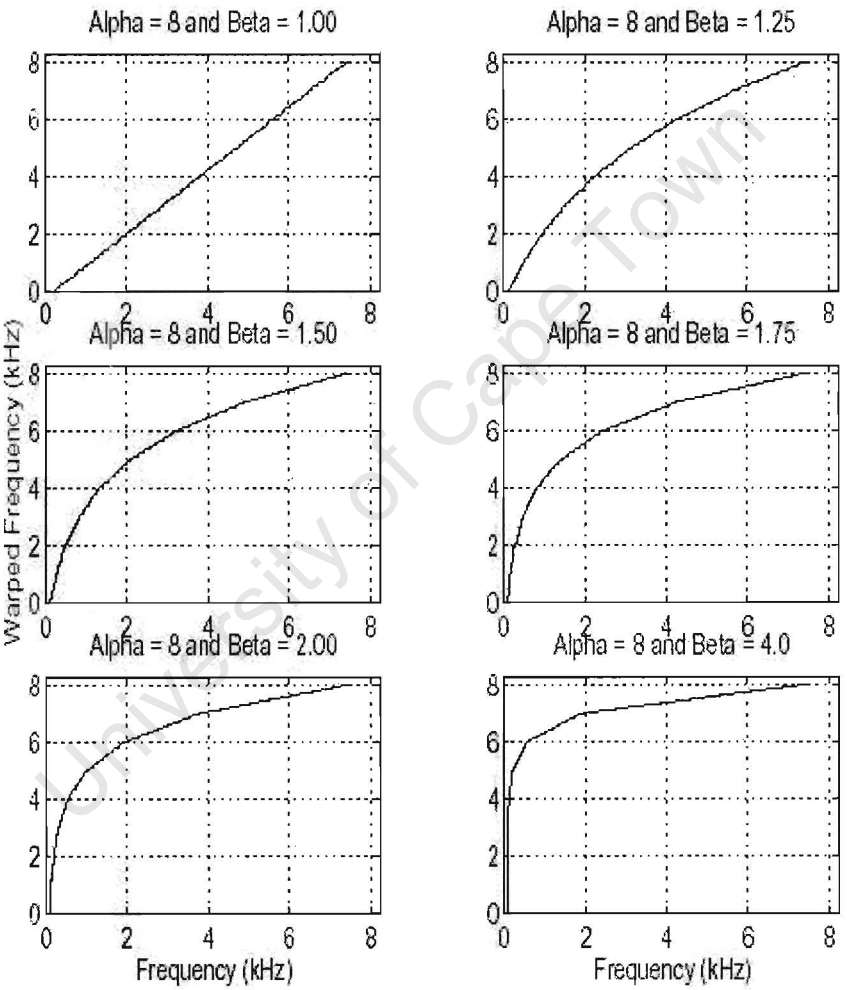

Figure 2.10: The varying spectral warping scales in PFS optimisation [53].

$$P(\omega) = Re[S(\omega)]^2 + Im[S(\omega)]^2 \tag{2.17}$$

The frequency scale ω is transformed into the Bark [8] frequency scale Ω by:

$$\Omega(\omega) = 6ln\left(\frac{\omega}{1200\pi} + \sqrt{(\omega/1200\pi)^2 + 1}\right) \tag{2.18}$$

The warped spectrum $P(\Omega)$ is convolved with the spectrum of the simulated critical-band masking curve, $\Psi(\Omega)$. This curve is a rough estimate of the shape of auditory filters. The critical band curve is given in equation 2.19 below.

$$\Psi(\Omega) = \begin{cases} 0 & for\ \Omega < -1.3, \\ 10^{2.5(\Omega+0.5)} & for\ -1.3 \leq \Omega \leq -0.5, \\ 1 & for\ -0.5 < \Omega < 0.5, \\ 10^{-1.0(\Omega-0.5)} & for\ 0.5 \leq \Omega \leq 2.5, \\ 0 & for\ \Omega > 2.5. \end{cases} \tag{2.19}$$

The convolution generates the critical-band power spectrum $\Theta(\Omega_i)$ according to equation 2.20. That is,

$$\Theta(\Omega_i) = \sum_{\Omega=-1.3}^{2.5} P(\Omega - \Omega_i)\Psi(\Omega) \tag{2.20}$$

This convolution process reduces the spectral resolution of the power spectrum more than $P(\omega)$. The critical band-spectrum is then pre-emphasised by the simulated equal-loudness curve, $E(\omega)$. This curve is the approximation to the non-equal sen-

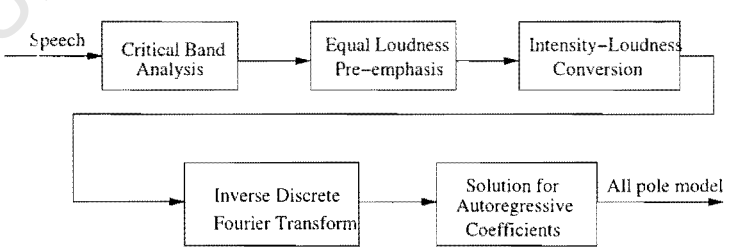

Figure 2.11: Perceptual linear predictive (PLP) speech analysis.

sitivity of human hearing at different frequencies. $E(\omega)$ simulates the sensitivity of hearing at about 40 dB. An example of this approximation is given in equation 2.21.

$$E(\omega) = \frac{(\omega^2 + 56.8 \times 10^6)\omega^4}{(\omega^2 + 6.3 \times 10^6)^2 \times (\omega^2 + 0.38 \times 10^9)} \tag{2.21}$$

The cube root of the pre-emphasised critical-band spectrum is calculated. This process is an estimated power law of hearing. This operation also reduces the spectral-amplitude variation for achieving the low order autoregressive modelling (see figure 2.11). Finally, inverse DFT is applied to the cube-root compressed spectrum and the autoregressive coefficients. These coefficients are then used as feature sets.

2.3 Back-ends

The feature vectors obtained from the feature extraction algorithms described in the previous section are evaluated at the back-end of the SiD system. The back-end is made up of speaker models and the decision logic. During training, the feature vectors obtained from the training utterance are used to generate the model for each speaker. In the test phase, the feature sets from the unknown speaker's utterance are extracted and then compared with each model of the enrolled speaker in order to obtain a set of scores. Clearly, the number of scores obtained is equal to the total number of enrolled speakers. Clustering algorithms such as k-means [65] and expectation maximisation (EM) [46] are used by classifiers in rendering the recognition decision because speech data changes with time as the speaker's voice varies with time. These clustering algorithms are generally used where the feature sets can easily discriminate between two speakers. Since there are no good enough feature sets for accurate separability of speakers, this investigation uses the clustering algorithms to update the feature patterns as the speaker varies his or her utterance [13].

For mathematical tractability, the pattern vectors are assumed to have multivariate Gaussian densities [7]. These Gaussian density parameters such as the mean and covariance are used to represent a speaker model during the training phase of the speaker identification process. The Gaussian modelling techniques such as hid-

den Markov models (HMM) [66] and Gaussian mixture models (GMM) [49] are used to model the feature vectors of individual speakers. Neural networks [13] are also a back-end that discriminates the feature patterns by way of creating decision boundaries between classes as appropriate. However, most researchers use GMM [46, 15, 49], neural networks [19, 34, 52, 60] and their derivatives. However, since, this study uses GMM as a back-end, a more detailed description of the GMM classifier is provided in section 2.3.1. Vector quantization (VQ)[13, 28, 33] is also popular in speaker identification. Other back-ends used in SiD include support vector machines (SVM) [31, 67], nearest neighbour (NN) and dynamic time warping (DTW) [13].

The decision logic compares the unknown feature vector, $\vec{x}$, with each model in the database and finds the probability that $\vec{x}$ represents the model, λ_j, which is the probability of the model given an unknown feature vector ($p(\lambda_j|\vec{x})$) [7], where λ_j represents a model for speaker *j*. If there are *N* models, then there will be *N* probabilities ($p(\lambda_j|\vec{x})$) which are called scores. The highest among the *N* probabilities is considered to be from the model of the speaker whose utterance has generated the feature vector, $\vec{x}$ and therefore that speaker becomes the identified one. Since $\vec{x}$ is the observation feature vector, $p(\vec{x}|\lambda_j)$ is obtained by using the Bayes' theorem as follows:

$$p(\lambda_j|\vec{x}) = p(\vec{x}|\lambda_j)\frac{p(\lambda_j)}{p(\vec{x})} \tag{2.22}$$

The decision rule is, for any unknown feature vector $\vec{x}$, choose the model λ_j that maximises $p(\vec{x}|\lambda_j)$. Let the *m* the feature dimension and $\vec{\mu_i}$ be the mean of all $\vec{x_j}$. The probability $p(\vec{x}|\lambda_j)$ is estimated using the probability density function (b_i) [7] for model *j* and it is expressed as shown in equation 2.23 below:

$$b_i(\vec{x}) = (2\pi)^{-m/2}|\Sigma_i|^{-1/2}\exp(-\frac{1}{2}(\vec{x}-\vec{\mu_i})^T\Sigma_i^{-1}(\vec{x}-\vec{\mu_i})) \tag{2.23}$$

where Σ_i is the covariance matrix for feature i. Now, $b_i(\vec{x})$ corresponds to $p(\vec{x}|\lambda_j)$ and $p(\vec{x}|\lambda_j)$ is maximised using ($\ln b_i(\vec{x})$) for computational simplicity. The constant $(2\pi)^{-m/2}$, from equation 2.23 can be ignored so that the expression that needs to be maximised is as follows:

$$\ln\left\{|\Sigma_i|^{-1/2}\exp(-\frac{1}{2}(\vec{x}-\vec{\mu_i})^T\Sigma_i^{-1}(\vec{x}-\vec{\mu_i}))\right\}. \tag{2.24}$$

The minus signs and the factor of $\frac{1}{2}$ can be left out in order to minimise what is called the *distance,* D_i, expressed in equation 2.25.

$$D_i(\vec{x}) = (\vec{x}-\vec{\mu_i})^T\Sigma_i^{-1}(\vec{x}-\vec{\mu_i}) + \ln|\Sigma_i| \tag{2.25}$$

This process is called the basic maximum likelihood criterion[7, 15] and D_i is not necessarily a distance by definition since $D_i(\vec{x}) = \ln|\Sigma_i|$ cannot be zero. Other simplifications are made to equation 2.25 so that the distance can be similar to the distance in the Euclidean space as shall be described in the following paragraph. In speech and speaker recognition, *distance* is used to compute the distance of the unknown input feature vector from the centroid of each class (word or speaker). Thus the class for which the distance is minimum is selected.

In practice, the distances used in speaker identification are mainly the Euclidean and Mahalanobis distances [7]. These distances are utilised based on the knowledge that the databases used in speaker recognition cannot always get reliable estimates of the covariance matrices. For this reason, the covariance matrix Σ_i may be assumed constant for all classes. Then the term $\ln|\Sigma_i|$ in equation 2.25 is constant and can thus be ignored. The result is the so-called Mahalanobis distance as expressed in equation 2.26. That is,

$$D_i(\vec{x}) = (\vec{x}-\vec{\mu_i})^T\Sigma^{-1}(\vec{x}-\vec{\mu_i}). \tag{2.26}$$

This distance measure is commonly used in speaker identification ***[7, 15]***. On the other hand, the Euclidean distance is given by:

$$D_i(\vec{x}) = \sum_{k=0}^{m}(x_k - x_{ik})^2 \tag{2.27}$$

assuming that the features are uncorrelated, Σ_i is the same for all classes and that all features have equal variances. The covariance matrices, Σ_i, are diagonal matrices whose elements are variances σ_i. In addition, x_{ik} is the k^{th} element of a feature vec-

tor i. Figure 2.12 shows a detailed back-end diagram which includes the decision logic block that utilises these distances.

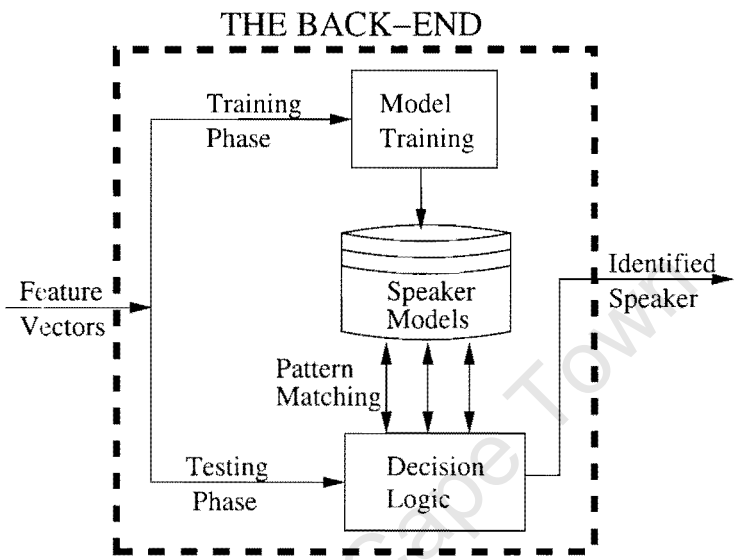

Figure 2.12: The SiD back-end.

2.3.1 Gaussian Mixture Models

The feature vectors are modelled by the GMM during the training phase of SiD. Each feature vector in a feature set is modelled using a d-variate Gaussian probability density function (pdf) with a state-dependent mean and covariance [49] shown in equation 2.28. It has already been mentioned in section 2.3 that the features are assumed to have multivariate-Gaussian densities. The characteristics of the type of feature vectors have hidden states that correspond for example to the perceptual scale or vocal tract configuration of a particular speaker. Equation 2.28 illustrates the pdf $(b_i(x))$ for state i as a function of the d-dimensional feature vector $\vec{x}$. That is,

$$l_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{d}{2}}\sqrt{\Sigma_i}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu_i})^T(\Sigma_i)^{-1}(\vec{x}-\vec{\mu_i})\right\} \tag{2.28}$$

The mean vector, $\vec{\mu_i}$, represents the expected feature vector while the covariance matrix , Σ_i, represents the correlations and variability of spectral features within state i [15]. The number of states is known as the number of mixtures [68]. The M-state Gaussian mixture speaker model, λ, is generated using a Gaussian mixture density function

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} c_i b_i(\vec{x}) \tag{2.29}$$

where M is the number of mixtures and c_i is the probability of the feature vector being in each state (also sometimes called the mixture weight). The mixture weights must add up to unity according to the constraint

$$\sum_{i=1}^{M} c_i = 1 \tag{2.30}$$

for the GMM pdf, $p(\vec{x}|\lambda)$, to be normalised [68]. The speaker model, λ, is represented by equation 2.31 below.

$$\lambda = \{c_i, \vec{\mu_i}, \Sigma_i\}, \quad for\ i = 1, 2...., M \tag{2.31}$$

The unsupervised clustering technique that is normally used for the GMM model parameter estimation is the expectation-maximization (EM) algorithm [15]. From the training data the EM algorithm maximally refines the model parameter estimates until it iteratively converges to the final value. This algorithm iteratively strives to find parameters which best fit the training data.

Subsequently, the identification or classification task is then performed on the set of speaker models. If there are S speaker models in the reference model database $\{\lambda_1, \lambda_2, \lambda_3..., \lambda_S\}$, the test speaker's feature vectors are compared to each model at a time. Finally, S scores are obtained from which the top scoring model is determined. This is done according to the maximum likelihood criterion. Let $\mathbf{X} =$

$\{\vec{x}_1, \vec{x}_2, ..., \vec{x}_F\}$ be the given feature set from a test utterance of a test speaker $\hat{s}$. The maximum posterior probability for X is then calculated as follows using Bayes' theorem [15]:

$$\hat{s} = \arg\max_{j=1}^{S} Pr(\lambda_j|\mathbf{X}) = \arg\max_{j=1}^{S} \frac{p(\mathbf{X}|\lambda_j)}{p(\mathbf{X})} Pr(\lambda_j). \tag{2.32}$$

This is done assuming that the prior speaker probabilities are equal. Then $\arg\max Pr(\lambda_j|\mathbf{X}) = \arg\max p(\mathbf{X}|\lambda_j)$ because $Pr(\lambda_j)$ is the same for all speaker models since all speakers are equally likely. The logarithmic likelihood L that the speaker is identified correctly can then be expressed as:

$$L(\mathbf{X}|\lambda_s) = \sum_{f=1}^{F} \log p(\vec{x}_f|\lambda_j) \tag{2.33}$$

where F is the total number of feature vectors (or frames) in the feature set of a test speaker. Finally, the decision rule for determining the identified speaker is

$$\hat{s} = \arg\max_{1 \leq j \leq S} L(\mathbf{X}|\lambda_j). \tag{2.34}$$

2.4 Limitations

Speaker identification systems are not popular in real life applications because of several factors that affect the identification rate. The feature sets derived from the speech signal for recognition purposes are generally not robust to both channel effects and acoustic noise. The confusion of enrolled speakers on a feature space increases as the population of speakers grows and therefore the SiD performance degrades. Speaker variability also affects the accuracy of SiD systems. Most SiD systems utilise commercial speech databases which have been acquired using different microphones or telephone handsets. Furthermore, the recording environment also introduce unwanted variations within the natural acoustic signal.

2.4.1 Noise

The most likely application for SiD is over the telephone network. This means that the telecommunication channel separates the user with the SiD system. However, channel noise is generally regarded as the most important factor that adversely affects the performance of SiD systems [1, 21, 46, 69]. Murthy et al [21] have looked further into the channel noise problem. They propose feature extraction methods which will try and overcome the noise. The training might be done using a close-talking microphone while the testing takes place over a telephone conversation. Scenarios of this nature leads to research on acoustic mismatch of training and testing data. Generally, feature extractors should have the capability of channel noise compensation as illustrated figure 2.13. Not surprisingly, there are many studies which address the noise in the context of speaker identification. However, only few are mentioned in this chapter. Grarcia and Mammone [22] use cepstral mean subtraction frequency warping for channel normalisation. On the other hand, Lo et al [24] perform channel compensation by removing channel mismatches using adaptive component weighting (ACW). Monte et al [23] address the noise by making use of self organising maps (SOM).

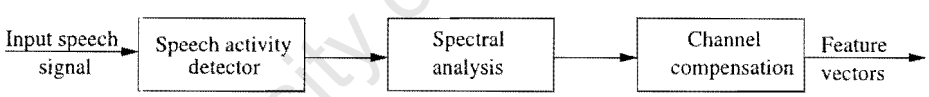

Figure 2.13: Channel compensation during feature extraction [15].

2.4.2 Large speaker populations

Large crowds of speakers engaged in different conversations can easily mislead the listener to identify the speaking person. By the same token, SiD systems are also subject to this limitation. In contrast, however, since the human auditory system is relatively robust to this phenomenon (the so-called cocktail party effect). The larger the number of speakers enrolling into the system, the more crowded the feature space becomes [20]. As a result, the difficulty of distinguishing speakers on the

feature space gets worse and hence the speaker misidentification error deteriorates. This observation is reported in several studies in the discipline of speaker identification [16, 17, 19]. Generally, SiD systems' performance decreases with speaker population. Some studies [18, 19, 64] (including this one) have investigated SiD performance as a function of population size. In addition, work has been done to make the SiD decision on a smaller set of speakers given a congested feature space [19, 28, 31]. The smaller set-based decision is identical to performing identification on a feature space with less and hopefully (more) separable classes.

2.4.3 Speaker variability

SiD systems could perform extremely well if each and every speaker could utter sentences in a very unique manner such that there was no similarity between his or her voice and other speakers' voices. Ramachandran et al [13] address differences in vocal tract anatomy and speaking style. The limitation is that SiD systems only exploit vocal tract anatomy and speaking style (e.g. accent) variation sources [13]. In addition, the identification rate is further reduced [33] when the enrolled speakers have the same age, accent and gender because the feature vectors occupy roughly the same region on the feature space. Campbell [1] highlights some of the inter- and intra-speaker variations that cause poor SiD performance. For instance, extreme emotional states such as rage and grief are some of the phenomena that alter a person's voice and thus SiD performance in general. Factors such as sickness, aging, and drunkenness also cause a speaker to sound different. Incidentally, some speech databases were recorded at different times (see table 2.1) to cater for investigations on speaker variability [10, 11, 12, 70].

2.4.4 Speech acquisition equipment

The choice of a microphone also affects the perception of speech and hence the performance of SiD systems. Thus some databases have numerous acquisition equipment (e.g. microphones or handsets). For instance, the so-called close-talking microphones eliminate channel noise and some undesirable acoustic effects from the recording stimuli. Certain studies were conducted in which the impact of a microphone on the performance of an SiD system were investigated [47].

2.4.5 Recording environment

State-of-the-art SiD systems have yield perfect performance if the database used was recorded using a close-talking microphone and the recording environment was quiet [16, 46, 63]. Investigations on how variable recording environments affect SiD performance also depend on the databases that allow for such a situation. Even the human ear struggles to identify a person if there is a lot of ambient noise. For example, it is not easy for two adjacent people watching a world cup soccer game to have a smooth conversation.

2.5 Possible applications

Possible applications of SiD include telephone banking , forensic applications as well as domestic applications [46, 69]. The entertainment industry can also benefit a great deal from the SiD technology. Interestingly, forensic investigations were some of the earliest applications of SiD [69]. In forensic applications, voice is used to link the identity of an unknown culprit with that of a suspect under interrogation. Furthermore, security applications such as access to buildings, documents, services and information in general are potential uses of SiD systems because the access can be made available only to authorised individuals [46]. In addition, telephone banking [21] could make use of a caller's voice to identify him or her in order for a particular transaction to take place. Telephone banking could also makes use of speaker verification for high level security [14]. Call centre technology [41] can be enhanced by making use of speaker identification especially when an agent needs the identity of a speaker. Text-dependent SiD is the suitable candidate for this application because the user could be prompted to say a particular text such as date of birth to ensure that he or she is the legitimate client. Finally, domestic applications that one could think of are those of commanding certain household electronic devices such as TV sets, radios and even lights. For instance, one could automatically switch the lights on and off over the telephone whenever one is away on holiday. This can be regarded as an extension of security applications.

As was mentioned previously, speaker identification is a difficult problem and hence commercial applications that exist today mainly utilise only speaker verification

[14]. However, SiD is sometimes used in conjunction with automatic speech recognition (ASR) for certain applications [20, 46]. The main disadvantage of SiD is that its performance degrades if speech is passed through the channel. The proposed features do not normally capture the desired speaker information which could be clearly identified by the application.

2.6 Summary

This chapter has introduced and described the structure of a speaker identification system. Some speech databases which are commonly used in speaker identification research were also discussed. In addition, detailed descriptions of the front-ends and back-ends of an SiD system were also provided. Examples of feature extraction and classification algorithms relevant to this work were also described. Moreover, some of the essential characteristics of feature extraction processes were highlighted. Factors the influence the performance of SiD systems (e.g. noise, speaker variability and enrollment population size) were also described. Finally, this chapter is concluded by describing some of the potential applications of SiD systems.

Chapter 3

Hierarchical speaker recognition systems

This chapter is a review of some previous work geared towards the improvement of recognition systems using hierarchical methods. It also illustrates different perspectives of hierarchical systems in various studies. The notion of hierarchy varies from one implementation to another depending on the *hierarchical element* of the recognition system. Some systems termed 'hierarchical' incorporate hybrid system designs at the front-end while others may utilise a combination of different classification methods. Section 3.1 highlights a few hierarchical systems that utilise signal processing-based methods with the aim of improving computation time and / or overall recognition accuracy. Section 3.2 briefly describes hybrid systems that have been referenced as motivation for the proposed N-best hierarchical system. Some of the work cited in section 3.2 does not necessarily reveal the term, 'hierarchical' but their implementation fits in with the contextual implication of this word in our proposal. Section 3.3 lists a variety of other hierarchical implementations in speech processing in order to illustrate that the proposed hierarchical methods are part of the diverse hierarchical paradigms in speech processing. A diagrammatic explanation of the group detection hierarchical system is illustrated in figure 3.1.

3.1 Front-end based hierarchical systems

Front-end orientated hierarchical systems refer to those that utilise signal processing methods as a way of grouping speaker models. Under these implementations recognition is performed on a smaller group of speakers instead of classifying a speaker from a large population in the model database. Using the same analogy, Abdulla and Kasabov [29] realised that their speech recognition system's performance was degraded by a speaker-attributed variability which was to do with gender. They therefore developed a gender identifier which split speakers into groups of male and female speakers. Their gender identifier first transforms the speech signal into the frequency domain using FFT. The pitch is then derived from the log magnitude spectrum. The value of the pitch frequency is compared with a threshold that differentiates female and male pitch. This process minimises the cross-sex confusion during the HMM classification. Our proposal in section 4.2 of chapter 4 duplicates this idea (see figure 3.1) even though our system is speaker identification.

On the other hand, Chengalvarayan [58] makes use of hierarchical sub-band linear predictive cepstral coefficients for improving the HMM based speech recognition. This approach splits the spectrum of the input signal into sub-bands. The LPCC are computed by performing the IDFT on the mel power spectrum of each sub-band. This method does not attempt to limit the number of speaker models on the feature space but the architecture used is hierarchical. The difference is that the sub-band groups are generated instead of model groups. Ellis [71] improves the performance of his speech recognition system by using the same idea as Chengalvarayan [58]. However, in these two systems classification is done on the whole set of trained data or speaker models whereas we aim to perform classification on smaller reference sets.

3.2 Back-end based hierarchical systems

Back-end based hierarchical systems refer to hybrid architectures that normally utilise uncorrelated modelling and decision-making techniques using two or more back-ends. Some studies [18, 19, 20, 31, 32] on hybrid implementations use the top N-best scores from the baseline classifier with an additional classifier to enhance

the recognition score. The main objective of making a decision on a small set of models is a characteristic of such systems' configurations. Most such systems use two back-ends. The performance improvement margin from the baseline to the N-best hierarchical system is not considerably large. The underlying problem is how to combine these classifiers. Other systems use a decision-making threshold which triggers the use of the second or alternative back-end.

Fine et al [31, 32] have carried out studies which use SVM to improve GMM scores for their SiD system. Their choice of SVM to enhance the GMM scores is prompted by the observation that SVM and GMM classification schemes exhibit uncorrelated errors at about the same performance level. In this work Fine *et al* report a decision-making threshold which chooses N-best scores produced by GMM and selects the ones with maximal scores. The SVM decision on the feature frames becomes final if the GMM classification decision is indecisive [31]. Under matched conditions, they first achieved relative error reductions of 25.7% and also 20.3% using two different types of microphones for training. The performance improvement was achieved on another similar system [32] by the same authors. That performance yielded a relative error reduction of 29.6% and 32.6% under similar conditions as in the first system. The population of 52 speakers from the Lincoln Lab Handset Database (LLHDB) was used for both these studies.

On the other hand, Ganchev et al [19] developed a system which uses GMM to improve their probabilistic neural networks (PNN) classifier on speaker recognition. A decision threshold is used to determine whether to use GMM or not. The N-best speakers from the PNN scores are further classified with GMM if the score is below or close to a threshold. Using a speaker population of 110, results reflect a reduction in SiD error from 3.44 to 1.38 when the system using this technique. This corresponds to a relative error reduction of 59.9%. The database used is Polycost corpus [10, 12].

Le Floch et al [47] combined GMM and autoregressive vector modelling (ARVM) in order to improve speaker recognition performance on telephone speech. They used 168 speakers from NTIMIT database. The scores from these two classifiers were normalised so that they can have one standard for comparison purposes. Their system had two modes, namely, competition and cooperation. For competition between GMM and ARVM, the best normalised measure was taken as a final decision using

certain weighting values which were empirically optimised. During cooperation the sum of normalised GMM and AVRM measures is calculated as the final decision. The final result reflects a GMM performance increase from 61.7 to 82.6. This approach differs from the one of Fine [31, 32] and Ganchev [19] because there are no N-best scores.

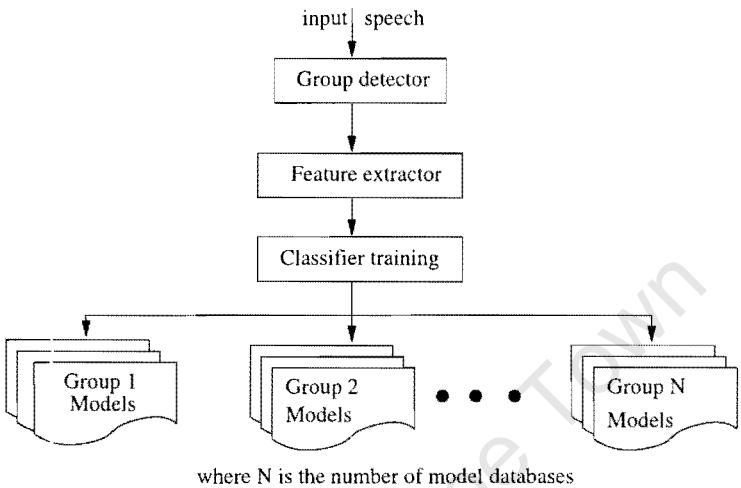

Figure 3.1: The notion of the proposed group detection hierarchical system.

3.3 Other hierarchical systems

Pan et al [28] implemented an on-line method for large population speaker identification. Their architecture is similar to the one proposed in this project except that their aim was to limit the processing time during both testing and training. Their system generates two databases of vector quantisation codebooks from which classification is performed. *Sivadas and Hermansky* [72] *configure their multi layer perceptron (MLP) hierarchically in order to replace the large monolithic MLPs with hierarchies of MLP experts in order to achieve robust GMM classification.* On the other hand, Beigi et al [27] utilise distances to build a hierarchical binary tree of models in order to minimise the computational time. In addition, Wu and Chen [33] aim to reduce identification time by implementing the three-level hierarchical speaker identification system using a so-called lateral inhibition Gaussian (LIG) network. Their work involves selecting the speaker candidate from feature sets forming

different clusters for each speaker. The second stage entails constructing the LIG network which then groups speakers into membership tables. The top N scoring speakers from the membership table are used for final identification. This method is similar to the N-best approach. There are many other hierarchical systems such as [73] that are designed with different objectives.

3.4 Summary

This chapter began with the description of systems which make use of front-ends for improving either speech or speaker recognition. Very little literature was referenced because most such systems do not attempt to reduce the number of speaker models in the feature space as is proposed in this project. Alternative methodologies in which a baseline system's score is further improved using another back-end which exhibits uncorrelated errors were also investigated. Finally, some studies which constitute hierarchical implementations were mentioned in the last section in order to emphasize that the hierarchical system is not standard but can be used in totally different implementations in speech processing. The implementation of the proposed hierarchical system based on some of the ideas highlighted in literature is presented in sections 3.1 and 3.2.

Chapter 4

The Hierarchical SiD System Design

This chapter describes the design of two types of hierarchical SiD systems. These are (a) the group detection hierarchical system and (b) the N-best list system. The group detection-based system is similar to Abdulla and Kasabov's speech recognition system [29] as explained in section 3.1. Although the front-end architecture is the same as in this reference, our study is on speaker identification. The N-best list-based system mainly borrows from the system proposed by Fine [31, 32] and Ganchev [19]. This chapter first explains the baseline speaker identification implementation so that its relationship with the proposed systems can be established.

4.1 The baseline system

The SiD used in this study extracts the parameterised feature sets (PFS) from a speech signal and uses them to construct Gaussian mixture models. Subsequently, GMM-based classification using maximum log-likelihood is used to classify the PFS from the test speaker. This is the framework on which the proposed hierarchical architectures are built. Figure 4.1 illustrates the architecture of the baseline SiD system.

As stated in Chapter 2, the parameterised feature sets are based on MFCC features. In this section the parameters used for our baseline system implementation will be mentioned. The back-end GMM number of mixtures is 32 and the design is exactly

as described in chapter 2.

4.1.1 Parameterised feature sets

The parameters of PFS are aimed at optimising useful characteristics of the mel-cepstral features. As mentioned in section 2.2.5 of chapter 2, the α and β parameters are used to fine-tune the performance of a speaker recognition system. Mashao and Baloyi [26] illustrated how SiD performance varies as a function of these parameters. This study assumes the optimal parameters to be $\alpha = 4.0$ and $\beta = 1.6$ because they have been found to yield a relatively high performance. Secondly, most preliminary experiments on clean speech and GSM speech were carried out using the same values for α and β [40]. It was found in several experiments [42, 63, 64] using the same system that 100 % speaker identification rate is achieved if clean speech is used. In addition, Mashao and Baloyi [26] investigated the effect of varying β and discovered that at $\beta = 1.7$ the comparatively high performance was obtained. This corresponded to using $\alpha = 4$. The feature dimension has been kept at 30 as in [16, 26, 64].

4.2 Group detection hierarchical SiD

The group detector-based hierarchical SiD system can be regarded as the most suitable way of solving the large population size problem in speaker identification ac-

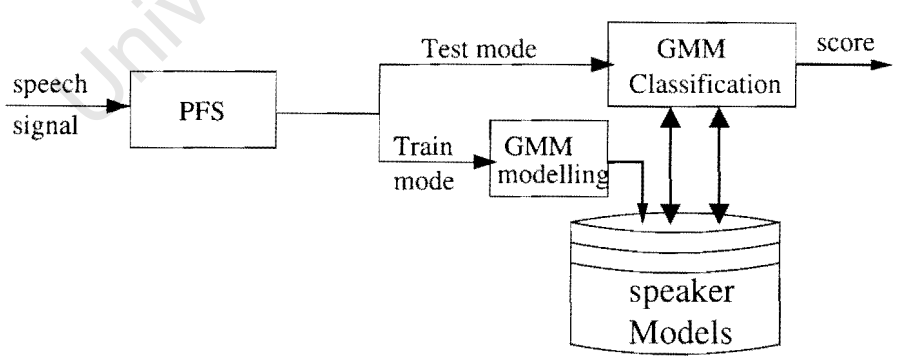

Figure 4.1: PFS -GMM speaker identification system

cording to the hypothesis described in chapter 1. As the hypothesis in figure 1.4 suggests, grouping the enrolled speakers into different smaller population sizes could yield better performance. The main question lies on the number of groups one can split the large population. The best way of dealing with this problem was to first start with two groups and find out how robust the proposed system was. The other initial step was to start with a smaller population of speakers for preliminary experiments. The proposed group detection hierarchical system architecture is illustrated in figure 4.2.

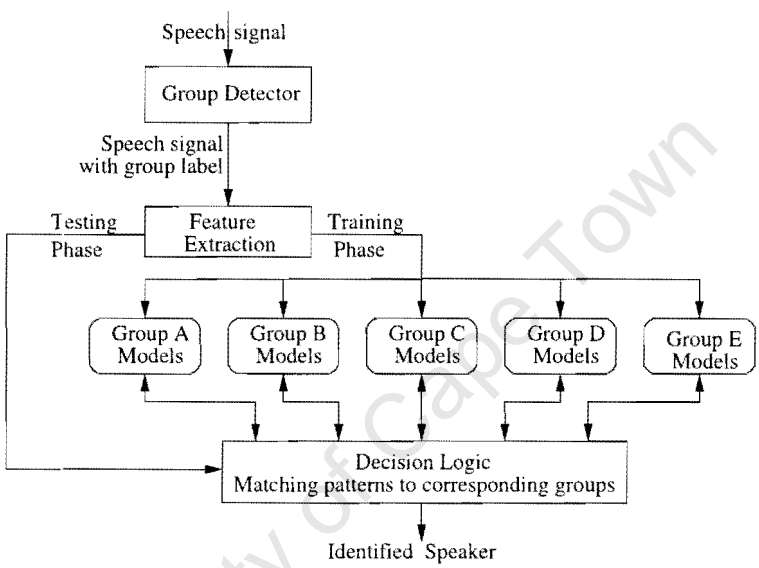

Figure 4.2: The Group detection hierarchical SiD system.

As illustrated in figure 4.2, the enrollment (training) phase of the system entails the labelling of the utterances so that the back-end can place the models in the correct group databases. During testing, speech from the test speaker is labelled by the group detector so that the classifier can compare the test speaker with a suitable pool of speakers in a corresponding database. This is advantageous because the feature space is now less congested.

The main challenge in this kind of implementation is the design of the group detector. The group detectors are all assumed to be perfect. However, it was later discovered that they do not consistently group speakers. Abdulla and Kasabov [29] used a gender identifier as their group detector and they found that it improved the

performance of their speech recognition system. There are several suggested characteristics that the group detector should maintain for robustness. The group detector must:

- be in such a way that group features and the baseline front-end features are uncorrelated.
- be text-independent.
- consistently group the speakers into their corresponding databases during testing.
- be a simple routine.

The information captured by the group detector should be different to the baseline features because speakers with similar voices are likely have a higher chance of being allocated to different groups. This is an advantage because similar voices could result in very similar model parameters that might cause confusion during testing. The consistent grouping of speaker models is necessary so that the test speaker's feature sets are not compared with the wrong set of models during testing. A simple algorithm that does this is desirable since the baseline SiD system is assumed to perform considerably well. The complex group detector might affect processing time and might even bring unnecessary degradations in the performance of the SiD system and therefore a simple one is preferable.

Several group detectors were implemented and tested. The first attempt was to group speakers according to gender and therefore a gender identifier was implemented. The second attempt was the use of a PLP-based group detector in the hope that PLP and PFS capture uncorrelated speaker information. Finally, it was realised that less confusion on the feature space might be achieved if the model databases could be separated according to dialect. This idea was motivated by the observations shown in Table 1 (appendix A) from preliminary experiments. The age group or vocal tract lengths could also be used to distinguish groups. Since the hierarchical approach has not been used much in SiD related studies, the author thought it was important to experiment with different group detectors in order to determine some of the main disadvantages of this setup.

4.2.1 The gender detector

The gender detector was implemented exactly as Abdulla and Kasabov [29] did. This includes windowing speech signal into 20ms frames and computing the cepstrum from which the peak value was determined. The peak value is either rejected or accepted depending on how big it is for the determination of the voiced speech frame. If the frame is voiced the the median filter is used for smoothing. The peak location is then computed and the pitch is calculated using the inverse of the location time. Finally the average pitch, $ave\, F_0$, is obtained using equation 4.1. That is,

$$ave\, F_0 = \frac{1}{N_v} \sum_{i=1}^{N_v} F_0(i) \tag{4.1}$$

where $F_o(i)$ is the fundamental frequency of the i^{th}frame and $N_v$is the total number of voiced speech frames. According to Abdulla and Kasabov, if $ave\, F_0 < 160Hz$ then the speaker is classified as male, otherwise the speaker is assumed female. The accuracy of this gender detector was measured and results are recorded in table 5.1. The output of this detector is a group label showing either male or female.

4.2.2 The PLP-based group detector

The PLP-based group detector uses a slightly different approach. The PLP was implemented according to Hermansky [61] except that the optimisation issues were not taken into consideration because the investigation was still at the initial stages. Figure 2.11 and section 2.2.6 in chapter 2 describe how this algorithm works. The resulting spectrum after the IDFT (see figure 2.11) is used to determine the groups. Samples 20 to 100 are used as a grouping segment. Peaks from sample 20 to 55 are added and the sum is stored for Group A identity. Group B is identified by the sum of peaks from samples 55 to 90 just to keep symmetry. The choice of these numbers is experimental. These sums are compared and if sum A is greater than sum B then the speaker belongs to group A and vice versa. The output is the group label showing A or B.

4.2.3 Dialect group detector

The dialect group detector was not fully designed because it was found that the dialect or accent identifiers do not yield high performances [2, 3, 4, 5, 6]. However some tests were carried out in order to find out by how much the baseline system could be improved if a dialect group detector were to work perfectly. The second attempt was to make use of the already known dialect labels in NTIMIT by using the baseline SiD as a dialect identifier. The GMM classifier is used to build the dialect models from the collective contribution of the speakers from each dialect region (eg. dr1) For example, speech from dr1 speakers is used to create model dr1. During identification, the test utterance from a speaker is compared with eight dialect models using the maximum likelihood criterion. Finally, the knowledge of dialect regions from NTIMIT database makes it possible to *force* the SiD system to compare the test speaker to only those who are from a dialect region during testing. This is done in order to ensure accuracy of the dialect group detector and also to investigate the margin of improvement if the number of model groups increases to a number larger than two.

4.2.4 The classification module

The GMM described in chapter 2 remains the same except that the labelling of parameters varies according to the model group in which the speaker's features belong. This means that the GMM for the group detection hierarchical system will represented by :

$$\lambda_A = \{c_{Ai}, \vec{\mu}_{Ai}, \Sigma_{Ai}\} \quad i = 1, 2, ...M \tag{4.2}$$

where λ_A is the model belonging to group A and M is the number of mixtures. M is the group label mentioned in section 4.2.2 above. The feature sets for group A are therefore $X_A = \{\vec{x}_{A1}, \vec{x}_{A2}, ..., \vec{x}_{AF}\}$ so that the maximum log likelihood is based on the scores of group A database using:

$$\hat{s}_A = \arg\max_{j=1}^{S} \sum_{f=1}^{F} \log p(\vec{x}_{Af}|\lambda_{Aj}). \tag{4.3}$$

4.3 N-best hierarchical SiD

The implementation of the N-best hierarchical SiD is as described in section 3.2 of Chapter 3. The ideal N-best list method should be in such away that the final decision is made from *N* speakers, where N is much less than the total number of enrolled speakers. Fine et al [31, 32] reported that certain thresholds were used to trigger the second scoring routine for the N-best identification process. The proposed method's training phase is similar to that of baseline SiD.

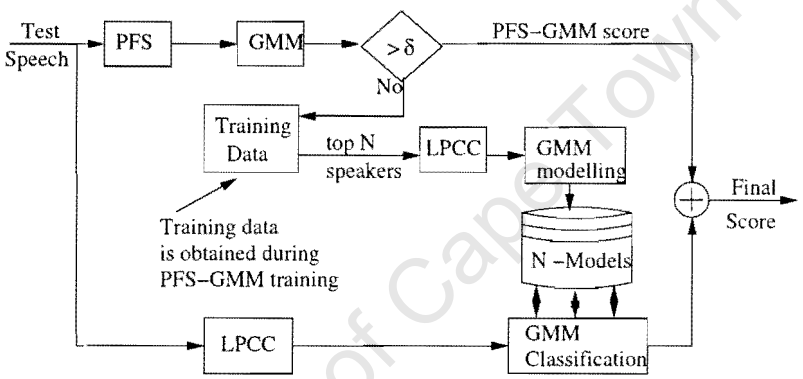

Figure 4.3: The test mode of SiD with LPCC based N-best scores.

The proposed architecture is similar to the SiD system of Fine [31, 32] and Ganchev [19] in that, they both perform recognition on the N-best list of speakers. It however differs from theirs because the N-best speakers are classified by using a new feature extractor instead of another back-end. Figure 4.3 shows the implementation of the proposed method during testing because training is similar to that of the baseline SiD system (PFS-GMM). A different feature extraction unit is proposed to retrain N-best scoring speakers using data acquired from the baseline training. Linear prediction cepstral coefficients (LPCC) are utilised for this purpose. The linear prediction (LP) order has been kept as 30 because the baseline system's feature vectors are also 30 dimensional. This similar dimensionality of feature sets makes it less complex to integrate the LPCC module into the baseline SiD system. The performance comparison is also more rational if the feature dimensions are equal. The LP order of

30 is high compared to most orders used in literature [1, 36]. This is not a problem because higher LP orders produce better performance [61].

The test speaker's utterance is now tested and a soft likelihood decision is made. This is because top N speakers form a very small set of models compared to the whole set of 630 models. Most researchers [18, 19, 31, 47] indicate that the correct identification by the baseline system should not be compromised. This study also tries to find ways of keeping the PFS-GMM (baseline) score as high as possible by making use of a threshold value δ. This threshold determines whether the PFS-GMM score should be final or if the system should further apply the LPCC-GMM step for final identification as demonstrated in figure 4.3. The motive behind the use of a threshold is that if the top 2 speakers are close to one another the confidence of the GMM classifier is low and therefore a second "opinion" from another module is necessary. This is why the enhancement happens for scores less than δ. Scores less than the threshold prompt the retraining of top N speakers' utterances while the rest of the scores are kept as final.

4.4 Summary

This chapter has illustrated different implementations of group detectors. These detectors make it possible for classification to be done on the limited number of speaker models and this results in the minimal crowding of the feature space. The problem with this method could be the correlation of features extracted by the group detector and PFS. The other drawback could be the robustness of group detection under text-independent conditions. The N-best hierarchical SiD is implemented using the LPCC (frond-end) as the score enhancing tool. It has been indicated that LPCC feature sets are used to enhance the PFS-GMM scores as opposed to most studies in which different back-ends are employed to improve performance. The results and the related performance issues are dealt with in chapter 5.

Chapter 5

The SiD Experiments and Results

This chapter is gives a detailed account of several experiments that were carried out in order to improve the baseline system's performance using the proposed hierarchical methods. Section 5.1 reports the results that were obtained when the group detection hierarchical system was tested using several group detectors. The results reported in this section simply serve to illustrate that the group detection algorithm could achieve great recognition performance provided it worked perfectly. The design of a group detector that performs accurately was not followed in detail but rather referred for future work.

Section 5.2 reports evaluations of the N-best hierarchical method. These results meet the objective of improving the overall speaker identification rate of the baseline system. The final results are those that appear at the end of each section of hierarchical methods. These results indicate the performance of the proposed SiD on 630 people. Finally, conclusions are made from these results.

The NTIMIT database has been consistently used in all experiments reported in this chapter. This was done in order to compare this study with previous similar SiD systems [16, 26]. The ten utterances for each speaker are referred to as sentences 0 to 9 throughout this chapter. Sentences 0 and 1 are the same for all speakers. The remaining 8 sentences are unique for each speaker.

Virtually all results are compared to the performance of the baseline SiD system (i.e. the PFS-GMM system). The signal processing of the baseline system starts by segmenting the speech signal every 10ms using a Hamming window of 20 ms. A

voice activity detection (VAD) is implemented by measuring signal energies. The VAD simply discards noise or unvoiced frames. The feature vector dimension has been kept at 30. The LPC order was also kept at 30 in order to match the PFS feature vector dimension. Several studies [19, 49, 47] discovered that a GMM classifier requires much training data for it to yield optimal performance. A little experiment was carried out in order to check the validity of their observation and the corresponding results (see Figure 5.1) were obtained by testing the baseline system on 38 speakers from dialect region 1 of NTIMIT. These results indicate that for GMM performance increases with training data.

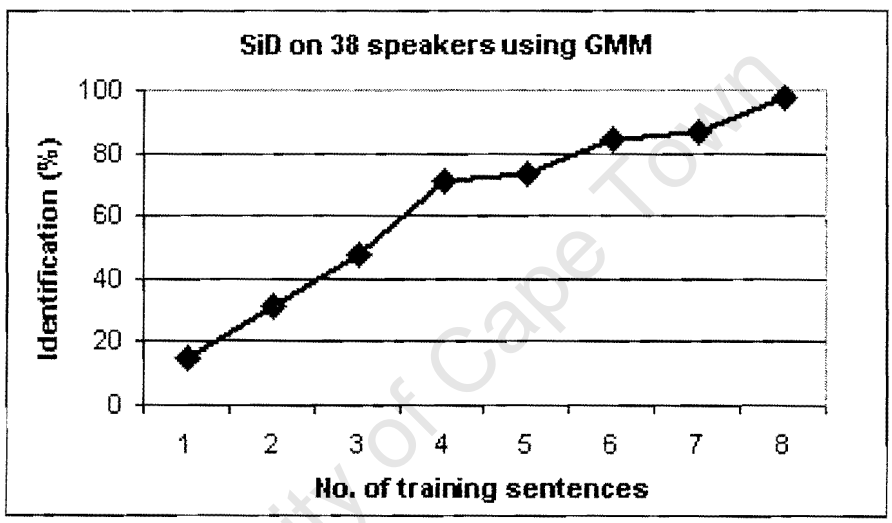

Figure 5.1: Training data influence on GMM classifier.

5.1 Group detection hierarchical SiD results

The group detection experiments are aimed at solving the problem of low identification accuracy due to a congested feature space. Group detectors were implemented and tested as described in Chapter 4. The first group detector designed was gender-based. This gender detection procedure led to the use of two groups of speaker models in all the experiments.

5.1.1 Gender detection

The gender detector was implemented as described in chapter 4. The gender detector test was the first to be implemented. In this test the whole NTIMIT database for all 10 utterances was utilised. Correct gender detection occurs when all 10 utterances are identified as coming from the same gender. If one of the 10 sentences resulted in a wrong gender of the speaker, the gender detector is considered to have failed. The performance of the detector was measured by dividing the number of completely successful gender detections with the total number of speakers in each dialect region. Table 5.1 shows the percentages of successful gender identification.

Dialect Region	Population Size	Correct Gender Id.	Gender iD (%)
dr1	49	41	83.7
dr2	102	83	81.4
dr3	102	78	76.5
dr4	100	78	78.0
dr5	98	83	84.7
dr6	46	38	82.6
dr7	100	85	85.0
dr8	33	28	84.8

Table 5.1: Performance of the gender classifier.

When the whole database was used, it was found that 81.8% of the time, the gender identifier fully classified the speaker according to gender under text-independent conditions. Table 5.1 shows that the identifier performs well enough considering that all utterances were used for testing. The gender identifier was then used to split the given speaker population into two model groups of male and female. The results are shown in figure 5.2. In this figure the population was divided into two equal gender groups. This was done in order for the system to get unbiased gender identification results. The first two utterances were used for testing while the remaining 8 were used for training.

The first two sentences of NTIMIT are the same for all speakers . It can be deduced from figure 5.2 that improvement occurs at speaker populations of 74 and 144. Since the SiD system under investigation is text-independent, a small pop-

ulation of 50 speakers (25 male and 25 female) was enrolled into the system and different test utterances were utilised for training as depicted in figure 5.3. The

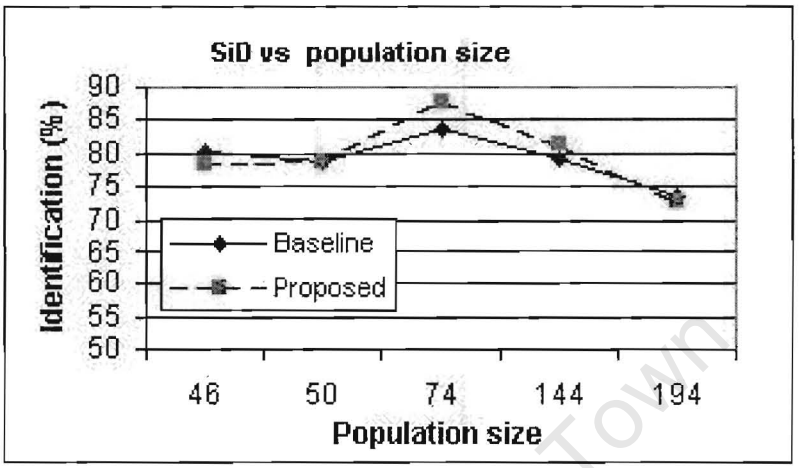

Figure 5.2: Gender identifier based SiD as a function of population size.

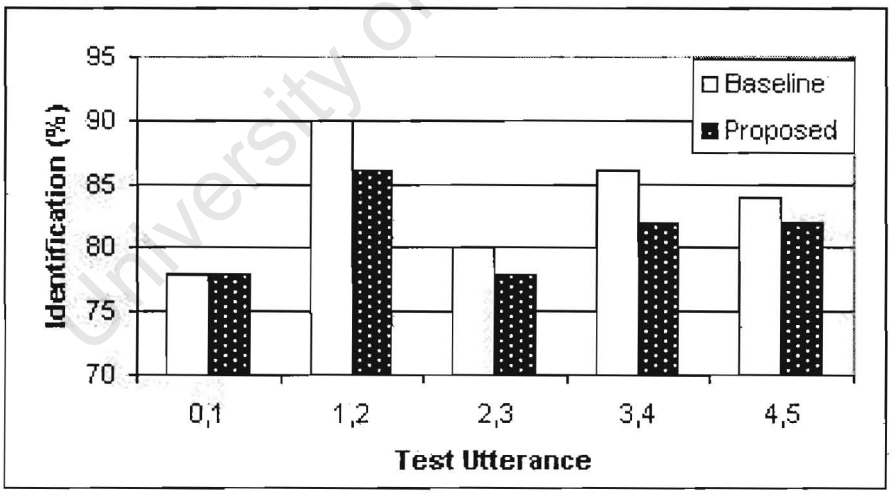

Figure 5.3: Gender hierarchical SiD using different test utterances on 50 speaker population.

purely text-independent gender based hierarchical SiD reflects no robustness as portrayed in Figure 5.3. Perhaps it might have been helpful to investigate the gender detector using a text-dependent recognition mode. However, the main limitation is that the NTIMIT database has no text-dependent capability [10]. A speaker does not produce the same acoustic signal if he or she speaks the same sentence at different times and thus the idea of text-dependent SiD. The same uttered text recorded at different times could have been utilised had there been a database that consists of text-dependent utterances.

It was assumed that a way to see the effectiveness of proposed group detection algorithm was to use *one of the ten sentences uttered by each speaker for during both training and testing* . This is rather an unrealistic way of dealing with real life problems because the group detection utterance is exactly the same signal during both training and testing. Using the same signal **guarantees** the existence of a system in which a test speaker will always be compared to the correct group of models and thus implementing a *perfect* group detector. The speakers were tested according to their dialect to investigate how the performance will change. Utterances 8 and 9 were used for testing while the remaining 7 sentences were used for training. In this experiment, the second utterance was used for grouping in both testing and training phases. Figure 5.4 shows the results obtained from this test. The population of speakers from different dialect regions (dr) used to obtain the results in figure 5.4 are shown in Table 5.2. The results in Figure 5.4 show performance up to dr6 because there was no improvement as the experiment progressed.

Dialect Region (dr)	dr1	dr2	dr3	dr4	dr5	dr6
Number of Speakers	38	76	76	68	70	35

Table 5.2: Number of speakers from different dialect region used for figure 5.4.

The results in Figure 5.4 indicate that the test of gender detector for higher population of speakers would not be fruitful. The reason that the perfect gender detector hierarchical SiD did not perform well is because of the ability of the baseline system's inherent ability to distinguish between the genders. The gender discriminability was tested and the confusion matrix was obtained. This is reported in appendix A and illustrates that 99.4% of the time the baseline SiD system is able to differentiate between male and the female speakers.

All NTIMIT female and male speaker's names start with an "f" and "m" respectively. This made it easier to measure the gender discrimination capability of the baseline system. It was then realised that the group detector that does not contain the information that the system already "knows" should be employed. The perceptual linear predication(PLP) analysis was then used as described in section 5.1.2 below.

5.1.2 PLP-based group detector

An investigation of the configuration that would best improve the overall recognition performance was continued by keeping several grouping sentences constant and testing with certain pairs of sentences. Figure 5.5 shows the average performance of a baseline SiD compared to the hierarchical system when the PLP-based *perfect* group detector is used. The *real* PLP-based group detection task was discontinued when it was realised that it yielded very poor performance during the initial tests and hence the corresponding results are not reported in this thesis.

The results in figure 5.5 signify that if a perfect group detector that splits the model

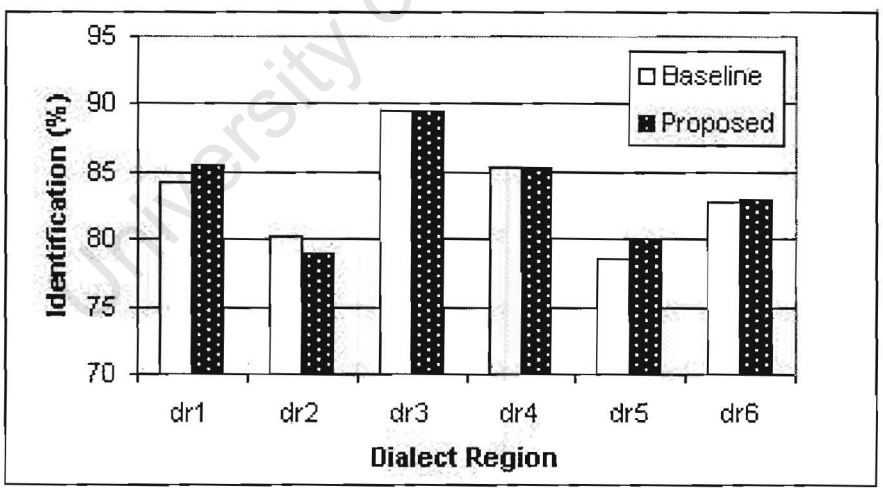

Figure 5.4: consistent gender detector hierarchical SiD performance.

databases into at least two groups were implemented, the SiD system's performance could improve considerably. The perfect hierarchical system yielded some improvement according to Figure 5.5. The PLP-based group detector's SiD performance of 75.7% was obtained when a population of 630 speakers had enrolled into the system (see Table 5.3). This result triggers the need to investigate what would happen if more than two groups existed. The use of 8 NTIMIT dialects was thus realised as the first step of this investigation.

5.1.3 Dialect detection

The 8 dialect regions of the NTIMIT database were assumed to be the eight model groups. During SiD testing the system was *forced* to compare the test speaker's features to the models which belong to his or her dialect region. This was also done using the knowledge of the dialect region names and again assuming a perfect group detection scenario. This test was done in order to investigate whether identification accuracy would increase by a large margin if the number of groups was to be increased to 8. It was found that 86.5% of the time the speaker was matched to the correct model.

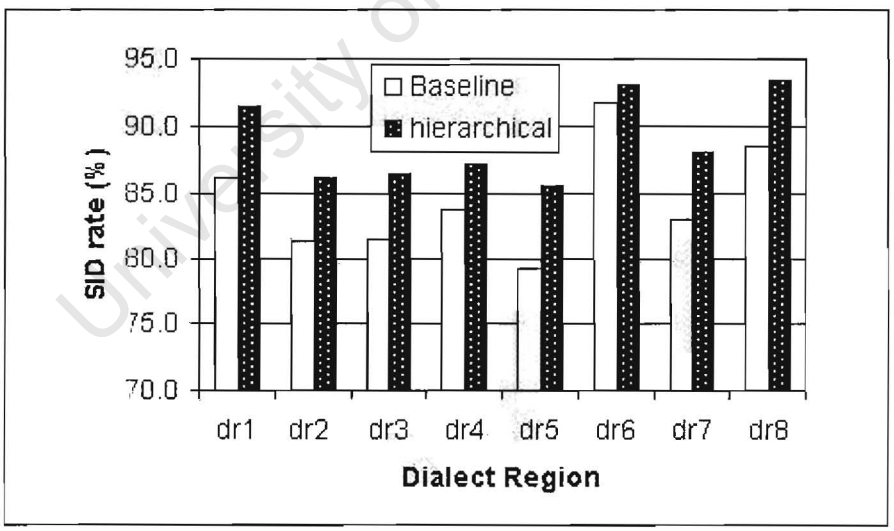

Figure 5.5: Perfect PLP SiD rates from all NTIMIT dialect regions obtained from Table 2 in Appendix A.

The second approach was the implementation of the dialect group detector itself in order to find out the dialect discrimination capability of the baseline system. This was done by implementing the dialect detector which is similar to the baseline PFS-GMM system whereby speech generated by all the speakers in a dialect region is used to create the corresponding dialect model. The GMM model for each dialect region was thus computed. This means that the total number of dialect models was 8. Finally, each speaker from each dialect region was tested such that the correct identification was when the GMM classifier identifies him or her as belonging to the correct dialect region. Figure 5.6 depicts the outcome of different test scenarios in which different pairs of test sentences were used. Sentence 0 and 1 were used for training in all cases (Figure 5.6) because they are the same for all the speakers. It was therefore logical to assume that the difference between the speakers' voices was due to dialect if they all utter the same sentence.

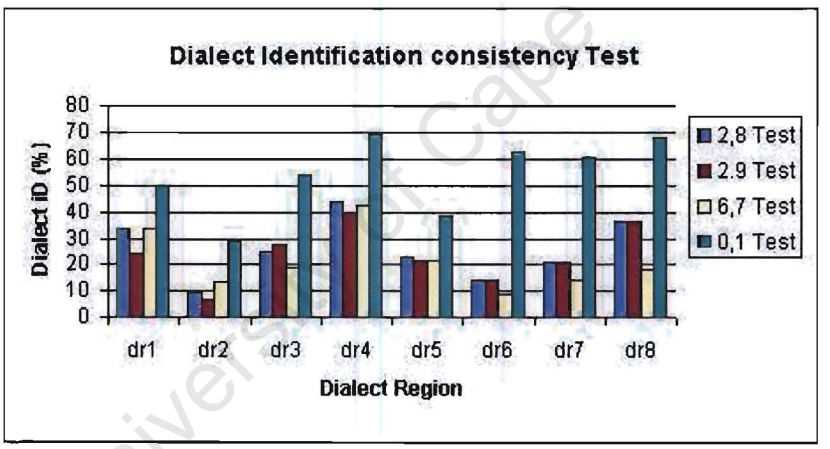

Figure 5.6: GMM dialect identifier using utterances 0,1 to train and {2,8; 2,9; 6,7; and 0,1) to test.

The results in Figure 5.6 demonstrate GMM's inability to consistently classify speakers according to dialect. The 0,1 (cyan) bar graph was obtained when sentences 0 and 1 were used for both training and testing. These results could have been better if the rest of the graphs were correlated. The non-uniform nature of graph heights in Figure 5.6 represents a vote of no confidence for the utilisation of a GMM-based dialect identifier as a group detector. Accent and dialect identification studies [4, 5, 6] also show low performance.

5.1.4 General performance of group detection hierarchical systems

Group detection hierarchical methods could be valuable for the improvement of speaker identification accuracy if more research were done to achieve nearly perfect group detectors. If the GMM model database of 630 speakers is split into at least into two groups, performance improves as seen in Figure 5.5 and also from the ideal dialect group detector which yielded 86.5% identification accuracy. Table 5.3 shows the SiD performance when sentences 8 and 9 were used to test the hierarchical system compared with the baseline system using two different group detectors. The rest of the sentences were used for training as in [16].

SiD architecture	SiD rate (%)
Baseline	69.5
PLP Group detector	75.7
Ideal dialect group detector SiD	86.5

Table 5.3: Summary of perfect group detectors' performance on 630 speakers.

Although Table 5.3 indicates improvements of 6.2% and 17.0% for PLP and ideal dialect detectors respectively, perfect group detection is assumed. These results may be indicating that an increase in the number of groups could cause a greater improvement on the SiD system's performance as a whole. The group detectors proved to lack robustness when different utterances are used. The results shown in Table 5.3 were obtained by running the process 3 times because ideal conditions prevail in the whole identification process. It was therefore decided that more statistical analysis was unnecessary. This is because the reported group detection SiD performance mainly results from ideal conditions. However, it illustrates a lot of desirable outputs if more research emphasis is put on perfecting the group detectors. Although the group detection hierarchical methods display the capability of solving the large population problem, they are not robust under text-independent conditions and therefore not sufficient for meeting the key objectives of this study. Another hierarchical configuration of the SiD implementation based on the N-best list of scores was tested which became a better method in terms of real system performance.

5.2 N-best list hierarchical SiD results

In Chapter 4 a description of the N-best hierarchical method was given. This section highlights the tests and the corresponding results. The LPCC-GMM enhanced results form the main finding of this work since the previous section reflects idealised situations.

5.2.1 LPCC-enhanced SiD results

In Chapter 2, characteristics of LPCC and PFS features were explained. The main difference between PFS and LPCC highlighted in chapter 4 motivates the use of LPCC feature sets to enhance the PFS-GMM (i.e. baseline) system's scores. The differences suggest the possibility that the errors due these front-ends are uncorrelated as is the requirement [19, 31, 32, 71] for most hybrid SiD system implementations. The LPCC-GMM block shown in figure 4.3 should be able to complement the PFS-GMM decision so that corrections can be made wherever PFS-induced misclassification occur. The threshold at which the LPCC-GMM architecture is triggered should also be determined as indicated in various implementations of hybrid systems [19, 31, 32]. Finally, the main results are based on the use of test utterances 8 and 9, which are also used is similar studies [16, 26] on SiD which utilise the whole NTIMIT database.

Correlation test

The first task in the N-best list experiment was to determine the extent to which the LPCC-related errors correlate with PFS-related errors. This test was done by performing identification on 100 speakers where PFS-GMM likelihood scores placed a test speaker either in position 0, 1, 2, 3,.. or 100. Since the winner is considered to be the speaker in position zero, it is necessary to experimentally verify how true that is by checking the speaker's label (name).

The correlation measurement is done by selecting all the test speakers that PFS-GMM ranks position 1. These selected speakers match the test speakers for this particular test The corresponding position 0 speakers are then trained together with

the position 1 speakers using LPCC-GMM as illustrated in Figure 4.3. If the PFS errors are correlated with LPCC then the position 1 speakers will always be in position 1 whereas if the speaker whose rank was 1 moves to position 0 after the being tested with the LPCC-GMM architecture then LPCC complements PFS [30]. The degree at which LPCC complements PFS is in fact the correlation test. If according to PFS-GMM Mr. B is a test speaker and scores position 1 with Mr. B model, then there is no inter-speaker confusion. Assuming that he scores position 0 with Mr. A's model. This means the system "thinks" Mr. A is the test speaker (wrong decision). In this case Mr. B and Mr. A's training data is used for LPCC-GMM training. If Mr. B (test speaker) now obtains the highest likelihood (position 0) then the LPCC have complemented the PFS-GMM decision which implies that PFS and LPCC have uncorrelated errors. Table 5.4 shows how far LPCC-GMM decision complemented the PFS-GMM decisions using a population of 100 speakers.

Test Utterance	PFS - No. of Speakers in Position 1	LPCC- No. of Speakers changed to position 0
0 , 1	11	7
2 , 3	5	4
4 , 5	6	4
6 , 7	9	7
8 , 9	8	5

Table 5.4: PFS - LPCC correlation test.

Table 5.4 shows that the PFS and the LPCC feature sets produce uncorrelated errors when used with GMM as the back-end. Row 2 in Table 5.4 shows that 11 test speakers were not successfully identified by the PFS-GMM system but instead they were thought to be the second highest (position 1) speakers. After the LPCC-GMM processing it is observed that 7 out of 11 speakers are now correctly identified. This complementing process therefore gives improvement on PFS-GMM performance. This finding makes it possible for the proposed N-best list hierarchical SiD to be tested using the LPCC-GMM module. These uncorrelated errors of the two front-ends are in line with what studies [19, 32] in N-best list methods suggest. It is observed from Table 5.4 that the LPCC does not fully *reverse* the PFS errors. The implication of this is that the reliable decision of when to complement the PFS-GMM misclassification errors has to be made. Further tests of determination of

decision thresholds were therefore carried out as elaborated in the next paragraph.

Choice of threshold (δ)

It was highlighted in Chapter 3 that choosing the decision threshold is an important step [32] in determining when to enhance the baseline score. The threshold, δ, used in this system was experimentally obtained in order to find the optimal SiD performance as explained in chapter 4. The threshold in this case is the difference between the likelihood scores of a speaker who scores the highest and the second highest scorer. This means that $\delta = L_0 - L_1$, where L is the log-likelihood score. If the difference is large it means that PFS-GMM decision is very confident. The initial N-best scoring tests were performed on small populations at $\delta = 50$. The first experiment was performed on dialect region 2 (dr2) of NTIMIT database whose speaker population is 102. There was no particular reason for choosing dr2 except for the population that was neither too small nor too large compared to a population of 630. It was also decided to keep N=2 for the N-best list of scores for initial experiments. Figure 5.7 shows how the 2-best list hierarchical system performs at $\delta = 50$.

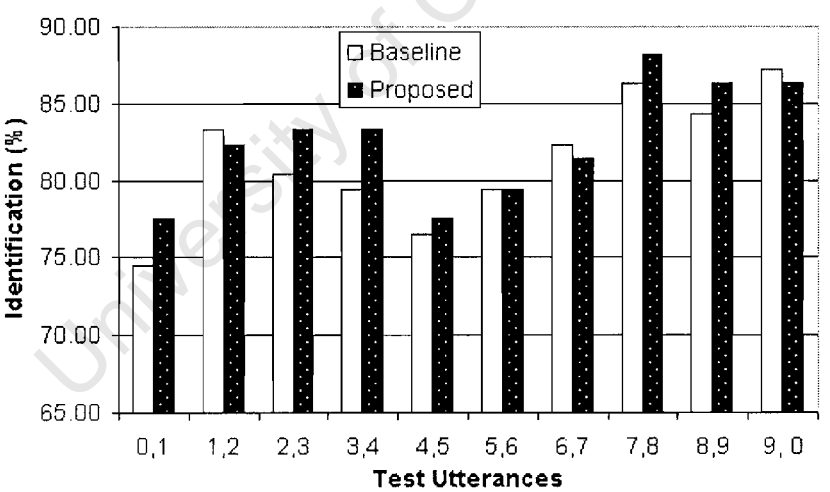

Figure 5.7: Proposed 2-Best SiD performance compared with baseline SiD for 102 speakers.

Figure 5.7 shows that 7 out of 10 times the LPCC-GMM module improves the base-

line performance. Test sentence pairs 1 and 2, 6 and 7 as well as 0 and 9 are negatively affected by the new system. This lack of robustness may be due to the choice of threshold or any another issues to be discovered. Chaudhari et al [18] suggest that the hybrid system should have an insignificant negative effect on the baseline score. It was therefore necessary to find the best threshold for larger populations. Keeping the testing data constant (sentences 8 and 9), the value of δ was varied and identification results are presented in Table 5.5. From these results $\delta = 30$ was found to be the optimum threshold for a population of 630 speakers. This threshold is then used for the remaining tests. The closeness of the SiD performances recorded at different thresholds in Table 5.5 shows that the proposed architecture in Figure 4.3 does not significantly deteriorate the baseline system's performance as suggested in [32].

Test Sentences	Threshold (δ)	Identification rate (%)
8 , 9	10	70.0
8 , 9	20	70.6
8 , 9	30	71.6
8 , 9	40	71.3
8 , 9	50	70.3
8 , 9	60	71.2
8 , 9	70	71.1
8 , 9	80	70.0
8 , 9	90	68.7
8 , 9	100	68.4

Table 5.5: Performance of proposed SiD at different thresholds on 630 speakers population.

Large population SiD performance

The large population speaker identification results are the key part of the findings since the main objective of this work was to improve large population SiD performance. All 630 speakers were enrolled into the SiD system for all tests in this section.

The first tests using $\delta = 30$ were performed by investigating how the system performs as the N-best list increases. This system reveals consistent improvement when N=2 according to table 5.6. When N=3 the SiD performance seems to be mostly

below the corresponding to N=2. These N=3 results show performance decreases with respect to the baseline system when sentences 0 and 1 are used for testing. It was therefore decided that the best 2 PFS-GMM scores (N=2) should be utilised for further LPCC-GMM processing since the corresponding performance reflects a consistent improvement.

Test utterance	Baseline SiD rate	Proposed SiD rate (N=2)	Proposed SiD rate (N=3)
0 , 1	64.8	65.2	64.3
2 , 3	64.6	65.9	65.6
4 , 5	66.8	67.8	68.1
6, 7	71.0	72.7	73.3
8 , 9	67.4	71.6	70.3
Average	67.4	68.6	68.3

Table 5.6: SiD rate (%) using N-best list.

The same experiment as the one which yielded table 5.6 was performed for ten pairs of test sentences and the results are shown in Figure 5.8. This was done in order to evaluate the robustness of the proposed SiD system under text-independent conditions. The identification rates shown in Figure 5.8 were obtained by using 2 sentences for testing and the other 8 NTIMIT utterances for training. From figure 5.8, the proposed 2-best hierarchical system seems to be fairly robust except when sentences 0 and 9 were used. The average baseline SiD rate is 67.2% and the proposed system produced 68.5% from the source data for Figure 5.8. The standard deviations for the baseline system and the proposed system are 2.65 and 3.04 respectively (See Appendix A table 4).

It was also necessary to explore how the proposed N-best hierarchical SiD system compares with the previous findings. The average performance of about 59.7% was obtained when the LPCC were used as the front-end, in replacement of the PFS as recorded in Table 5.7. This performance is lower than what has been achieved by most SiD systems (See Table 5.11) and therefore the LPCC-GMM architecture is not considered as a baseline system in this investigation. It is used instead as an

enhancement module for the baseline SiD. It is observed from Table 5.7 that the proposed method improves the PFS-GMM baseline results on the whole. The standard deviation of the proposed system is lower than the baseline which indicates a reliable improvement. The proposed N-best list retraining can only improve the system performance up to a certain level because if the test speaker's model is not among the top N, then misclassification is guaranteed. The author therefore investigated how best the system could have performed if the LPCC-GMM moved all speakers' second highest scores to best score. The fifth column of Table 5.7 shows the SiD performances obtained when this occurs. This means at N=2 the proposed SiD cannot perform better than 77.6% on average.

Similar systems such as those of Baloyi [64] and Reynolds [16] show a plot of the SiD rate as a function of population size using the NTIMIT database. Other researchers [18, 19, 21] also plotted the SiD rate as a function of population size. The best solution should reverse or minimise misclassification error. It was therefore necessary to investigate what characteristic and trend the proposed system would provide. This experiment was performed and Figure 5.9 shows the results. These results are averages obtained from 5 trials as indicated in Table 5, appendix A. Only

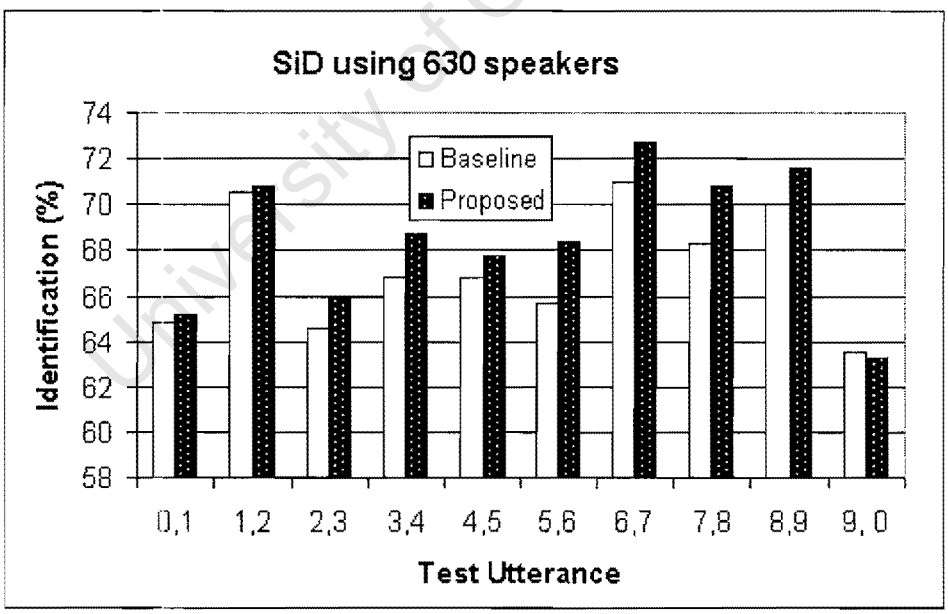

Figure 5.8: Performance of the SiD using different training data.

five trials were used for each test because it is already apparent from Table 5.7 that the N-best SiD system yields an improvement improvement in performance. Once again sentences 8 and 9 were used for testing while 0 to 7 were used for training phase. Population sizes were varied from 10, 100, 200, 300, 400, 500, 600 and 630 as in[16, 64] and were used to observe how identification error varied as a function of population size.

Figure 5.9 shows that the *(N=2)-best* scenario does improve the identification accuracy as the population grows. The performance degradation from a proposed system displays the relatively uniform trend with respect to the baseline. This is because the back-end is GMM for both LPCC and PFS. Furthermore, the classification, clustering and decision criteria are the same. The standard deviations were calculated for each SiD rate and Figure 5.10 shows the corresponding results. It can be observed from Figure 5.10 that the standard deviation of the proposed system is generally lower than that of the baseline SiD system. When the population size was 10, the identification rate is always 100% for both the proposed and baseline SiD systems which indicates that SiD performance is high on small populations.

No. of Trials	LPCC-GMM SiD rate	PFS-GMM baseline SiD rate (%)	Proposed 2-best SiD rate(%)	Maximum 2-best SiD rate(%)
1	60.3	70.0	71.6	77.9
2	59.2	70.5	71.1	77.6
3	60.0	69.8	71.7	77.0
4	60.5	68.3	70.8	77.8
5	61.1	69.0	70.5	78.6
6	59.8	68.6	70.8	78.1
7	58.7	69.8	70.2	77.5
8	59.7	69.5	71.7	77.3
9	58.6	69.5	71.3	77.3
10	58.6	68.6	71.0	76.8
Average	59.7	69.4	71.1	77.6
Standard Deviation	0.863	0.711	0.512	0.528

Table 5.7: Average SiD rate using sentence 8 and 9 to test.

The N-best SiD systems uses the relative error reduction (RER) criterion [32] to find out the degree by which the proposed system reduces the baseline identification error. The identification error is given by, $100\% - SiD\,rate$. It has been seen in Chapter 3 how some of the N-best list systems make use of this way of computing the RER. Equation 5.1 shows how to calculate the percentage RER. That is,

$$\%RER = \frac{Baseline\,SiD\,error - Proposed\,SiD\,error}{Baseline\,SiD\,error} \times 100\%. \qquad (5.1)$$

The RER was calculated for this study using the same data as that of Figure 5.9. Results in Figure 5.11 show that the new system relatively improves the identification rate by a larger margin at populations of 100 and 200 speakers. The first RER value of zero should strictly be *undefined* (division by zero) according to equation 5.1, but it has been included as one of the points on the RER graph to indicate *no need for*

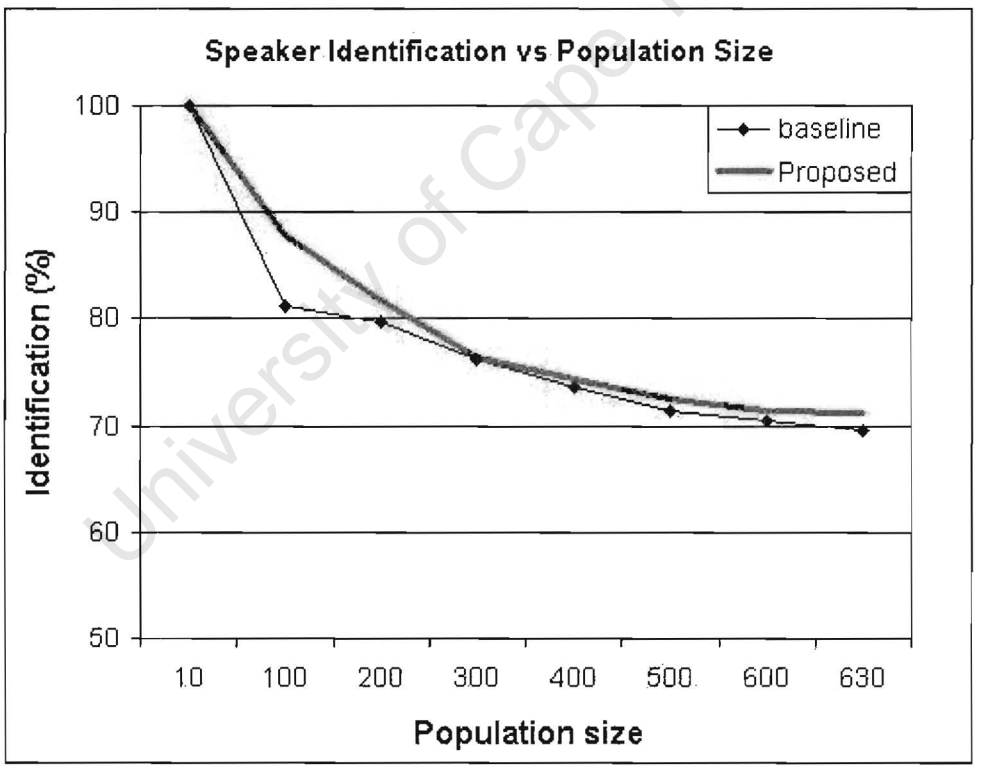

Figure 5.9: Average SiD rate as a function of population size

improvement when 10 speakers enroll into the system. From 300 speakers onwards, the RER increases. This demonstrates the ability of the proposed system to perform better under large population tests.

Statistical significance test

The N-best hierarchical method displays a consistent improvement at N=2 according to results shown in Table 5.7. It is appropriate however, to investigate whether the proposed N-best method significantly differs from the baseline system. The McNemar's test [44] is chosen because the test utterances are the same for both PFS-GMM and LPCC-GMM configurations and, secondly, the speaker identification decision

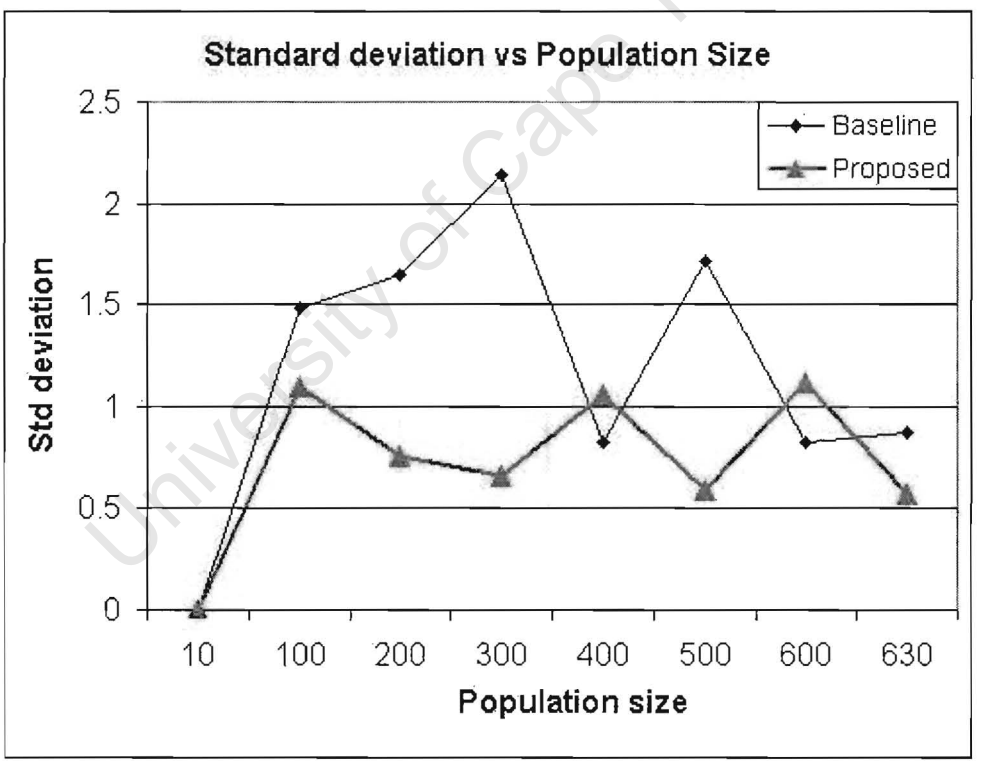

Figure 5.10: The standard deviation obtained from data in figure 5.9

is either *correct* or *incorrect*. This makes McNemar's a good candidate [44] for test of statistical significance.

The McNemar's test assumes two algorithms, A and B. The joint performance of A and B can be expressed in four different ways. In a given set of speaker models outcomes of the two systems, that is, the baseline (A) and N-best systems (B) have four possibilities during identification.

- N_{00} Number of speakers correctly identified by A , and correctly identified by B
- N_{01} Number of speakers correctly identified by A and incorrectly identified by B
- N_{10} Number of speakers incorrectly identified by A and correctly identified by B
- N_{11} Number of speakers incorrectly identified by both A and B

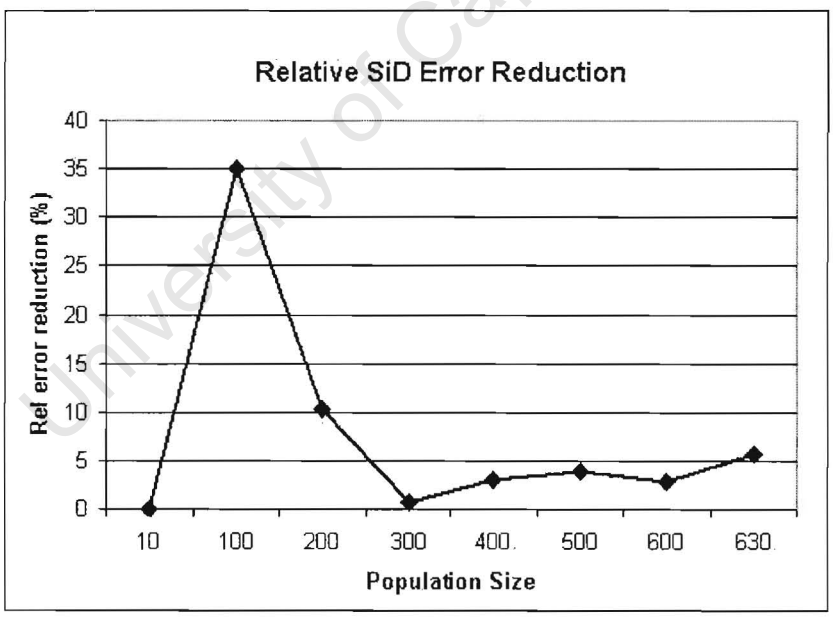

Figure 5.11: Relative error reduction for both baseline and the proposed N-best hierarchical SiD

Table 5.8 summarises the above information. The null hypothesis that A and B have equal performance [74] is used. This means the probability of one of these systems making an error is $\frac{1}{2}$. The test considers the condition where only one of the systems makes an error. The total number of speakers in this case is $K = N_{10} + N_{01}$. The N_{00} and N_{11} are not necessary for the McNemar's test. The number of observed speakers in K is equal to k. $k = n_{10} + n_{01}$. The null hypothesis now tested using a binomial distribution and the probabilities are calculated as follows:

$$P = 2 \sum_{m=n_{10}}^{k} \binom{k}{m} \left(\frac{1}{2}\right)^{k} \quad when\ n_{10} > \frac{k}{2} \tag{5.2}$$

$$P = 2 \sum_{m=0}^{n_{10}} \binom{k}{m} \left(\frac{1}{2}\right)^{k} \quad when\ n_{10} < \frac{k}{2} \tag{5.3}$$

The null hypothesis is rejected if the value of P is less than some chosen significance level, α. Typical values of α are 0.05, 0.01 and 0.001 [44].

		B	
		Correct	Incorrect
A	Correct	N_{00}	N_{01}
	Incorrect	N_{10}	N_{11}

Table 5.8: The McNemar's test parameters

The values of N_{10} and N_{01} were counted as the number of scores obtained from both baseline and the N-best hierarchical system. Since the top 2 speakers are those that the second algorithm needs, the scores were counted from the top speakers. The value of N_{10} was obtained by counting the number of the top two scoring speakers that the baseline system did not identify. The value of N_{01} was found by counting the number of the speakers who were correctly classified by the PFS-GMM system but incorrectly identified by the LPCC-GMM module. Then the results were obtained by using equations 5.2 and 5.3 as shown in table 5.9. The values of P from this table indicate that the PFS-GMM and LPCC-GMM configurations are significantly different only at trial numbers 3, 6 and 8 if a 10% level of significance is considered.

N-best system results discussion

The general performance of the system using test utterances 8 and 9 is provided in Table 5.10 in comparison with what was reported in literature under the same test conditions. The results shown in Table 5.10 also summarise what was reported in Table 5.7 with respect to the two systems that are similar to the baseline SiD.

The important metric of the hybrid system is the relative error reduction (RER). Therefore Table 5.12 illustrates how the RER of the proposed system compares to some of the previous hybrid SiD systems. The first observation is the differences between the system set-ups shown in Table 5.12. The general difference between this study and those that it compares to is the population size. The population of 100 speakers shows the highest RER that was observed in Figure 5.11. The work of Le Floch et al [47] differs from our proposed system because their ARVM and GMM are combined in a *competitive* and *cooperative* [47] way for decision making. Their

Trials	N_{10}	N_{01}	P
1	30	21	0.262
2	25	19	0.451
3	36	19	0.030
4	29	19	0.193
5	29	19	0.193
6	39	24	0.0769
7	25	18	0.360
8	41	25	0.064
9	33	28	0.609
10	24	16	0.268

Table 5.9: Statistical significance test results based on Table 5.7.

Implemented SiD System	SiD Performance (%)
Reynolds [16]'s SiD	60.7
Mashao [26] @ $\beta = 1.6$	68.9
LPCC based SiD	59.7
PFS-Baseline @ $\beta = 1.6$	69.4
Proposed 2-best hierarchical SiD	71.1

Table 5.10: The SiD rate on 630 speakers

work however uses NTIMIT like the authors. There is are no N-best list scoring protocols in their study. The polycost database used by Ganchev et al [19] contains the personal identity numbers (PINs) and 10 digits starting from 0 to 9. Their system could be regarded as text-dependent. This could be the reason for a very high RER. The rest of the RER values found in this research are not included in Table 5.12 because they are for higher speaker population sizes as opposed to the small populations used in the literature for N-best SiD systems. The other unique feature of this study is the use of the front-end module for complementing the baseline score.

Authors	Year	Corpus	Front-end	Back-end	Max. Population	SiD (%)
Le Floch *[47]*	1996	NTIMIT	LPCC	AVRM/ GMM	168	82.6
Le Floch [48]	2001	NTIMIT	LPCC	AVRM	630	58.0
van Vuuren [59]	1996	NTIMIT	LPCC	GMM	168	70.9
Fakotakis [52]	1999	NTIMIT	RASTA PLP	ANN	510	74.9
Besacier [75]	2000	NTIMIT	MFCC	Statistical Modeling	630	42.0

Table 5.11: Some previous SiD performances on NTIMIT database

Study Reference	Year	Speech Database	Speaker Population	Complementary modules	RER (%)
Fine [31]	2001	LLDB	52	GMM/SVM	25.7
Fine [32]	2001	LLDB	52	GMM/SVM	32.6
Le Floch [47]	1996	NTIMIT	168	ARVM/GMM	31.0
Ganchev [19]	2002	Polycost	110	PNN/GMM	59.9
Proposed SiD	2003	NTIMIT	100	PFS/LPCC	35.1

Table 5.12: The RER from different hybrid SiD systems

5.3 Summary

This Chapter has reported the experiments carried out in this work and the corresponding results. Two types of approaches were demonstrated and their results were observed. The first type of results in section 5.1 have shown limitations of the proposed SiD group detection hierarchical method. These results however have shown that the proposed method could improve the baseline if a perfect group detector could be implemented. After *forcing* robustness of the group detectors it was observed (see Table 5.3) that the baseline system's performance could be improved considerably. Some key objectives such as the improvement of the performance of the baseline system were not fully met by the group detection approach. The N-best hierarchical method results have been reported and some statistical tests were also performed. The relative error reductions from literature were compared with that of the proposed SiD system.

Chapter 6

Conclusions and future work

Two hierarchical methods were implemented in this study. Some literature has been reviewed in order to distinguish the meaning of "hierarchical systems" from different perspectives. The proposed methods were evaluated as outlined in Chapter 5. This work was carried out with the main objective of improving the baseline SiD system's performance. The second most important objective was to find ways of addressing the large population problem.

The group detection hierarchical method was initially considered the best candidate for solving the large population problem because the over-crowded feature space is split into smaller sub-sets which enable the model parameters to be more separable. The method limits the interspeaker variability. Observing that the first approach lacks robustness, a second method was proposed which uses the top N scores from the 630 available scores. This method was named the N-best hierarchical method which proved reliable for improving the baseline system. The N-best system met the main objective of improving performance according to Table 5.10. The improvement has not been statistically significant but in terms of relative error rate it competes with previously implemented SiD system (see table 5.12).

6.1 SID on NTIMIT speech

Speaker identification performance on NTIMIT speech has been found to be 75.5% and 86.5% when utilising the perfect PLP based group detector and an ideal dialect identifier, respectively. The average SiD rate of 71.1% is achieved by the second hierarchical method which is based on the N-best list of scores. The comparative literature values of 60.7% [16] and 68.6% [26] are slightly below the baseline SiD performance of 69.4% and therefore, the N-best list hierarchical SiD yields considerable improvement. The LPCC-GMM system also achieves 59.7% SiD rate. This is probably the reason why it causes the improvement on the baseline system. The proposed system performance was compared with previous work in Table 5.11 and we find that not many systems use large populations but their performance is relatively lower than achieved by the author's proposals. This means that on the large population investigations, the proposed SiD system is useful.

6.2 Hierarchical SiD methods

Improvement of the baseline system performance is the common goal for both proposed hierarchical systems. The goal has been met as described in section 6. The only drawback is that, although the group detection system is capable of solving the large population problem, it does not consistently classify speakers into their respective groups. The N-best system is robust as shown in Table 5.7 and Figure 5.8. The trend of decreasing performance as a function of increasing population size is still observed as shown in Figure 5.9. This observation shows the improved baseline which indicates that the objective of improving the baseline system using the whole NTIMIT database has been met. The relative error reduction in Figure 5.11 and Table 5.12 indicate the robustness of the proposed N-best hierarchical method. The McNemar's test however, does not fully indicate statistical significance in the performance of the proposed system.

6.3 Recommendations

After implementing and evaluating the two proposed hierarchical methods the following recommendations are made:

1. The group detection method did not fully meet the objectives of this study because of the ideal group detection algorithms. Therefore it is recommended that a text-dependent SiD system be set up for further evaluations. A dialect identification or language identification system should be merged with the SiD system to LPCC perform the group detection for obtaining a potentially high identification rate.

2. A full study on how to choose the best threshold in the N-best hierarchical system should be carried out using the optimised SiD system so that the standardised way of determining thresholds is established.

3. Both the group detection and the N-best hierarchical systems can be joined together to form a hybrid SiD system that perhaps performs better than either of the two.

4. This study has not investigated the noise issues as well as the cross-sex identification including many other problems. Therefore it is recommended that the pattern recognition study which directly deals with Gaussian overlaps on the feature space should be considered.

5. The statistical significance test has not seen much recognition in the filed of speech processing [44] and therefore it would be beneficial to investigate better ways and methods that are suitable for measuring the statistical significance levels suitable for speaker recognition systems.

Appendix A
Source Data for Graphs

This appendix lists the tables containing the source data for graphs in chapter 5. Values in these tables are those that were recorded during the experiments. The first section displays a table which reflects the cross-dialect and cross-gender confusion matrix. The second section contains tables of the SiD rate values when a perfect PLP group detector was used. The last section presents the data for the N-best list hierarchical system performance. In this section data for SiD performance as a function of population size is included.

Gender and dialect confusion

The gender and dialect discrimination capabilities of the baseline SiD system was tested by observing both the cross-dialect and cross-gender confusion during identification. All dialect regions in the NTIMIT database were used. Every time the speaker was identified, both the gender and dialect were checked using the known speaker labels from the NTIMIT database. Table 1 shows the outcome of the experiment. Each number in table 1 represents the number of speakers which the system *"thinks"* they belong to the two corresponding groups. For example, in dialect region 1 (dr1), the system correctly placed 20 male speakers as male and in the same dialect region The same thing happened with the 12 female speakers. However, the system correctly classified 2 female speakers from dr1 as female, but it categorised them as dr2 speakers. Similar explanation applies to all other dialect regions.

PLP based SiD performance

Upon realising from the gender detector and also from initial test that uncontrolled group detection hierarchical SiD yields low performance, perfect PLP based group detector was implemented and the results were checked for one dialect region at a time. Table 2 shows the number of speakers in each dialect region. Tables 3 (a) and (b) are results of the perfect PLP group detector for the baseline and the proposed SiD systems respectively. This experiment was done with varying training and testing sentences as shown the tables 3 (a) and (b).

		dr1		**dr2**		**dr3**		**dr4**		**dr5**		**dr6**		**dr7**		**dr8**	
		m	**f**	**m**	**f**	**m**	**f**	**m**	**f**	**m**	**f**	**m**	**f**	**m**	**f**	**m**	**f**
dr1	**m**	20		3		3		0		2		0		3		0	
	f		12		2		0		2		0		1		0		1
dr2	**m**	0		58		5		4		1		2		1		0	
	f		2		20		2		4		1		0		2		0
dr3	**m**	3		5		55		4		4		1		6		1	
	f		1		1		14		2		2		1		2		0
dr4	**m**	0		5		8		52		2		1		1		0	
	f		2		1		2		23		1		0		2		0
dr5	**m**	1		5		2		2		47		2		3		0	
	f		2		1		1		1		29		1		1		0
dr6	**m**	0	**1**	1		1		2		0		21		3		1	
	f		0		0		0		1		1	**1**	13		0		0
dr7	**m**	1		3		7		4		4	**1**	0		54		0	
	f		0		2		2		2		0		2		18		0
dr8	**m**	0		0		0		3		2		1		2		14	
	f		0		0	**1**	1		0		0		1		0		8

Table 1: Confusion matrix showing gender and dialect discrimination of SiD on NTIMIT database.

dialect region	dr1	dr2	dr3	dr4	dr5	dr6	dr7	dr8
population size	49	102	102	100	98	46	100	33

Table 2: Number of Speakers in each NTIMIT dialect region.

2-Best based SiD performance

This section tabulates source data for graphs that display the SiD system performance to address two different aspects SiD evaluation. First, different testing utterances or sentence pairs are used in order to acknowledge that the proposed system

Test	Dialect Region							
Sentences	dr1	dr2	dr3	dr4	dr5	dr6	dr7	dr8
0 , 1	85.7	74.5	75.5	83.0	75.5	91.3	79.0	84.8
1 , 2	89.8	83.3	80.4	83.0	77.6	93.5	87.0	93.9
2 , 3	83.7	80.4	83.3	85.0	76.5	89.1	88.0	81.8
3 , 4	77.6	79.4	83.3	81.0	79.6	91.3	87.0	87.9
4 , 5	91.8	76.5	83.3	86.0	79.6	89.1	84.0	93.9
5 , 6	83.7	79.4	83.3	82.0	83.7	87.0	88.0	84.8
6 , 7	93.9	82.4	84.3	85.0	82.7	95.7	84.0	87.9
7 , 8	93.9	86.3	83.3	84.0	77.6	97.8	80.0	93.9
8 , 9	83.7	84.3	83.3	84.0	83.7	89.1	78.0	87.9
0 , 9	77.6	87.3	75.5	85.0	77.6	93.5	75.0	87.9
Average iD	86.1	81.4	81.6	83.8	79.4	91.7	83.0	88.5

(a)

Test	Dialect Region							
Sentences	dr1	dr2	dr3	dr4	dr5	dr6	dr7	dr8
0 , 1	91.8	82.4	85.3	88.0	81.6	91.3	89.0	93.9
1 , 2	93.9	90.2	88.2	83.0	85.7	93.5	87.0	97.0
2 , 3	91.8	84.3	84.3	83.0	81.6	89.1	92.0	90.9
3 , 4	85.7	85.3	90.2	90.0	83.7	91.3	90.0	93.9
4 , 5	93.9	86.3	85.3	87.0	88.8	93.5	87.0	90.9
5 , 6	89.8	81.4	88.2	86.0	91.8	95.7	95.0	97.0
6 , 7	93.9	87.3	88.2	90.0	88.8	97.8	89.0	90.9
7 , 8	95.9	89.2	87.3	85.0	82.7	97.8	82.0	93.9
8 , 9	93.9	87.3	85.3	86.0	86.7	89.1	89.0	90.9
0 , 9	83.7	87.3	82.4	94.0	83.7	91.3	81.0	93.9
Average iD	91.4	86.1	86.5	87.2	85.5	93.0	88.1	93.3

(b)

Table 3: (a) Baseline (PFS-GMM) SiD performance (b) PLP based hierarchical SiD performance.

indeed improves the baseline system independent of varying testing sentence. Table 4 shows the SiD performances when test sentence pairs are varied. It should be noted however, that all sentences are different for each speaker except for utterances 0 and 1. This means that sentences 8 and 9 of a speaker are unique to that talker.

Test Utterance	Baseline (PFS-GMM)	Proposed (2-best system)
0 , 1	64.8	65.2
1 , 2	70.5	70.8
2 , 3	64.6	65.9
3 , 4	66.8	68.7
4 , 5	66.8	67.8
5 , 6	65.7	68.4
6 , 7	71.0	72.7
7 , 8	68.3	70.8
8 , 9	70.0	71.6
0 , 9	63.5	63.3
Average	67.2	68.5
Std. Dev.	2.649	3.0337

Table 4: Source data of text-independence Test for N-best system.

The second aspect of large population SiD system is the experiment to findout how identification varies with increase in enrolled speaker population size. The results of this experiment are recorded in table 5 (a) and (b) for the baseline and proposed SiD system respectively. Finally table 6 shows the relative error reductions (RER) after the baseline was improved. The RER figures were first obtained by calculating the identification errors made by both the baseline and the proposed system across the selected population sizes. Finally the RER values were obtained by using equation 5.1.

Test Utterance	Population Size							
	10	**100**	**200**	**300**	**400**	**500**	**600**	**630**
8 , 9	100	79.0	81.5	76.0	75.0	69.2	70.2	70.0
8 , 9	100	82.0	77.5	77.7	73.0	71.2	71.8	70.5
8 , 9	100	81.0	78.5	76.0	73.5	70.4	70.0	69.8
8 , 9	100	81.0	80.5	73.0	73.5	72.4	70.7	68.3
8 , 9	100	83.0	80.5	78.6	73.0	73.6	69.7	69.0
Average	100	81.2	79.7	76.26	73.6	71.36	70.48	69.52
Std. Dev.	0	1.483	1.643	2.140	0.822	1.711	0.823	0.870

(a)

Test Utterance	Population Size							
	10	**100**	**200**	**300**	**400**	**500**	**600**	**630**
8 , 9	100	88.0	82.0	75.3	73.3	73.4	72.2	71.6
8 , 9	100	89.0	81.5	76.7	73.5	72.2	72.2	71.7
8 , 9	100	88.0	81.0	77.0	75.5	71.8	72.0	71.7
8 , 9	100	88.0	81.5	76.5	74.3	72.4	70.2	70.8
8 , 9	100	86.0	83.0	76.7	75.5	72.6	70.0	70.5
Average	100	87.8	81.8	76.44	74.42	72.48	71.32	71.26
Std. Dev.	0	1.095	0.758	0.662	1.055	0.593	1.119	0.568

(b)

Table 5: (a) PFS-GMM baseline performace figures (b) 2-best System performance figures as a function of population size.

	Population Size							
	10	**100**	**200**	**300**	**400**	**500**	**600**	**630**
Baseline iD error	0	18.8	20.3	23.74	26.4	28.64	29.52	30.48
Proposed N-best iD error	0	12.2	18.2	23.56	25.58	27.52	28.68	28.74
Difference	0	6.6	2.1	0.18	0.82	1.12	0.84	1.74
Relative Error Reduction	-	35.106	10.344	0.758	3.106	3.911	2.846	5.709

Table 6: Relative Error Reduction as a function of population size.

Bibliography

[1] J. P. Campbell, "Speaker recognition: A tutorial," in *Proceedings of the IEEE*, vol. 85, pp. 1437–1462, 1997.

[2] T. Hazen and V. Zue, "Automatic language identification using a segment-based approach," in *Proceedings of Eurospeech*, pp. 1303–1306, 1993.

[3] K. Berkling and E. Barnard, "Language identification with inaccurate string matching." in *Proceedings of ICSLP*, 1996.

[4] L. M. Arslan and J. H. L. Hansen, "Language accent classification in american english," in *Speech Communications*, vol. 18, pp. 353–367, 1996.

[5] D. Rojas. "Probabilistic identification and quantification of linguistic affinity," Master's thesis, University of Edinburgh, 2002.

[6] W. H. Tsai and W. W. Chang, "Chinese dialect identification using an acoustic-phonotactic model," in *Proceedings of Eurospeech*, pp. 367–370, 1999.

[7] T. Parsons, *Voice and Speech Processing*. McGraw-Hill, 1986.

[8] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*. Prentice Hall PTR, 1993.

[9] R. L. Klevans and R. D. Rodman, *Voice Recognition*. Artech House, 1997.

[10] J. Campbell and D. A. Reynolds, "Corpora for the evaluation of speaker recognition systems," in *Proceedings of IEEE ICASSP*, pp. 2247–2250, 1999.

[11] J. Godfrey, D. Graff, and A. Martin, "Public databases for speaker recognition and verification," in *ESCA Workshop on Automatic speaker recognition 1994*, pp. 39–42, 1994.

[12] H. Melin, "Databases for speaker recognition: activities in cost250 working group 2," in *COST250 Workshop on speaker recognition in telephony 1999*, 1999.

[13] R. P. Ramachandran, K. R. Farrell, R. Ramachandran, and R. J. Mammone, "Speaker recognition - general classifier approaches and data fusion methods," in *The Journal of the Pattern Recognition Society*, pp. 2801–2821, 2002.

[14] D. Reynolds, "An overview of automatic speaker recognition technology," in *Proceedings of ICASSP, IEEE*, vol. IV, pp. 4072–4075, 2002.

[15] D. Reynolds, "Automatic speaker recognition using gaussian mixture models," in *The Lincoln Laboratory Journal*, vol. 8, pp. 173–192, 1995.

[16] D. A. Reynolds, "Large population speaker identification using clean and telephone speech," in *IEEE Signal Processing Letters*, vol. 2, pp. 46–48, 1995.

[17] D. A. Reynolds, "Effects of population size and telephone degradations on speaker identification performance," in *Proceedings of the SPIE Conference on Automatic Systems for the Identification and Inspection of Humans*, July 1994.

[18] U. V. Chaudhari, J. Navratil, G. Ramaswamy, and S. Maes, "Very large population text-independent speaker identification using transformation enhanced multi-grained models," in *Proceedings of IEEE ICASSP*, May 2001.

[19] T. Ganchev, A. Tsopanoglou, N. Fakotakis, and G. Kokkinakis, "Probabilistic neural networks combined with gmms for speker recognition over telephone channels," in *Proceedings of 14 Internationl Conference on DSP*, vol. II, pp. 1081–1084, July 2002.

[20] S. H. Maes, "Conversational biometrics," in *Proceedings of Eurospeech*, 1999.

[21] H. A. Murthy, F. Beaufays, L. Heck, and M. Weintraub, "Robust text-independent speaker identification over telephone channels," in *IEEE Transaction of Speech and Audio Processing*, vol. 7, pp. 554–568, 1999.

[22] A. A. Garcia and R. J. Mammone, "Channel-robust speaker identification using modified-mean cepstral mean normalization with frequency warping," in *Proceedings of IEEE ICASSP*, vol. 1, pp. 325–328, 1999.

[23] E. Monte, J. Hernando, X. Miro, and A. Adolf, "Text independent speaker identification on noisy envrionments by means of self organizing maps," in *Proceedings of ICSLP*, vol. 3, 1996.

[24] T. F. Lo, M. W. Mak, and K. K. Yiu, "A new cepstrum-based channel compensation method for speaker verification," in *Proceedings of Eurospeech*, vol. 2, pp. 775–778, 1999.

[25] NYNEX, "http://www.ldc.upenn.edu/catalog/docs/ldc93s2/ntimit.txt.".

[26] D. J. Mashao and N. T. Baloyi, "Improvements in the speaker identification rate using feature-sets on a large population database," in *Proceedings of Eurospeech*, vol. 4, pp. 2833–2836, 2001.

[27] H. S. M. Beige, S. H. Maes, J. S. Sorensen, and U. V. Chaudhari, "A hierarchical approach to large-scale speaker recognition," in *Proceedings of Eurospeech*, vol. 5, pp. 2203–2206, 1999.

[28] Z. Pan, K. Kotani, and T. Ohmi, "A on-line hierarchical method of speaker identification for large population," in *IEEE Proceedings of Nordic Signal Processing Symposium*, pp. 33–36, 2000.

[29] W. H. Abdulla and N. K. Kasabov, "Improving speech recognition performance through gender separation," in *Proceedings of the 5th Biannual Conference on Artificial Neural Networks and Expert Systems*, pp. 218–222, 2001.

[30] L. Lerato and D. J. Mashao, "Enhancing gmm scores of speaker identification using complementary feature sets," in *Proceedings of the 14th Annual Symposium of the PRASA*, pp. 91–95, 2003.

[31] S. Fine, J Navratil, and R. A. Gopinath, "Enhancing gmm scores using svm "hints"," in *Proceedings of Eurospeech*, 2001.

[32] S. Fine, J. Navratil, and R. Gopinath, "A hybrid gmm/svm approach to speaker identification," in *Procedings of IEEE ICASSP*, 2001.

[33] C. Wu and J. Chen, "Text-independent speaker identification based on small training data and fast search algorithms," in *Journal of Information Science and Engineering*, vol. 11, pp. 73–87, 1995.

[34] M. F. BenZeghiba and H. Bourlard, "User-customized password speaker verification based on hmm/ann and gmm models," in *Proceedings of ICSLP 200d2*, pp. 1325--1328, 2002.

[35] B. L. Pellom and J. H. L. Hansen, "An efficient scoring algorithm for gaussian mixture model based speaker identification," in *IEEE Signal Processing Letters*, vol. 5, pp. 281–284, 1998.

[36] J. D. Markel and A. H. Gray, *Linear Prediction of Speech*. Springer-Verlag, 1976.

[37] D. J. Mashao and J. E. Adcock, "Utterance dependent parametric warping for a talker independent hmm-based recognizer," in *Proceedings of IEEE ICASSP*, pp. 1235--1238, 1997.

[38] S. Grassi, L. Besacier, A. Dufaux, M. Ansorge, and F. Pellandini, "Influence of gsm speech coding on the performance of text-independent speaker recognition," in *European Association for Signal Processing EUSIPCO*, 2000.

[39] E. T. S. institute, "www.etsi.org." GSM CODEC.

[40] L. Lerato and D. J. Mashao, "Evaluation of speaker identification using gsm data," in *Southern African Telecommunications and Network Applications Conference*, vol. 1, pp. 75–78, 2002.

[41] L. Lerato and D. J. Mashao, "Call centre speaker identification using telephone and gsm data," in *Southern African Telecommunications and Network Applications Conference*, 2003.

[42] L. Lerato and D. J. Mashao, "Hierarchical approach for improving speaker identification," in *Proc. of 13th Annual Symposium of the PRASA*, pp. 51–55, 2002.

[43] A. Mehrotra, *GSM system engineering*. Artech House, London, 1997.

[44] L. Gillick and S. J. Cox, "Some statistical issues in the comparison of speech recognition algorithms," in *Proceedings of IEEE International Conference on Acoustics and Speech Siganl Processing*, pp. 532–535, 1989.

[45] A. Toutios and K. G. Margaritis, "Development of a text-dependent speaker identification system with the ogi toolkit," in *Proceedings of the Conference on AI, SETN-2002*, pp. 525–530, 2002.

[46] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, pp. 18–32, 1994.

[47] J. L. L. Floch, C. Montacie, and M. J. Caraty, "Gmm and arvm cooperation and competition for text-independent speaker recognition on telephone speech," in *Proceedings of ICSLP*, vol. 4, 1996.

[48] J. L. L. Floch, C. Montacie, and M. J. Caraty, "Speaker recognition experiments on the ntimit database," in *Proceedings of Eurospeech*, pp. 379–382, 1995.

[49] D. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture models," in *IEEE transactions on Speech and Audio Processing*, vol. 3, pp. 72–83, 1995.

[50] S. Srivastava, "Fundamentals of linear prediction." The lecture: Mississippi State University Elec.Eng., 1999.

[51] R. A. B. Soria and E. F. Cabral., "Combining neural networks paradigms and mel-frequency cepstral coefficients correlations in a speaker recognition task," in *Proceedings of the International Conference on Signal Processing Applications and Technology*, 1996.

[52] N. Fakotakis, J. Sirigos, and G. Kokkinakis, "High performance text-independent speaker recognition system based on voiced/unvoiced segmentation and multiple neural nets," in *Proceedings of Eurospeech*, 1999.

[53] D. Mashao, *Computations and Evaluations of an Optimal Feature-set for an HMM-based Recognizer*. PhD thesis, PhD. Brown University, 1996.

[54] K. Gopalan, T. R. Anderson, and E. J. Cupples, "A comparison of speaker identification results using features based on cepstrum and fourier-bessel expansion," in *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 289–294, 1999.

[55] Y. Shao and D. Wang, "Co-channel speaker identification using usable speech extraction based on multi-pitch tracking," in *Proceedings of IEEE ICASSP*, vol. II, pp. 205–208, 2003.

[56] W. Dhaes and X. Rodet, "Discrete cepstrum coefficients as perceptual features," in *Proceedings of International Computer Music Conference*, 2003.

[57] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing mel-frequency cepstal coefficients on the power spectrum," in *Proceedings of the IEEE ICASSP*, pp. 73–76, 2001.

[58] R. Chengalvarayan, "Hierarchical subband linear predictive cepstral (hslpc) features for hmm-based speech recognition," in *Proceedings of IEEE ICASSP*, vol. 1, pp. 409–412, 1999.

[59] S. van Vuuren, "Comparison of text-independent speaker recognition methods on telephone speech with acoustic mismatch," in *Proceedings of ICSLP*, pp. 1788–1791, 1996.

[60] C. Tanprasert, C. Wutiwiwatchai, and S. Sae-tang, "Text-dependent speaker identification using neural networks on distinctive thai tone marks," in *NECTEC Technical Journal*, vol. 1, pp. 249–253, 2000.

[61] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990.

[62] D. J. Mashao, "Comparing svm and gmm on parametric feature-sets," in *Proceedings of the 14th Annual Symposium of the PRASA*, pp. 15–20, 2003.

[63] D. J. Mashao, "Parallel processing for the auditory feature-set of a speaker recognition system," in *Proceedings of 2002 SCI-ISAS*, 2002.

[64] N. T. Baloyi, "Comparison of features for large population speaker identification," Master's thesis, University of Cape Town, 2000.

[65] V. Faber, "Clustering and the continuous k-means algorithm," in *Los Alamos Science Journal*, no. 22, pp. 138–144, 1994.

[66] A. Cohen and V. Lapidus, "Unsupervised text independent speaker classification," in *Proceedings of the International Conference on Signal Processing Application and Technology*, pp. 1745–1799, 1996.

[67] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and the kernel-based learning methods*. Cambridge University Press, 2000.

[68] H. Huang and C. Hsu, "Recognizing 100 speakers using homologous naive bayes," in *Proceedings of the 7th PRICAI*, 2002.

[69] F. A. Westall, R. D. Johnston, and A. V. Lewis, "Speech technology for telecommunications," in *BT Technol Journal*, vol. 14, pp. 9–27, 2001.

[70] J. Ortega-Garcia, J. Gonzalez-Rodriguez, and V. Marrero-Aguiar, "Ahumada: A large speech corpus in spanish for speaker characterization and identification," in *Speech Communication Journal*, pp. 255–264, 2000.

[71] D. Ellis, "Improved recognition by combining different features and different systems," in *Proceedings of AVIOS-2000*, 2000.

[72] S. Sivadas and H. Hermansky, "Hierarchical tandem feature extraction," in *Proceedings of IEEE ICASSP*, vol. I, pp. 809–812, 2002.

[73] S. R. Waterhouse and A. J. Robinson, "Classification using hierarchical mixture of experts," in *Proceedings IEEE Workshop on Neural Networks for signal Processing*, pp. 177–186, 1994.

[74] M. Forsyth, "Discriminating observation probability (dop) hmm for speaker verification," in *Speech Communication*, pp. 117–129, 1995.

[75] L. Besacier, J. F. Bonastre, and C. Fredouille, "Localization and selectin of speaker specific information with statistical modeling," in *Speech Communication*, pp. 89–106, 2000.

[76] I. Magrin-Chagnolleau, G. Gravier, M. Seck, O.Boeffard, R. Blouet, and F. Bimbot, "A further investigation on speech features for speaker characterization," in *Proceedings of ICSLP*, 2000.

[77] K. Chen and H. Chi, "A method of combining multiple probabilistic classifiers throuth soft competition on different feature sets," in *Neurocomputing journal*, pp. 227–252, 1998.

[78] V. Radova and J. Psutka, "An approach to speaker identification using multiple classifiers," in *Proceedings of IEEE ICASSP*, pp. 1135–1138, 1997.

[79] L. Mothae and D. J. Mashao, "Using a polynomial approximation based impulse noise suppression algorithm for robust speaker verification," in *Proceedings of the 14th Annual Symposium of the PRASA*, pp. 85–89, 2003.