LOOK WHO’S TALKING NOW: THE CHALLENGE OF 
TELEPHONE CONVERSATIONS

A Dissertation Proposal

submitted to
the Temple University Graduate Board

in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Engineering

by
Uchechukwu Ofoegbu
September, 2006

Dr. Robert Yantorno
Advisor
Director of Speech Processing Lab
Temple University

Dr. Dennis Silage
Committee Member
Director of System Chip Design Lab
Temple University

Dr. Iyad Obeid
Committee Member
Department of Electrical Engineering
Temple University

Dr. Kwatny
Committee Member
Department of Computer Science
Temple University
ABSTRACT

Given a collection of speech utterances from different individuals, the task of determining WHO spoke WHEN is usually referred to as speaker recognition/discrimination. This problem is generally encountered in speech applications such as speaker identification, verification, indexing/clustering and speaker counting. In some of these applications – such as speaker identification and verification, the utterances are available and usually consist of considerable amount of data, while in others – such as speaker indexing/clustering and count, each speaker’s utterance is embedded in the midst of others, and the amount of data could be large (like in the case of broadcast news data) or small (like in the case of telephone conversations). Additionally, some applications involve differentiating between speakers using a priori information about the speakers (supervised speaker recognition), while others involve working directly with the test data (unsupervised speaker recognition). Speaker discrimination in telephone conversations is generally unsupervised, with only short lengths of homogeneous speaker utterances available. This is a very challenging task as speaker recognition systems are known to perform better when more data is available, and even more so when the process is supervised. The goal of this research is to address the challenge of limited data and no a priori speaker information by investigating distance measure and speaker-differentiating features and to develop a speaker recognition system which is optimized for telephone conversations. Distance measures (including the Mahalanobis distance, the T-Square statistics, the Kullback-Leibler distance, the Bhattacharyya distance and Levene’s test), features (such as the Linear Predictive Cepstral Coefficients, The Mel-Scale Frequency Cepstral Coefficients and the first and second derivates) and data enhancement techniques for improved speaker recognition performance with telephone data will be investigated. In addition, the formulation of a novel distance measure will be discussed, which is derived from a weighted combination of the outputs from currently existing distance measures optimized for maximum inter-speaker and minimum intra-speaker variation. A decision-level fusion technique will also be investigated. The applications of this research include such things as: criminal detection and forensics, commercial services, military activities and terrorist identification.
TABLE OF CONTENTS

ABSTRACT ................................................................................................................................................... ii
Chapter 1: INTRODUCTION .........................................................................................................................1
  1.1. An Overview of Speaker Discrimination ......................................................................................... 1
  1.1.1. Types of Speaker Discrimination .............................................................................................. 1
  1.1.2. Application of Speaker Discrimination ...................................................................................... 2
  1.2. Distances for Speaker Discrimination ........................................................................................... 5
  1.3. Features for Speaker Discrimination ............................................................................................. 7
  1.4. Application of Fusion in Speaker Discrimination .......................................................................... 8
  1.5. Research Goal: The Challenge in Telephone Conversations ....................................................... 10
  1.6. Scope of Research ....................................................................................................................... 11
  1.7. Proposal Outline ......................................................................................................................... 12

Chapter 2: ANALYSIS OF DISTANCE MEASURES FOR SPEAKER DISCRIMINATION ............... 15
  2.1. Introduction ................................................................................................................................... 15
  2.2. Distance Measures ....................................................................................................................... 16
     2.2.1. Mahalanobis Distance ........................................................................................................ 16
     2.2.2. Hotelling’s T-Square Statistics ......................................................................................... 17
     2.2.3. Kullback-Leibler Distance ................................................................................................. 17
     2.2.4. Bhattacharyya Distance .................................................................................................... 18
     2.2.5. Levene’s Test ..................................................................................................................... 19
  2.3. Analysis of Distance Measures for Speaker Discrimination ....................................................... 20
     2.3.2. Procedural Set-Up ............................................................................................................. 20
     2.3.3. Distance Analysis ................................................................................................................ 22

Chapter 3: ANALYSIS OF FEATURES FOR SPEAKER DISCRIMINATION ...................................... 28
  3.1. Introduction ................................................................................................................................... 28
  3.2. Cepstral Features ......................................................................................................................... 29
     3.2.1. Cepstral Analysis .............................................................................................................. 29
     3.2.2. Linear Predictive Cepstral Coefficients ........................................................................... 30
     3.2.3. Mel-Scale Frequency Cepstral Coefficients ..................................................................... 32
CHAPTER 1
INTRODUCTION

1.1 An Overview of Speaker Discrimination

Speaker discrimination, which is synonymous with speaker recognition, has been a major, rapidly evolving aspect of speech processing for over four decades. It differs from its counterpart, speech recognition, in that speaker discrimination is the process of recognizing speakers using their voice characteristics while speech recognition involves identifying what is being said. The process of speaker discrimination is primarily based on determining and utilizing acoustic characteristics of speech that are observed to be different for different individuals. These acoustic features generally contain information about oral anatomy, such as the shapes and sizes of the mouth as well as the throat, and prosodic attributes including pitch, loudness, tempo, rhythmic patterns and accent. Speaker discrimination can be grouped into various types, and used in several applications, as explained below.

1.1.1. Types of Speaker Discrimination

Speaker discrimination can be classified into three main types, depending on the style of speech input into the speaker recognition system, which are as follows:

Text dependent speaker discrimination: This is the art of differentiating speakers by selecting and observing characteristics from specific words or phrases which are repeated by all participating speakers. In this case, the system is trained to recognize specific speakers from pre-defined utterances, and is most likely to fail if a different utterance is given. This input style is generally employed in applications were speakers need to be verified for security purposes.
Text prompted speaker discrimination: This approach is very similar to text dependent speaker discrimination, as the system recognizes speakers based on specific words or phrases. However, in this case, the system prompts each speaker to give a new utterance, which is selected randomly for each prompt. This approach is normally used in situations where there is the possibility of imposters.

Text independent speaker discrimination: With this method, speakers are differentiated regardless of what is being uttered by the participants. The system is trained such that the speaker-different features do not depend on any particular set of phonemes, words, phrases or sentences. This is the most complex form of speaker discrimination, as it assumes the least amount of knowledge of user input.

1.1.2. Applications of Speaker Discrimination

Several well-known speech processing applications are based on differentiating between speakers, the most important of which are briefly described below.

Speaker Identification: This is the process of determining which speaker, amongst a group of registered speakers, corresponds to a given test utterance. Speaker identification systems generally require a priori information about all participating speakers, and the identification process is usually performed in two stages: the training and the testing stages. During training, the system extracts features or distinguishing voice characteristics of individuals by forming models from their utterances. These utterances are referred to as the training dataset. During testing, a different set of utterances, from the same group of speakers is used. This time, the system is expected to examine each test utterance and determine, based on distance measurements, probabilistic inferences or statistical measures, who the participating speaker is which produced the utterance. Speaker Identification (SID) can be either open-set or closed-set. With open-set SID, the test utterance could be either one of the enrolled speakers, or an unknown speaker. With closed-set SID, the test utterance is expected to have been produced by one of the enrolled participants. The most common technique that has been employed in speaker identification is the use of Gaussian Mixture Models, which was initiated by [Reynolds, 1997]. This method involves representing each speaker using a number
(N) of Gaussian models, and is based on the assumption that a speaker’s utterance contains more than one (or about N) classes of speech or phonemic groups, and each model corresponds to each group. The use of Neural Networks for Speaker Identification was investigated by [Rudasi and Zahorian, 1991], and the use of a more complex technique, Hidden Markov Models, was employed by [Matsui and Furui, 1992]. Results, however, proved the GMM methods to be superior to these methods, which require either the same or a greater amount of computational complexity. Simpler approaches to speaker identification, which involve formation of single models (Gaussian or non-parametric), and the use of the minimum distance criterion in determining the identity of speakers have also been examined. Such methods were studied in detail by [Iyer et al., 2006d]. Presently, speaker identification systems are able to yield up to 96% accuracy in identifying speakers even in noisy conditions. Several methods for enhancing degraded speech in order to increase identification accuracy have been investigated, the most recent being the detection and extraction of ‘usable’ (or non-degraded) speech portions and the elimination of ‘unusable’ speech portions from observed speaker utterances. This method was initiated by [Yantorno, 1998], and several techniques for identifying usable or unusable speech segments have been presented by [Krishnamachari, et al., 2000], [Yantorno et al., 2000], [Yantorno et al., 2001], [Krishnamachari et al., 2001], [Lovekin, et al., 2001a], [Lovekin, et al., 2001b], [Smolenski et al., 2002a], [Smolenski, et al., 2002b], [Smolenski, et al., 2002], [Kizhanatham and Yantorno, 2002], [Chandra and Yantorno, 2002], [Iyer and Yantorno, 2003], [Kizhanatham et al., 2003], [Sundaram et al., 2003], [Shao and Wang 2003], [Ofoegbu et al., 2004], [Iyer et al., 2004], [Shao and Wang 2006]. All these proposed techniques showed significant improvement in the performances of SID systems. SID systems may be applied in criminal detection and forensics, automated customization, pilot-base communications and terrorist activity detection.

Speaker Verification: This application is very similar to SID, except that the goal, in this case, is to either accept or reject the identity claim of a speaker based on comparisons of a test utterance with previously enrolled utterances of the presumed speaker. Comparison methods and enhancement techniques for speaker verification (SV) systems do not vary significantly with those for SID systems. Some applications of SV systems include criminal investigations, secured control access systems for services such as private
database access, voice dialing, telephone banking or shopping, information services, voice mail, security control for confidential information areas, and remote access to personal computers.

**Speaker Change-Point Detection:** This is a relatively recent branch of speaker recognition, which involves detecting start and end points of homogeneous speaker utterances in conversations. For some change-point detection systems, information about speakers participating in the observed conversation is obtained *a priori*, and the system is trained with this information. In several cases, however, systems are designed to detect change-points without any knowledge of speaker information, which makes the problem more complex. Some techniques that have been proposed for speaker change-point detection include the use of the Bayesian Information Criterion [Chen and Gopalakrishnan, 1998], [Jitendra et al., 2004], Vector Quantization techniques [Liu and Kubala, 1999], [Jorgensen et al., 2006], energy of the speech signals [Kemp et al., 2000], Support Vector Machines [Kartik et al., 2005], Universal Background Models [Wu et al., 2003a], [Wu et al., 2003b], Bayesian Fusion methods [Lu et al., 2002], and distance-based measures [Chen and Gopalakrishnan, 1998], [Gish et al., 1999], [Adami et al., 2002]. Speaker change point detection is generally used as an initialization step for more difficult tasks such as speaker clustering/indexing [Chen and Gopalakrishnan, 1998], [Zhou and Hansen 2000], speaker count, and other blind speaker recognition applications (where information of about the speaker is unknown).

**Speaker Indexing:** This is task is synonymous to speaker diarization and often also referred to as speaker clustering. It involves examining multi-speaker speech data and determining who is speaking, and when they are speaking. Speaker indexing is also a relatively recent branch of speaker-recognition and can be classified into two categories: supervised and unsupervised speaker indexing. The former category involves having some information about the speakers participating in the observed speech data; whereas, in the latter, this information is lacking. Consequently, the problem of unsupervised indexing is more difficult than supervised speaker indexing, and has been addressed using techniques such as Neural Networks [Roy 1997], Generic Speaker Modeling [Kwon and Narayanan, 2004], Bayesian Information Criterion [Delacourt and Wellekens, 2002], [Zhou and Hansen 2000], Support Vector Machines [Ferganal et al., 2006], distance measurements [Iyer et al., 2006a], [Iyer et al., 2006b], and a combination of two or more
techniques [Delacourt and Wellekens, 2002]. Speaker Indexing is generally performed for broadcast news data, conferences, telephone conversations, and other multi-speaker events. The primary application of speaker indexing is in monitoring and mining speech data. For instance, meetings or conferences could be monitored and archived for later access for review purposes and/or by interested individuals who were unable to attend such meetings.

**Speaker Count:** This application involves determining the number of speakers participating in a conversation (most likely without having any *a priori* information about any of the speakers). Speaker count systems can be employed in criminal activity detection; for example, in prisons, where three-way calling is prohibited, detecting the presence of a third speaker in recorded conversations could be helpful in identifying violators. Speaker count systems can also be utilized in speaker indexing procedures where the number of speakers is unknown. Some approaches that have been applied to speaker count include distance-based methods [Ofoegbu et al., 2006a], [Ofoegbu et al., 2006b] and generic modeling [Iyer et al., 2006a].

### 1.2. Distances for Speaker Discrimination

All speaker recognition applications are based on differentiating between speakers. Although several non distance-based methods such as GMM, HMM and Neural Networks have been used widely especially in speaker recognition [Chaudhari et al., 2001], [Reynolds 1992], [Naik, 1990], [Rudasi and Zahorian, 1991], [Matsui and Furui, 1992], many techniques used in performing these tasks usually rely heavily on distance/divergence measures, be the statistical or probabilistic. Some common distance measures used for speaker recognition include Euclidean, Mahanalobis and City-block distances [Ong and Yang, 1998]. For speaker indexing and change-point detection, Kulback-Leibler (KL) and Gaussian-Likelihood-Ratio (GLR) [Delacourt and Wellekens, 2002] distances have been used.

Some previous analyses of distance measures in speaker discrimination are as follows. Ong and Yang (1998) performed a comparative study of the use of distance measures as well Gaussian probability density
estimates in speaker identification. Their results showed that, in general, probability estimates yielded higher results than distance measures, however, this was probably due to the fact that, the distance measures studied, which included the Euclidean, City block and Mahanalobis distances, were simplified models of the Gaussian probability estimates themselves. Moreover, when a weighting function was applied to the Euclidean distance, it outperformed all other features. For speaker indexing, [Delacourt and Wellekens, 2002] made an effort to improve the performance of the Bayesian Information Criterion (BIC) technique developed by [Chen and Gopalakrishanan, 1998] by introducing a modified version which involved combining decisions of the GLR and KL distances as well as four similarity measures in order to tentatively detect speaker change points, which were then refined using the BIC algorithm. This modification was shown to enhance considerably the speaker indexing system. One observation made by Delacourt and Wellekens was that the distance-based parameters varied significantly with varying data lengths. An effort was also made [by Ofoegbu et al., 2006] to determine the number of speakers in a telephone conversation using the T-Square statistic.

In addition to the employment of distances in speaker recognition systems, several studies have been performed to observe and compare the speaker recognition performance of several distance measures in various conditions. Some of these studies include the following: Gray and Markel (1976) presented a review of distance measures for speech processing. The Root Mean Square (RMS) log spectral distance, cepstral distance, likelihood ratio, and a cosh measure were investigated, a detailed explanation of these measures and their behaviors was given but there was no discussion or conclusion about their implementation in speaker discrimination or even any speech processing application. Gray et al. (1980) performed another speech processing distance study which was very similar to the former, except that more distances were observed, and a more careful and comprehensive analysis was performed. Bimbot et al. (1995) later investigated the use of similarity measures for SID which include the Gaussian likelihood measure, the Arithmetic-geometric sphericity measure, and the symmetrization measure. The research also involved a comparison of the performance of these measures in SID systems, the only factors varied in the experiments being the distances themselves and the relative amounts of training and testing data. Moreover, the distance study focused mainly on SID, and not on the general speaker discrimination task. Souza (1977)
presented a study of statistical distances for differentiating short segments. The distance measures studied included Itakura’s $\chi^2$ test, Quenoille’s test and the regression test. Very recently, a comprehensive study of distances for different speaker recognition applications in several conditions was performed by Iyer et al. (2006). The Mahanalobis distance, Kulback-Leibler (KL) distance, T-Square statistic, Euclidean distance, Hellinger distance, Bhattacharyya distance, Gaussian Likelihood Ratio (GLR), L-infinity and Levene distances were examined, and the results suggested that if applied judiciously, simple distance-based systems could yield results which are comparable to more computationally complex systems such as GMM, HMM and Neural Networks. Distance-based systems are also much easier to analyze, and much more flexible and less sensitive to unfavorable conditions (such as limited data sizes or lack of a priori information) than the more complex systems.

1.3. Features for Speaker Discrimination

Extraction of features is a major stage in the speaker discrimination process. It entails converting the observed speech signals to parametric representations, commonly referred to as feature vectors, which are then analyzed and processed by means of classifiers (distance-based or non-distance-based). The following are some desirable attributes of speaker discriminating features:

- **Computational simplicity:** this attribute is based on Occam’s razor, which states: “when you have two competing theories which make exactly the same predictions, the one that is simpler is the better”.

- **Frequent occurrence in speech:** this characteristic is necessary because scarcity or irregularities in speech signals can lead to recognition errors or even system failures.

- **High Inter-speaker but low intra-speaker variations**

- **Robustness to noise, channel distortions or other unfavorable conditions**

- **Difficult to falsify:** Sometimes speakers may try to alter their voices to avoid being identified, but features that will remain the same in spite of these alterations are usually very desirable.

Speaker discrimination features can be classified into ‘high-level’ and ‘low-level’ features [Quatieri, 2004].
**High-Level Features:** These include voice characteristics such as clarity, roughness, animation, energy and prosody (i.e. pitch intonation, accent and articulation rate) [Reynolds, 1992], [Voiers, 1964]. It was observed that such high-level features provide effective perceptual indications of speaker differences [Voiers, 1964]. However, due to their sensitive nature, these attributes are quite difficult to extract automatically, and are therefore not commonly used in speaker recognition. [Quatieri, 2004].

**Low-Level Features:** These features are easier to compute because they are acoustic in nature. They include spectral features, cepstral features and their derivatives, short-time pitch, glottal flow derivatives, and other minor temporal properties. Spectral features consist of features extracted from short-time spectral measurements performed on the speech waveform. The measurements are taken on short frames of speech due to the nonstationary nature of speech which causes its characteristics to change after short durations. Examples of such features are Linear Predictive Coding (LPC) coefficients, Line Spectral Pairs (LSPs), Perceptual Linear Predictive (PLP) coefficients and the Log Area Ratios (LARs). The LSPs, PLP coefficients and LARs are more robust derivations of Linear Predictive Coding, which is based on representing speech signals as a weighted sum of past samples. Other features which are based on Linear Predictive Analysis have been investigated by Ramachandran et al. (1995). Spectral features are generally very effective in representing speech signals, and have been widely used in speech and speaker recognition systems since the 1970s. On the other hand, spectral features are obtained by taking the Inverse Fourier Transform (IFT) of the logarithms of the Short-Time Fourier Transform (STFT) of speech signals. This process is carried out in order to separate slowly varying formant coefficients (vocal tract information) from fast varying harmonics (excitation information) observed in the speech spectrum by taking advantage of the fact that the convolution of two signals in time is equivalent to multiplication in frequency, and then using the logarithmic operation converts the multiplication to addition. Cepstral features have been found to be very effective in preserving speaker-dependent properties of speech signals, and are more widely used in speaker recognition systems than spectral features, and have been studied extensively for speaker recognition [Mammone, 1996]. More state-of-the-art speaker recognition systems utilize cepstral features than any other types of features. The most commonly used cepstral features are the Linear Predictive
Cepstral Coefficients (LPCCs), which are usually derived from LPC coefficients, and the Mel-Scale Frequency Cepstral Coefficients (MFCCs), which are based on auditory perception of speech signals. [Oppenheim and Schafer, 1989]. First and second order derivatives of cepstral features are also used sometimes in conjunction with the cepstral coefficients so as to represent the dynamics of the cepstrum. They are generally referred to as delta and delta-delta coefficients, respectively.

Other low-level features can be computed directly from the time waveform without any transformations, and are usually combined with spectral or cepstral features in order to improve their speaker discriminative capabilities. More detailed descriptions of some of the low-level features are given later in this proposal.

1.4. Application of Fusion in Speaker Discrimination

Over the past two decades, several speaker recognition features, classifiers, and systems have been developed, most of which are complimentary, or at least uncorrelated to one another. As a result, in the past decade, some techniques have been investigated for fusing these features, classifiers or systems in order to utilize the different information from each one. Some of these techniques are discussed below.

Higgins et al. (1999) introduced multi-spectral multi-score multi-source fusion approach in which the speech signal is first filtered into several sub-bands and the output of each filter is separately modeled by linear prediction cepstral coefficients. The models are then matched against the test models and scores are combined using the sum rule of information fusion. Kinnunen et al. (2003) presented classifier-based fusion method for speaker verification whereby a combined match score for the unknown speaker was determined using a reliability-based weighted sum of multiple supplementary classifiers. Kinnunen et al. (2004) also developed a feature-based fusion technique which involves feature-level and decision-level fusion of various spectral features. Kajarekar (2005) also devised a classifier-based fusion system which performed speaker recognition based on a combination of scores from four different Support Vector Machines.
Fusion of generative and discriminative classifiers (generative being those that entail learning a model of the joint probability of the inputs and the label and making predictions based on Bayes probability rules and discriminative classifiers those that compute the posteriori probability directly or map the inputs to their labels directly [Ng and Jordan, 2002]) was introduced by Campbell et al. (2004), who proposed novel techniques for fusing Support Vector Machine (SVM) classifiers with Gaussian Mixture Model (GMM) classifiers to improve speaker recognition, and also proved that both techniques were complimentary in nature. Scheffer and Bonastre, (2006) also presented a new method with the same basic idea, whereby combined the UBM-GMM (Universal Background Models – Gaussian Mixture Models) classifier with SVM for speaker detection.

In this research, a technique for fusing information from distance measures is proposed. A ‘feature’-level fusion method (the ‘features’ being the distances) is considered, where the measures are assigned weights computed such that a linear combination of the weighted distances would yield minimum intra-speaker and maximum inter-speaker variation. A decision-level fusion approach is also discussed, which involves combining the decision outputs of the measures based on weights determined from the individual speaker discrimination performance of each distance.

1.5. Research Goal: The Challenge of Telephone Conversations

The main goal of this research is to address the challenge of differentiating speakers participating in telephone conversations. Discriminating speakers participating in telephone conversations is a more complex task than other speaker recognition endeavors for the following reasons:

1. *Speakers change rapidly*: during telephone conversations, speaker turns could last less than 1 second (or even less), and usually last for an average of about 1 second [Iyer *et al.*, 2006a], [Ofoegbu *et al.*, 2006]. As a result, the length of homogeneous speaker utterances (that can be
used in creating speaker models or forming speaker-consistent feature vectors) is limited to about
1 second for telephone conversations, since using more data could lead to errors caused by
forming speaker models from speech data obtained from more than one speaker [Iyer et al.,
2006b], [Ofoegbu et al., 2006b]. Research has shown that, given only brief utterances (1 second or
less), humans can recognize speakers with an accuracy of about 54% on average [O'Shaughnessy,
1999]. For SID and SV methods, at least 5 seconds of speaker consistent data is available for
speaker model formation. Also, in most cases, long consecutive speaker utterances can be
extracted from broadcast data or conference speech for indexing, making the speaker
differentiation process less challenging than with telephone conversations. Previous research has
shown that in speaker recognition systems, where data limitation is not an issue to be concerned
about, the use of more data results in better performance; but in telephone conversations, the
contrary is the case [Iyer et al., 2006d].

2. Lack of a priori information: Unlike in SID, SV and speaker indexing of broadcast or conference
data (in some cases), speech data from speakers participating in telephone conversations is usually
unavailable, so the system cannot be trained with any speaker-dependent parameters. Moreover,
for speaker-indexing, the number of speakers recorded is sometimes known to the system. This is
not always the case, especially when handling telephone conversations such as those made from
prisons. This lack of information poses a challenge for speaker recognition systems which have to
deal with telephone conversations.

3. Adverse conditions: Like with any speech produced in uncontrolled environments, working with
telephone data generally involves dealing with unfavorable conditions such as co-channel speech,
channel distortions, frequency alterations, background noise, and so on. These could result in
adverse degradation in the speaker recognition performance if not effectively addressed by the
system.

In this research, all the above mentioned challenges will be addresses, and a novel, speaker
discrimination technique will be introduced for telephone conversations or speech data with similar
characteristics.
The primary speaker discrimination approach considered in this research is the use of distance measures. One important investigation performed in this research is the formulation of a robust distance measure, which is created using an optimized combination of select distance measures, taking into account the correlation and dependency of the measures. Examination of features and features combinations will also be addressed in this research, as well as techniques for enhancing features for desirable performances even in unfavorable conditions.

1.6. Scope of Research

This research lends its relevance to several applications including criminal detection and forensics, commercial services, military activities and terrorist identification.

1. Criminal Detection: In order to effectively manage inmate telephone privileges, three-way calling by inmates is prohibited in most federal prisons in the United States. It is almost impossible, however, for prison officials to simultaneously monitor all telephone conversations to determine if a three-way call has been placed. The implementation of automatic three-speaker detection systems could therefore be helpful in overcoming this problem, and this involves counting the number of speakers participating in telephone conversations, an aspect of speaker discrimination. Moreover, this research could be implemented in speaker tracking of inmate conversations to ensure that prison officials are notified when unidentified or suspicious individuals are contacted by inmates. Furthermore, automatic detection of keywords from recorded inmate telephone conversations could be helpful in determining violation of inmate telephone regulations. For instance, inmates are prohibited from performing any financial transactions over the phone; therefore, detection of a long string of numbers could indicate that the inmate has violated this rule. Speaker indexing of these recorded data could enhance or simplify the process of recognizing these specific keywords. In forensics, recorded telephone conversations are sometimes examined for criminal evidence, and speaker recognition systems could be effective in the process. A
database of prisoners’ voices could be created from their conversations, and used in developing voice-prints, which could serve as an alternative to fingerprints in forensics and/or even replace the use of fingerprints in future, especially since recorded telephone conversations are easily obtainable by law enforcement officials.

2. **Commercial Services**: with the vast technological advancements of the present day, several service providers that deal with automated telephone conversations require/implement automatic speech and speaker recognition technology for personalized contact with customers, as this generally increases positive consumer response.

3. **Military Activities**: in addition to telephone conversations, this research could also be applied to conversations or speech data that contain only short homogenous speaker utterances, such as air-related military conversations, pilot-control tower communications or detection of unidentified speakers on pilot radio channels.

4. **Terrorist Identification**: Terrorists’ telephone conversations are sometimes recorded by federal, state and local law enforcement agents and examine in order to prevent future terrorist attacks. These conversations could also be used in identifying suspected terrorist from their voices when other evidence is unavailable or insufficient.

### 1.7. Proposal Outline

This proposal is organized as follows: In Chapter Two, an in-depth presentation of the distance measures that will be considered in this research is given, and the performances of these distance measure in discriminating speakers are analyzed. A detailed analysis of some speaker recognition features is presented in Chapter Three, and an investigation of their speaker differentiating capabilities is also performed. In Chapter Four, a detailed description of process of forming data models to represent speakers is given, and, in the same chapter, speech data are also examined in order that portions of speech which are useful for speaker recognition are detected and extracted. In Chapter Five, correlation analysis of the distance measures is performed and the development of a novel distance which takes several distances into account.
is proposed. The development of application systems is described in Chapter Six. Finally, the research that has been completed to date is summarized in Chapter Seven, and an outline of future research plans is also given in the same chapter.
CHAPTER 2
ANALYSIS OF DISTANCE MEASURES FOR SPEAKER DISCRIMINATION

2.1. Introduction

This research involves the investigation of distance-based techniques for differentiating speakers. These distance measures include the Mahalanobis distance, Hotelling’s T-Square statistics, the Kullback Leibler distance, the Bhattacharyya distance and the Levene distance. All distances were chosen due to the fact that the feature sets used in representing speakers are multivariate random variables; therefore distance measures which take all variables into consideration are preferred since they generally utilize information not only from the mean values, but also from the covariance matrices of the feature vectors. The above mentioned distance measures have all been successfully applied in speaker recognition systems for applications such as SID [Ong and Yang, 1998], [Gish and Schmidt, 1994], [Iyer et al., 2006b], Speaker change-point detection and indexing [Delacourt and Wellekens, 2002], [Ofoegbu et al., 2006b], [Iyer et al., 2006b], and speaker count [Ofoegbu et al., 2006a].

In this chapter, the distance measures will be presented in detail, and their basic speaker differentiation performance will be analyzed and compared. The LPCCs will be used as features for this initial distance analysis; in subsequent chapters, all features examined in this research will be introduced, and the distance measures will be examined and compared using each of the features, and combinations of two or more amongst them.

The following notations will be applied. The random variables:
\[ \mathbf{X} = [X_1, X_2, \ldots, X_p] \] (2.1)
\[ \mathbf{Y} = [Y_1, Y_2, \ldots, Y_p] \] (2.2)

denote the two multivariate random variables - of lengths \( n_x \) and \( n_y \) and number of features equal to \( p \) - to be compared. For all distances to be valid, the feature vectors are required to have pdfs \( f_x(\mathbf{X}) \) and \( f_y(\mathbf{Y}) \) which follow the multivariate Gaussian distribution:

\[
f_x(\mathbf{X}) = \frac{1}{(2\pi)^{n_x/2} |\Sigma_x|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_x)^T \Sigma_x^{-1} (x - \mu_x) \right\}.
\] (2.3)

In order to be considered valid distance metric, each distance measure, \( Q(X, Y) \), between two random variables, \( \mathbf{X} \) and \( \mathbf{Y} \), their pdfs or their parameters \( \{\mu_x, \Sigma_x\} \) and \( \{\mu_y, \Sigma_y\} \), where \( \mu_x \) is the mean of the random variable and \( \Sigma_x \) is the covariance matrix, is required to and do satisfy the following properties:

\[ Q(X, Y) \geq 0, \] (2.4)
\[ Q(X, Y) = 0 \text{ iff } X = Y, \] (2.5)
\[ Q(X, Y) = Q(Y, X), \] (2.6)
\[ Q(X, Y) \leq Q(X, Z) + Q(Z, Y) \] (2.7)

2.2 Distance Measures

2.2.1. Mahalanobis Distance

The Mahalanobis distance [Mahalanobis, 1948] is simply a modified version of the Euclidean distance. Note, the Euclidean distance does not take into account the correlations of the dataset, and is sensitive to the scale of the measurements. With the Mahalanobis distance, on the other hand, each dimension is given a weight which is inversely proportional to its variance (in order words, the covariance matrices of the random variables are taking into consideration during the distance computation). The
Mahanalobis distance, which was one of the first distance measures applied in speaker recognition, is expressed as:

\[
Q_{\text{MAHANALOBIS}}(X,Y) = (\mu_x - \mu_y)^T \Sigma^{-1} (\mu_x - \mu_y) \tag{2.8}
\]

where \(\Sigma\) is the covariance matrix of the two random variables combined.

\subsection*{2.2.2. Hotelling’s T-Square Statistics}

The Hotelling’s T-Square Statistic [Manly, 1994] is a multivariate generalization of the t-test, which is commonly used in comparing the means of two univariate random variables. Hotelling’s T-Square Statistic can be expressed as:

\[
Q_{\text{TSQ}}(X,Y) = \frac{n_x n_y}{n_x + n_y} \sum_{i=1}^{p_x} \sum_{k=1}^{p_y} (\bar{x}_i - \mu_x) (\bar{y}_i - \mu_y) C_{ik} (\bar{x}_k - \mu_x) (\bar{y}_k - \mu_y). \tag{2.9}
\]

where \(C_{ik}\) is the element in the \(i^{\text{th}}\) row and \(k^{\text{th}}\) column of the inverse of \(C\), the pooled estimate of the covariance matrix for both populations, expressed as:

\[
C = \frac{(n_x - 1) \Sigma_x + (n_y - 1) \Sigma_y}{n_x + n_y - 2} \tag{2.10}
\]

Note that the T-Square statistic is simply the square of the T-test, thereby taking into account the correlation of all features in the set simultaneously. It is also a scaled representation of the Mahanalobis distance; the scaling factor is derived from the sizes of the two random variables. Significantly large values of the T-Square statistic indicate more separation between the feature sets being compared.

\subsection*{2.2.3. Kullback Leibler Distance}

The Kullback Leibler distance, which is used mainly in information theory and pattern recognition, belongs to a class of distance measures which compute the separation of two pdfs based on the dispersion of the
likelihood ratio with respect to one of the densities [Basseville, 1989]. The distances in this class are derived from the equation:

\[
Q(X, Y) = g\left( E_x \left\{ f\left( \frac{p_y(x)}{p_x(x)} \right) \right\} \right)
\]  

(2.11)

where \( g \) is a function which is continually increasing on the Real Line, \( E_x \) is the expectation of the random variable \( X \), and \( f \) is a continuous convex function on the positive Real Line. The Kullback Leibler distance is derived by assigning:

\[
f(x) = (x-1) \log(x), \quad \text{and} \quad g(x) = x
\]

(2.12)

(2.13)

resulting in the expression:

\[
Q_{\text{KULLBACK}} = \int_S (p_y(\xi) - p_x(\xi)) \log \left( \frac{p_y(\xi)}{p_x(\xi)} \right) d\xi,
\]

(2.14)

where \( S \) denotes the probability space spanned by the feature vector sets compared. When the random variables are assumed to be Gaussian (Equation (2.3)), the expression of the Kullback Leibler (KL) distance becomes:

\[
Q_{\text{KULLBACK}} = \frac{1}{2} (\mu_y - \mu_x)^T (\Sigma_x^{-1} + \Sigma_y) (\mu_y - \mu_x) + \frac{1}{2} \text{tr}(\Sigma_x^{-1} \Sigma_y + \Sigma_y^{-1} \Sigma_x - 2I)
\]

(2.15)

where \( I \) is the identity matrix. It can be observed that under the assumption of equal covariance matrices for both feature sets, the KL distance is equivalent to the Mahanalobis distance.

### 2.2.4. Bhattacharyya Distance

The Bhattacharyya distance [Bhattacharyya, 1943] belongs to the same class as the KL distance. In this case however, the functions \( f(x) \) and \( g(x) \) in Equation (2.11) are expressed as:
The general form of the Bhattacharyya distance is given by:

\[
Q_{BHATTACHARYYA} = \log\left(\rho(p_x(x), p_y(y))\right),
\]

(2.18)

where

\[
\rho(p_x(x), p_y(y)) = \sqrt{p_x(\xi)p_y(\xi)d\xi}
\]

(2.19)

represents the Bhattacharyya coefficient (geometrically interpreted as the cosine of the angle between the pdfs of the two random variables being compared) which measures the amount of separation between the two feature vectors. If the random variables are assumed to be Gaussian, the Bhattacharyya distance may be expressed as:

\[
Q_{BHATTACHARYYA} = \frac{1}{4}(\mu_y - \mu_x)^T (\Sigma_x + \Sigma_y)^{-1} (\mu_y - \mu_x) + \frac{1}{2} \log\left(\frac{\Sigma_x + \Sigma_y}{2\sqrt{\Sigma_x \Sigma_y}}\right).
\]

(2.20)

### 2.2.5. Levene’s Test

Levene’s Test [Levene, 1960] is a robust statistical distance measure which can be derived from the T-Square statistics [Manly, 1994]. The basic idea of this measure is a comparison of variation of points between the two feature sets, and the concept behind the method is to determine if there is a notable difference between the mean deviations of the two random variables being compared using a t-test, after initially converting the original data to absolute deviations from the median [Schultz, 1983]. This distance, denoted by \(d_{LEVENE}\), is computed using a two-step procedure as follows:
Step 1: Each set of points is transformed along each vector into absolute divergence from the mean vector such that the variation comparison is performed by simply comparing the means of the transformed feature vectors.

Step 2: The T-Square Statistic is then applied on the transformed features.

2.3. Analysis of Distance Measures for Speaker Discrimination

In this section, the performances of the distance measures in distinguishing speakers are investigated by observing the differences between intra-speaker and inter-speaker distance values. Pdfs of these distances are obtained from the data, plotted and analyzed.

2.3.1. Procedural Set-up

The HTIMIT database, which consists of 10 utterances each from 384 speakers (192 male and 192 female), recorded over 3 different telephone channels (Reynolds, 1997) was used for the initial analysis of the distance measures. Each utterance is about 5 seconds in length, and two of the 10 utterances were the same for all the speakers. A telephone-based speech database was used because this research is focused on speaker discrimination for telephone conversations. Fourteenth (14th) order LPCCs (which will be discussed in further detail later in this proposal) were used as speaker differentiating features. Distances measures were examined as follows for three different cases.

1. Intra-Speaker Distances: This involved computing distances between features from different utterances from the same speaker. For each speaker in the HTIMIT, the intra-speaker distance was computed as follows:
   (i) First, two utterances were selected at random from each of the 10 utterances
   (ii) The utterances were then broken down into frames of 30 milliseconds with no overlap.
   (iii) 14th order LPCCs were computed for each frame for both utterances
   (iv) The distance between utterances was computed using each of the distance measures.
The intra-speaker distance computation procedure is illustrated in the block diagram shown in Figure 2.1.

**Figure 2.1:** Block diagram showing the intra-speaker distance computation procedure

2. **Inter-Speaker, Same Utterance Distances:** In this case, distances were computed between features from different speakers saying the same sentence. The inter-speaker, same utterance distances were computed as follows:

   (i) For each speaker in the HTIMIT database, one utterance was chosen from the two identical utterances for all speakers.

   (ii) A different speaker was then chosen at random from the same database, with care taken to ensure that the same speaker was not chosen; the same utterance used in step (i) above was selected for this second speaker.

   (iii) The utterances were then broken down into frames of 30 milliseconds with no overlap.

   (iv) 14th order LPCCs were computed for each frame for both utterances.

   (v) The distance between both utterances was computed using each of the distance measures.

The above procedure is illustrated in the block diagram shown in Figure 2.2.

**Figure 2.2:** Block diagram showing the inter-speaker, same utterance distance computation procedure
3. **Inter-Speaker, Different Utterance Distances**: In most cases (especially those with which this research is concerned) different utterances from different speakers will be available for comparison. Hence a third type of distances, computed between features from different utterances from different speakers was observed. These distances were computed as follows:

(i) For each speaker in the HTIMIT database, one utterance was chosen at random.

(ii) A different speaker was then chosen at random from the same database, with care taken to ensure that the same speaker was not chosen; and an utterance from this speaker was chosen, also at random.

(iii) The chosen utterances were then broken down into frames of 30 milliseconds with no overlap.

(iv) 14th order LPCCs were computed for each frame for both utterances

(v) The distance between both utterances was computed using each of the distance measures.

The above procedure is illustrated in the block diagram shown in **Figure 2.3**.

![Figure 2.3: Block diagram showing the inter-speaker, different utterance distance computation procedure](image-url)

### 2.3.2. Distance Analysis

The probability density functions (pdfs) of the distances obtained using the procedures illustrated in **Figures 2.1 – 2.3** were estimated based on their histograms, and the pdfs for the three cases discussed above were plotted against one another.
The histograms of the Mahalanobis distances for the intra-speaker distances; inter-speaker, same utterance distances; and inter-speaker different utterance distances, as shown in Figure 2.4 were observed to fit the Gaussian distribution with parameters \( \{\mu, \sigma\} \) expressed as:

\[
f_x(X) = \frac{1}{\sigma \sqrt{2\pi}} \exp \left( -\frac{(x - \mu)^2}{2\sigma^2} \right)
\]

(2.21)

The intra-speaker distances are indexed SSDU (Same Speaker Different Utterances), while the inter-speaker, same utterance distances and inter-speaker different utterance distances are defined as DSSU (Different Speaker Same Utterance) and DDSU (Different Speaker Different Utterances), respectively.

\[\text{Figure 2.4: Probability Density Functions for Intra-Speaker (SSDU), inter-speaker same utterance (DSSU) and inter-speaker different utterances (DSU) Mahalanobis distances.}\]

From Figure 2.4, it can be observed that a significant amount of separation exists between intra-speaker (SS) and inter-speaker distances (DS). Regardless of the fact that the same speech is uttered by each speaker for the same-speaker inter-speaker distances, their distribution are seen to be reasonably distinguishable, indicating that the Mahalanobis distance is a reliable speaker discriminating distance measure. The inter-speaker different utterance case (DSU), which is the case most likely to occur, shows
even further separation from the intra-speaker distances. These observations provide motivation for the use of the distance for this research; nevertheless, further analysis of the data suggests that the speaker discriminative capability of the distance can be enhanced. This will be discussed in subsequent chapters of this proposal.

The estimated pdfs for the T-Square distances are shown in Figure 2.5. The distances were observed from their histograms to follow the Gamma distribution expressed as:

\[
\gamma = f(x | a, b) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{x/b}
\]  

where \(\Gamma(a)\) is the Gamma function evaluated at \(a\) and the parameters \(a\) and \(b\) are given by the equations:

\[
a = \frac{\mu^2}{\sigma^2}; \quad b = \frac{\sigma^2}{\mu}
\]

\(\mu\) and \(\sigma\) are the standard representations of the mean and standard deviation of the distances.

![T-Square Statistic Comparisons](image)

**Figure 2.5**: Probability Density Functions for Intra-Speaker (SSDU), inter-speaker same utterance (DSSU) and inter-speaker different utterances (DSDU) T-Square statistics.

The KL and Bhattacharyya distance histograms were observed to follow a Gaussian distribution, and the pdf estimates computed using Equation (2.21) are shown in Figures 2.6 and 2.7.
For the both the distances, the intra-speaker distance and inter-speaker same speaker distance pdfs are practically indistinguishable as can be noted from Figure 2.6 and 2.7. The separation between the intra-speaker and inter-speaker different utterance distances is also not as significant as in the Mahalanobis
distance and T-Square statistics. The question now becomes, “of what good are these distances, then?” As will be proven, both distances aid in distinguishing speakers when combined with other distances, as they do offer some speaker discrimination (especially when the data is enhanced). Moreover, since the derivation of these distances is based on different factors than the other distances, appropriate combinations of the distances could result in the exploitation of different information which could yield much higher speaker differentiation performance.

Like the T-Square statistics, the Levene’s Test histograms were observed to best fit the Gamma distribution, and their pdfs were estimated using Equation (2.22) and (2.33). Note that this similarity is inevitable seeing that the computation of the Levene’s Test relies heavily upon the T-Square statistics. The estimated pdfs of the Levene’s Test values are shown in Figure 2.8 below.

![Figure 2.8: Probability Density Functions for Intra-Speaker (SSDU), inter-speaker same utterance (DSSU) and inter-speaker different utterances (DSDU) Levene’s Test values.](image)

Figure 2.8 above clearly suggests that Levene’s Test possesses hardly any ability to distinguish between speakers whether the speech uttered by both speakers are the same or different. This lack of separability is, however, also investigated in this research, and data enhancements for the improvement of the performance of Levene’s Test, as well as combination of Levene’s Test with other distance measurements, is proposed.
In this chapter, the distance measures to be considered in this research were introduced, and an analysis of their basic speaker discrimination capabilities was performed. The LPCCs were used as features for this initial analysis; however, analyses of several features will be carried out in the next chapter using all the distances. Feature and data enhancement techniques will also be examined and proposed in subsequent chapters.
CHAPTER 3
ANALYSIS OF FEATURES FOR SPEAKER DISCRIMINATION

3.1. Introduction

In this chapter, the use of different features and their combinations is explored for effective speaker discrimination performance. As discussed in Chapter One, several features have been successful exploited for speaker recognition; however, the concentration in this research will be on low level features such as cepstral features – MFCCs, LPCCs and their 1st and 2nd derivatives. These features were chosen because, since they are low level features, they are easy to compute. Moreover, the cepstral features are widely known to yield outstanding speaker recognition performances and are still used in state-of-the-art speaker recognition systems [Ferrer et al., 2006]. A detailed introduction of each feature set and its computation is presented in this chapter, as well as a basic analysis of their speaker discriminative capabilities using the distances described in Chapter Three. Also, investigation of the performances these features when combined with one another will also be carried out.

It must be noted that this chapter does not deal with data enhancements for the improvement of feature performances. Such topic will be covered in Chapter Four. This chapter is merely concerned with presenting and comparing the basic performances of the features in distinguishing speakers. Experiments are carried out as outlined in Chapter Three, with the only difference being the features that are used.
3.2. Cepstral Features

3.2.1. Cepstral Analysis

The word "Cepstrum" was derived from the word “Spectrum”, and is used to represent the process of observing frequency domain signals as though they were time domain signals. This process, which was introduced by Borget et al., (1963), involves a transformation of the data such that the x-axis values, which depict the variations in the frequency spectrum of the signals, are in units of seconds. These values are referred to as quefrencies. The cepstrum characterizes the periodicity of the frequency response of a signal.

The derivation of cepstral values is as follows. Let $S_n$ be the speech signal being observed, and $E(e^{i\theta})$ and $H(e^{i\theta})$ be the excitation and linear filter components obtained from the source decomposition of the observed signal [Deller et al., 2000]. A frequency domain representation of the speech signal can be expressed as:

$$S(e^{i\theta}) = H(e^{i\theta})E(e^{i\theta})$$  \hspace{1cm} (3.1)

Here the envelope of the power spectrum is symbolized by the linear filter $H(e^{i\theta})$, and in speech processing, the coefficients of this filter are considered to contain information about the formants, which contain vocal tract information that characterize the individual speakers.

Taking the natural logarithm of Equation (4.1) above, the following equation is obtained:

$$\log(|S(e^{i\theta})|) = \log(|H(e^{i\theta})|) + \log(|E(e^{i\theta})|)$$  \hspace{1cm} (3.2)

With this transformation, the formant information, which is represented by slow varying components, are contained in the low frequencies, while the excitation (fast varying) components are contained in the high frequencies. These two distinct components can therefore be separated by taking the Inverse discrete Fourier Transform (IDFT) of the log spectrum, and “liftering” (a term coined to represent the filtering procedure in the “time domain”). This process is illustrated in Figure 4.1 below, which is an adaptation from [Deller et al., 2000].
Figure 3.1: Illustration of cepstral analysis. The circle represents the desired vocal tract information obtained by liftering. (Adapted from [Deller et al., 2000].)

Figure 3.1 also illustrates the motivation behind cepstral analysis, which involves converting the problem of a nonlinear combination between the excitation and vocal tract information to a linear problem, thereby enabling the two components to be more easily separated.

The derivations of the LPCCs and MFCCs are based on this analysis, even though their computations do not directly portray the computational steps explained above. The computations of the features are explained in the following two sub-sections.

3.2.2. Linear Predictive Cepstral Coefficients

The LPCCs are derived from a combination of cepstral analysis (illustrated in Figure 3.1) and linear prediction analysis. Linear prediction analysis is applied extensively in speech processing today. It is based on a source-filter model of speech signals in which the filter is assumed to be an all pole linear filter which represents the vocal tract of a person [Makhoul 1975]. The concept behind linear prediction is the
estimation of speech sample, \( s(n) \), using a weighted linear combination of \( p \) previous samples of the same signal, and is given by:

\[
\tilde{s}(n) = \sum_{i=1}^{p} w_i s(n - i)
\] (3.3)

where \( w \) is the weighting function. This yields an all-pole, \( p \)th order Finite Impulse Response (FIR) filter representation of the vocal tract, with coefficients equal to \( w \). [Quatieri, 2004]. The prediction error can then be obtained as:

\[
e(n) = s(n) - \sum_{i=1}^{p} w_i s(n - i)
\] (3.4)

The prediction coefficients can be determined by minimizing \( e(n) \) for short speech frames. The error minimization is generally carried out using the Levinson-Durbin recursive procedure [Haykin 2002], which was first proposed by [Levinson, 1947] and then modified by [Durbin, 1960]. This technique exploits the Toeplitz structure [Gray, 2006] of the correlation matrix of the speech samples (which are input to the all-pole filter) in determining the minimum error solution for an \( i \)th order filter using information from the \((i-1)\)th order predictor [Quatieri, 2002].

In choosing the model order for the all-pole filter, the vocal tract, the source and the radiation from the lips are generally considered as the important components. Orders ranging from about 8 to 14 have been explored in LPC modeling of speech [Quatieri 2002], [Ojala and Lakaniemi, 1998], [El-Jaroudi and Makhoul, 1991]. The key is to utilize the minimum order which will give sufficient information about the speech or speaker, as desired.

The LPCCs are obtained recursively from the Linear Prediction Coding (LPC) coefficients as follows:

Let \([a_0 \ a_1 \ a_2 \ldots \ a_p]\) be the LPC coefficients with order equal to \( p \); the LPCCs of order \( m \) are expressed as:
\[ c_0 = \ln E^2 \]  

\[ c_m = a_m + \frac{1}{m} \sum_{k=1}^{m-1} (m-k) a_k c_{(m-k)} \]  \( 1 \leq m \leq p \)  

\[ c_m = \sum_{k=1}^{m-1} \frac{-(m-k)}{m} a_k c_{(m-k)} \]  \( p \leq m \leq N \)

where \( E \) is the energy of the signal.

### 3.2.3. Mel-Scale Frequency Cepstral Coefficients

The MFCCs were motivated by the idea of building models based on how the human auditory system analyzes of speech. This is standard in most speech processing applications. The Mel frequency scale, depicted in Figure 3.2, was introduced by Davies and Mermelstein (1980), and is characterized by its ability to selectively weight the frequencies in the power spectrum of the signals such that the weights of the low order cepstral coefficients, which are affected by the entire spectral slope, are generated on a logarithmic scale. The high-order cepstral coefficients are assigned weights which are generated on the linear scale since the coefficients are more sensitive to noise than to the spectral slope. The Mel cepstrum is obtained by the following steps [Quatieri, 2002]:

Step 1: The discrete Short-time Frequency Transform (STFT) of frames (or windows) of the speech signal are computed as:

\[ F(n, \omega_k) = \sum_{m=-\infty}^{\infty} s(m) w(n-m) \exp(-j m \omega_k) \]  

where \( s(m) \) is the speech waveform, \( w(n) \) is the windowing function and \( \omega_k = (2\pi k)/N \), with \( N \) being the STFT length.
Step 2: The magnitude of the STFT is then passed through a Mel-scale frequency filter bank which is comprised of a series of the filters that follow the Mel-frequency scale. The Mel-scale frequencies are computed as:

$$mel(f) = 2595 \log_{10}(1 + f/700)$$

(3.9)

This function is illustrated in Figure 3.2 for the frequency range of 0-10KHz

![Logarithmic Plot of Mel-Scale Frequency](image)

**Figure 3.2**: Mel-scale Frequency Function

Step 3: The energy of the output of each Mel-Scale filter is then computed using Equation (3.10), where $\eta_i(\omega)$ represents the frequency response of the $i^{th}$ filter and $A$ and $B$ denote the lower and upper indices over which each Mel-scale filter possesses a nonzero value.

$$E(n,i) = \frac{1}{\lambda_i} \sum_{k=A_i}^{B_i} |\eta_i(\omega_k)F(n,\omega_k)|^2$$

(3.10)
The parameter $\lambda$ is a normalizing function for the filters which ensures that the energy obtained when the input is a flat spectrum input is equal for all filters. This function is expressed as:

$$\lambda_i = \sum_{k=A_i}^{B_i} |\eta_i(\omega_k)|^2$$  \hspace{1cm} (3.11)

Step 4: The frame-by-frame Mel cepstrum, represented by the real cepstrum computed from the Mel-scale energy, is then determined. Taking advantage of the even property, the exponential function in the inverse transform can be replaced by the cosine function, which results in the Discrete Cosine Transform (DCT), which is preferable in speaker recognition procedures [Quatieri, 2002], as it bears significant resemblance to the Karhunen-Loeve transform [Zelinski and Noll, 1977] which possesses the desired feature of being able to decorrelate the cepstral coefficients. The Mel cepstrum is computed using the DCT as:

$$C(n,m) = \frac{1}{R} \sum_{i=0}^{R-1} \ln\{E(n,i)\} \cos(2\pi i m / R)$$  \hspace{1cm} (3.12)

where $R$ is the number of filters.

### 3.2.4. Delta Cepstral Coefficients

In differentiation of speakers, a desired characteristic of features is channel invariance, which is not a property to be found in instantaneous features such as the cepstral coefficients. Features which reflect dynamic information about the speech signals, such as the speaking rate, are usually channel invariant and are valuable in speaker recognition systems. These dynamic characteristics can be obtained from the first and second derivatives of the cepstral coefficients, which are usually referred to as the delta and delta-delta cepstral coefficients, respectively. [Furui, 1981], [Soong and Rosenberg, 1985]. In order to simplify the computation of the delta and delta-delta coefficients, polynomial approximations of the first and second derivatives of the cepstral coefficients, $c_{m}$ are estimated as [Bimbot et al., 2004]:
\[ \Delta c_{m,n} = \frac{\sum_{k=-P}^{P} k c_{m,n+k}}{\sum_{k=-P}^{P} |k|} \]  

(3.13)

\[ \Delta \Delta c_{m,n} = \frac{\sum_{k=-P}^{P} k^2 c_{m,n+k}}{\sum_{k=-P}^{P} k^2} \]  

(3.14)

where \( P \) is the order of the polynomial (classically 9) and \( n \) is the index of the current frame. The delta and delta-delta coefficients are generally used in combination with the cepstral coefficients in order to enhance their discriminative capabilities by exploiting the ability of the delta coefficients to capture dynamic information.

### 3.3. Proposed Work: Analysis of Features for Speaker Discrimination

In the previous chapter, the distances introduced were analyzed using the LPCCs as features. The same analysis is proposed in this research for the MFCCs, as well as various combinations of the LPCCs, MFCCs their delta and delta-delta coefficients. Combination of the features will be performed such that comparisons of the combined sets will enable the determination of:

1. Whether or not the combination of both feature sets will result in an improvement in speaker discrimination as compared with when only one feature set with the same number of coefficients is used.

2. Whether or not the delta and delta-delta coefficients contribute significantly to the speaker differentiating ability of the features.
In subsequent analyses of the speaker recognition distances and features, inferences were made by observing the distributions of the intra-speaker and inter-speaker distances. In order to make valid comparisons, however, a numerical value for the separability of the distributions for each feature-set and distance is proposed. One way to assign this numerical value is the T-test, which is a measure of similarity between two sample means. Some properties of the T-test include:

1. It is robust to the Gaussian distribution especially for large sample sizes (20 or more samples) and when the two samples to be compared have approximately equal values. [Carter et al., 1979].
2. Although there is an equal variance assumption, differences in the variances of the two compared samples will not significantly affect the performance of the test provided that the ratio of their actual variances are within a ratio range of 0.4 to 2.5 [Manly 1994].

These properties are very favorable to this analysis due to the fact that not all the distances can be approximate using the Gaussian distribution, as explained earlier, and as could be observed from Figures 2.5 – 2.8, the variances of the intra-speaker and inter-speaker distances are not always equal. Nevertheless the sample sizes to be compared are always equal and the ratio of the two sample variances fall mostly within the range specified in (2) above.

Let \( \mu_1 \) and \( \mu_2 \) represent the means of the intra-speaker and inter-speaker distances, \( n_1 \) and \( n_2 \) represent their length (which is 384 in each case) and \( \sigma_1^2 \) and \( \sigma_2^2 \) represent their variances. The T-test can be expressed as:

\[
T = \frac{\mu_1 - \mu_2}{\sqrt{s^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}
\]  
\[
(3.15)
\]

where \( s^2 \), a pooled estimate of the variances of the two distributions, is given by:

\[
s^2 = \frac{(n_1 - 1)\sigma_1^2 + (n_2 - 1)\sigma_2^2}{n_1 + n_2 - 1}
\]  
\[
(3.16)
\]

A higher value of \( T \) indicates greater separation between the classes being compared.
Evaluation of the speaker discrimination performance of the LPCCs, MFCCs and various combinations of the features with their delta and delta-delta coefficients will be performed using the T-test and other tests of dissimilarity which will indicate how separated the intra-speaker and inter-speaker distance measurements are. Enhancement of feature combinations using Principal Component Analysis is also proposed.
CHAPTER 4
DATA ANALYSES FOR TELEPHONE CONVERSATIONS

4.1. Introduction

In this chapter, results of the investigation of data enhancement techniques with specific application to telephone conversations or other speech utterances that involve short speaker segments will be presented. Also, the idea of speech segmentation and voiced speech detection is presented, along with its role in speech enhancement. The formulation of data “models” for speaker discrimination in telephone conversations will also be investigated. Finally, the concept of usable speech will be discussed in relation to this research, and some existing usable speech extraction techniques will be presented. The effect of eliminating “unusable” speech from speech signals will also be examined.

4.2. Speech Classification and Extraction of Voiced Speech

4.2.1. Speech Classification

Speech is generally classified into several categories depending on the nature of the speech source (whether it is periodic, noisy, impulsive or all three combined), the shape of the vocal tract, the time domain waveform, and the spectrogram. [Quatieri 2002]. However, the most common classification of speech, which takes into consideration all the factors listed above, is the voiced-unvoiced classification. Voiced speech, is produced by an air flow of pulses caused by the vibration of the vocal cords. The resulting signal
could be described as quasi-periodic waveform with high energy and high adjacent sample correlation. On the other hand, unvoiced speech, which is produced by turbulent air flow resulting from constrictions in the vocal tract, is characterized by a random aperiodic waveform with low energy and low adjacent sample correlation.

Due to their noise-like nature, information from unvoiced speech has been shown to offer little or no useful contribution to speaker recognition systems. As a matter of facts, studies have shown that speaker recognition systems are sometimes improved by eliminating unvoiced portions speech before processing [Benicassa and Savic 1998], [Yantorno, 1998], [Lovekin, et al., 2001a]. In practice, however, the entire speech signals are generally used in speaker recognition systems, without any removal of unvoiced speech, since it results in a more complex system, and may not yield any significant improvement. Moreover, for applications (such as speaker identification) which involve a considerable amount of data, having unvoiced speech might actually yield a slight increase in the performance of the system [Iyer 2006c].

4.2.2. Voiced-Unvoiced Classifiers

Several voiced-unvoiced classifiers have been developed over the past several decades. The most common measure is the frame-by-frame energy measure [Atal and Rabiner, 1976], which is based on the difference in amplitude (and therefore, energy) between voiced and unvoiced speech. Another traditional measure, the zero-crossings approach, also developed by Atal, et al. (1976), involves counting the number of times the signal crosses the x-axis, and is based on the knowledge that unvoiced speech signals, being more noise-like in nature, oscillate much faster than voiced speech signals. Therefore, the zero-crossing rates of voiced signals should be lower than those of unvoiced signals. Voiced speech detection has also been developed using the first order reflection coefficient and the residual energy of the speech signals [Childers, 2000]. The reflection coefficient, obtained by modeling the vocal tract as a concatenation of tubes, determines the amount of volume-velocity reflection that can be found at the intersection of two tubes. Due to its high energy, voiced speech possesses a high amount of volume-velocity as compared to unvoiced speech. Significant information in speech is usually contained in the first coefficient, hence the use of the first order
reflection coefficient. The residual energy measure was also developed by Childers (2000). And is defined as the energy of the signal that has been inverse-filtered using the LPC (Linear Predictive Coding) coefficients. The chaotic nature of an unvoiced speech signal results in a low residual energy as compared to a voiced speech signal. Although all these measures, and a combination of some of them with one another, have been shown to successfully classify speech into voiced and unvoiced speech, more complex, state-of-the art classification systems have been developed, which take into account the nonlinearities and dynamics of the speech signals, making them more robust to noise and other interferences. Some of these classifiers involve the application of clustering analysis and Principal Component Analyses (PCA) on some of the traditional characteristic features extracted from the data [Smolenski, 2005]. Another recently developed voiced-unvoiced classifier involves the exploitation of Taken’s embedding theorem [Takens, 1981], which states that a state space representation topologically equivalent to the original state space of a system can be obtained from a single dimension. This theorem has been applied in several signal processing applications including speech processing [Terez, 2002], [Schreiber, 1995], [Kantz and Schreiber 1998]. Several speech classification measures have been developed based on this theorem [Ofoegbu 2004], [Ofoegbu et al., 2004]. In this research, voiced speech classification is performed using one of such measures, the curvature measure [Smolenski, 2004], which was developed using the Serret-Frenet theorem [Rahman & Mulolani, 2001]. A detailed description of voiced-unvoiced classification process via the curvature measure is given in [Ofoegbu, 2004], [Ofoegbu et al., 2004].

4.3. Formation and Analysis of Data Models

In telephone conversations, speakers change rapidly and speaker change-points are seldom known a priori (except it specifically determined). Therefore, in this research, speakers are represented on a phoneme basis, as it is quite improbable for two speakers’ speech to be contained in the same voiced phoneme. Speaker homogeneous utterances are usually expected to consist of more than one voiced phoneme. Models are formed from speech segments created by concatenating $N$ consecutive voiced phonemes, with $N > 2$, as illustrated in Figure 4.2 (for $N = 3$).
In determining the number of segments used in forming the homogeneous speaker segments (or ‘models’), one has to take into account the application of the system. In general, the greater the data size of the speech utterances compared, the better the speaker recognition performances. Nevertheless, as mentioned earlier, some applications do require that short data lengths be used in order to avoid overlapping speakers in the same model [Ofoegbu et al., 2006b].

**Figure 4.2** shows the intra-speaker and inter-speaker distributions for $N$ (Number of voiced phonemes) = 20. The Mahalanobis distance was used and the LPCCs were used as features. Each homogeneous speaker model was obtained by concatenating 20 consecutive voiced phonemes from as many utterances of the same speaker as was required to yield 20 phonemes. 1000 intra-speaker and inter-speaker distances were computed.

A clear separation can be observed between the intra-speaker and inter-speaker distances for $N = 20$. The same trend was observed for all distance and with the LPCCs used as features (figures are not shown).
Figure 4.3: Distances obtained using homogenous speaker models formed from 20 voiced phonemes.

It is, however, very impractical to assume 20 voiced phonemes will be available for each homogeneous speaker utterance in telephone conversations. Note that the 20 phonemes had to be extracted from three or more 5-second utterances of the same speaker from the HTIMIT database. In practical conversations, the speaker’s utterances are generally about 2 seconds in length on average [Iyer et al., 2006b]. It is therefore important to determine an appropriate number of phonemes, which would yield sufficient differentiation between intra- and inter-speaker distances, and also prevent grouping together of segments from two different speakers to form one model. In other words, the least number of segments with adequate separation is desired.

Figure 4.4 shows the means (circles) and standard deviations (horizontal bars) of the Mahalanobis distances for the intra- and inter-speaker comparisons, and it illustrates the effect of data size on speaker discrimination. One thousand comparisons were observed for each value of \(N\) (number of segments). Each voiced segment (or phoneme) was of an average length of 200 milliseconds. From Figure 2, it is observed

---

**Figure 4.4**: Means and standard deviations of the Mahalanobis distances for intra- and inter-speaker comparisons.
that an increase in data size results in an increase in speaker separability. Additionally, all values of \( N \) below 5 result in an overlap in the standard deviations of the intra- and inter-speaker distances; thus, in this research, 5 segments (resulting in a total length of about 1 second) are used in forming speaker models.

**Figure 4.4:** Comparison of Mahalanobis Distances for different data sizes. X-axis represents the number of segments used to form each model.

**Figure 4.5** shows the intra-speaker and inter-speaker distributions for \( N = 5 \) using the same parameters as in **Figure 4.3**. In this case, the separation between the intra-speaker and inter-speaker distances is shown to have decreased significantly from what was obtained for \( N=20 \), as expected, however, a considerable amount of separation does exist. For quantitative purposes, it might be worthwhile to note that the T-value between the two classes (intra- and inter-speaker) is about 44 for \( N = 5 \) and 85 for \( N = 20 \).

In this research, techniques to compensate for this performance degradation due to data unavailability are proposed.
Figure 4.5: Distances Obtained using homogenous speaker models formed from 5 voiced phonemes.

4.4. Usable Speech Analysis

Speech produced when two or more speakers are talking simultaneously through the same communication channel is known as co-channel speech. Instances of co-channel speech are very commonly encountered in telephone conversations, where speech utterances from different speakers tend to overlap. The different speaker’s speech can easily be separated by the human auditory system, but automated speech processing systems are usually not as effective. As a result, the performances of speaker recognition systems are generally degraded in the presence of co-channel speech, as information from more two (or more) speakers are sometimes used in forming speaker models, thereby rendering the features incapable of distinguishing the various speakers. Previously, the problem of co-channel speech was resolved by enhancing the portions of speech produced by the desired speaker and repressing the unwanted portions [Hanson and Wong, 1984] [Morgan, et al., 1997]. In order for this to be performed, the sinusoidal components of each speaker had to be obtained using least squares estimation. However, this method failed to provide accurate separation of
co-channel speech when the frequencies of both speakers were not sufficiently different, moreover, a priori knowledge of the pitch of the speakers, which was not always available, was required [Quatieri and Danisewicz, 1988].

The concept of “usable speech” was presented by [Yantorno, 1998] as a solution to the co-channel speech problem. The idea of usable speech is derived from the fact that not all portions of speech corrupted by co-channel interference are unusable for speech processing techniques. When the energies of the two overlapping utterances are approximately equal, certain portions still exist in co-channel speech in which the energy of one speaker is greater than the energy of the other speaker. These portions are termed “usable” while the other portions (with both high energies overlapping) are termed “unusable”. The use of only ‘usable’ portions of speech has been shown improve the performance of speaker identification systems [Lovekin, et al., 2001b], [Iyer et al., 2004]. With co-channel speech, an accuracy of about 40% was obtained in speaker identification [Yantorno, 1998], however, research has shown that with the use of extracted usable speech from co-channel speech, the accuracy of the speaker identification system could increase significantly (up to 90%) [Lovekin et al., 2001b], [Iyer et al., 2004].

### 4.4.1. Target to Interferer Ratio (TIR) Based Usable Speech

Co-channel speech can be labeled usable or unusable based on the ratio of the energy of the target speaker to that of the interfering speech, i.e. the target-to-interferer ratio (TIR) – the target being the speech from the speaker of interest, and the interferer being the interference). As mentioned earlier, usable speech occurs when the energy of one speaker is higher than that of the other, for instance, when the voiced speech from one of the speakers occurs when there is either unvoiced speech or occasional speech breaks (silence) from the other speaker. Speech segments having TIR magnitude above a pre-determined threshold (generally about 20dB) can be considered usable for further speech processing applications such as speaker identification. It must be noted that TIR is only available in artificial situations (for training and testing the system) but never in co-channel speech situations such as telephone conversations or conferences. Some existing usable speech detection techniques which detect usability using measures that correlate with TIR.
include [Krishnamachari, et al., 2000], [Yantorno et al., 2000], [Yantorno et al., 2001], [Lovekin, et al., 2001b], [Iyer et al., 2004], [Krishnamachari et al., 2001], [Kizhanatham and Yantorno, 2002], [Kizhanatham et al., 2003], [Smolenski, et al., 2002a], [Smolenski, et al., 2002b], [Sundaram et al., 2003], [Chandra and Yantorno, 2002], [Iyer et al., 2003], [Ofoegbu et al., 2004].

4.4.2. Application Based Usable Speech

A different approach to determining usability in co-channel speech was introduced by [Iyer et al., 2004], and also investigated by [Khanwalkar et al., 2004], [Khanwalkar et al., 2005]. The technique involves labeling speech frames as usable or unusable based on the output of the Speaker Identification (SID) system. Speech frames, used by the speaker identification system, that result in a sufficiently close match when compared to the model of the target speech were considered usable, while the others were considered unusable. This usability criterion was determined by first computing the distances between each frame of speech to the trained models, and then defining speech frames with distances smaller than an assigned threshold as usable. Using this approach, the differences in the normalized distances between the test vector and the best two speaker models are utilized in usability determination. Although this technique was designed for speaker identification, the same idea can be extended to other speaker recognition applications such as speaker indexing, where usability can be defined based on the system’s performance, and the usability detection can be evaluated using groundtruth information. The application-based usable speech was recently investigated for speaker indexing, and results proved that it was able to improve the accuracy of the indexing system [Ofoegbu et al., 2006b]. Detection and removal of unusable co-channel speech for the enhancement of speaker indexing systems will be addressed in the next chapter, which deals with the development of application systems involving telephone conversations.
CHAPTER 5

PROPOSED WORK: DEVELOPMENT OF APPLICATION SYSTEMS

5.1 Introduction

In this chapter, the development of a speaker indexing and a speaker count system for telephone conversations is proposed. These systems are based on the distance computation, features and speech enhancement techniques presented in preceding chapters. Evaluations of the systems after they have been fully developed will be performed using artificial conversations from the HTIMIT database and actual telephone conversations from the SWITCHBOARD conversations database [Godfrey et al., 1992]. A combined speaker indexing system which incorporates the speaker count system is also proposed.

5.2. Unsupervised Speaker Indexing System

A novel approach to speaker indexing, referred to as the Restrained-Relative Minimum Distance (RRMD) Approach is proposed. This technique involves selecting reference models from the conversation and then matching other models from the conversation to the references based on a constrained minimum distance approach. The RRMD method is described below.

5.2.1. The Restrained-Relative Minimum Distance Approach to Speaker Indexing

The proposed RRMD speaker indexing technique is described in the following steps:
Step 1: All pair-wise distances between homogeneous speaker models (formed as described in Section 4.2) are computed.

Step 2: The two models with the maximum difference between them are selected to be the two reference models, as this ensures that they are from two different speakers.

Step 3: Each of the other models are matched to the reference models based on the following conditions

(i) The Restraining Condition:

This condition is based on likelihood ratio testing of each distance in order to determine if it follows an intra-speaker or an inter-speaker distance distribution. It was observed in Chapter 2 that the distances computed, using the distance measures considered in this research, can be approximated by either the Gaussian distribution (Mahalanobis, KL and Bhattacharyya) or the Gamma distribution (T-Square and Levene’s). Let \( \alpha_1 \) and \( \beta_1 \) be the intra-speaker parameters and \( \alpha_2 \) and \( \beta_2 \) be the inter-speaker distribution parameters – which would represent the mean (\( \alpha \)) and the variance (\( \beta \)) for the Gaussian case, and the parameters \( a (\alpha) \) and \( b (\beta) \) given in equation (2.33). Representative values of \( \alpha_1 \) and \( \beta_1 \) and \( \alpha_2 \) and \( \beta_2 \) can be computed for all the distance measures using a significant amount of data from a standard data base (such as the HTIMIT). Therefore, given a distance value, \( x \), computed between two models, the two models can be said to be from the same speaker if the intra-speaker likelihood or probability, \( f(x|\alpha_1,\beta_1) \) is greater than the inter-speaker likelihood, \( f(x|\alpha_2,\beta_2) \), and from different speakers otherwise. A Distance Likelihood Ratio (DLR) can thus be defined as:

\[
DLR = \frac{f(x | \alpha_1, \beta_1)}{f(x | \alpha_2, \beta_2)}
\]  

(5.1)

If the single-speaker and different-speaker cases are assumed to have equal probability, then a DLR value above 1 will indicate that both models are from the same speaker and if the DLR is below 1, then both models are from different speakers. If the DLR between the test model and the reference models to which it is closest is greater than 1, then both models can be
considered to be from the same speaker. If this condition fails, matching is restrained and Condition (ii) below may be checked.

(ii) **The Relative Distance Condition**: This involves computing the difference between the distances of the observed model from both reference models. In other words, let $d_{\text{min}}$ and $d_{\text{max}}$ be the distance from the observed model to its closer and farther reference models, respectively. The relatively distance parameter can be defined as:

\[
D_{\text{rel}} = d_{\text{min}} - d_{\text{max}} \tag{5.2}
\]

If the value of $D_{\text{rel}}$ is greater than a threshold, suitably a value around the difference between the means of the intra- and inter-speaker T-Square distribution, then the restraint may be lifted, and the observed model may be considered to be from the same speaker as its closest reference model. The Relative Distance condition is illustrated in **Figure 5.1**, which shows the intra- and inter-speaker T-Square distributions for $N$ (number of voiced phonemes) = 5. The vertical dashed-lines on each distribution represent estimated means.

![Distribution of T-Square Statistics for N = 5](image_url)

**Figure 5.1**: Illustration of Relative Distance Conditions for Speaker Model Matching.
**Step 4:** Models that fail to meet both requirements may not be given any index, but could instead be considered unusable or undecided, while each reference model, as well as all models matching that same reference, is given an index of either 1 or 2 to represent the first or second speaker in the conversation.

### 5.2.2. Next Steps

In this research, analysis will be performed on the SWITCHBOARD telephone conversations database to determine an appropriate value for $N$, the number of voiced phonemes used for each speaker homogeneous model, and to illustrate the necessity for limiting the data size used in forming models from telephone conversations, especially when speaker change points are unknown. All the distance measures, as well as features, will be considered. In this case, the RRMD indexing technique will not be applied. Instead, a simple Minimum Distance matching method will be applied. Groundtruth data will be obtained from transcriptions provided by the Mississippi Switchboard Transcription Project [Hamaker et al., 1998].

Using the optimal value of $N$ determined from the process described above, speaker indexing will then be performed using the RRMD technique, with the Drel threshold, which controls the amount of undecided speech samples, varying between 0 and 200 based on the distribution shown in Figure 5.1. Two types of error were determined, one being the indexing error and the other error being based on the proportion of undecided speech samples that were not labeled as co-channel speech from the ground truth transcription, as compared to the total number of ground truth speaker samples.

The proposed RRMD speaker indexing system was designed with the assumption that the number of speakers participating in the conversation is known *a priori*, however, it is possible to encounter situation in which more than two speakers are present in the conversation, and the number of speakers is unknown. In order to apply the RRMD approach, a speaker count system must be implemented first, and then the number of reference models to be matched can be increased based on the number of speakers estimated by the speaker count systems. Development of speaker count systems for telephone conversations could be
very challenging especially when no information about the conversation is known. In the next section, a novel speaker count system, which is based on the same idea as the RRMD, approach is proposed.

5.3. Speaker Count System

A three-speaker detection system, based on a Residual Ratio Algorithm (RRA), was recently introduced [Ofoegbu et al., 2006c]. The algorithm involved eliminating two speakers from a conversation and observing the relative amount of speech remaining. In this section, a generalized form of the RRA, referred to as the Generalized RRA (GRRA), is proposed where a speaker count of up to K speakers can be determined.

5.3.1. The Generalized Residual Ratio Approach

A detailed description of this technique is given below:

i. Speech models are formed from a given conversation as explained in Section 4.2, with N equal to 5.

ii. All pair-wise distances for all models in the conversation are computed.

iii. A reference model is chosen at random, and DLR tests are performed (as described in the previous section) between this model and all others. Every model with a DLR > 1 is considered to belong to the reference speaker, and eliminated from the conversation along with the reference model itself, and the Residual Ratio – the ratio of the size of residual speech to the original size of the conversation - is determined. This completes the first elimination round.

iv. Step iii is repeated for the second round; however, the following procedure is taken in order to ensure that the new reference model is not one of those that belong to the first reference speaker but were erroneously mismatched in the first round: the ratio of size of the speech that was matched to the second reference to the total amount of speech is observed, and the process is repeated until this ratio is greater than a chosen threshold determined a priori. Once this condition is satisfied, the Residual Ratio for the second round is determined.

v. Step iv is repeated until the $(K-1)^{th}$ round.
Ideally, all reference models should belong to different speakers, and all models from the \(k^{th}\) \((k = 1, 2, \ldots, K-1)\) reference speaker should be eliminated in the \(k^{th}\) round, and if there are \(k\) speakers, the Residual Ratio after the \(k^{th}\) round should be zero. In practice, however, some models may be mismatched in the elimination rounds, with some models belonging to the references being missed, and some models being wrongfully eliminated as illustrated in Figure 5.4. This figure shows the elimination stages of an artificial conversation simulated by concatenating speech signals from two different speakers from the HTIMIT database. In other words, the first half of the conversation consisted of speech from one speaker while the second half consisted of speech from another speaker.

![How the Residual Ratio Algorithm Works for Two-Speaker Conversations](image)

**Figure 5.4**: Illustration of the Residual Ratio Algorithm for Two-speakers. Original Speech (top panel), remaining speech after first elimination round (middle panel), remaining speech after second elimination round (bottom panel).

Note that all the speech segments from Speaker 1 were eliminated in the first round, and almost all of Speaker 2’s speech segments were eliminated in the second round. One segment from Speaker 2 was not removed, meaning that it was not matched with the reference mode. In spite of such errors, it is expected that the number of segments unmatched (residual segments), as compared to the total number of segments
in the conversation, will be considerably higher for three-speaker conversations, since most of the segments from the first two speakers encountered would be eliminated. In other words, the residual for a three-speaker conversation is expected to be higher than the residual for a two-speaker conversation. This is illustrated in Figure 5.5, which shows the same process as in Figure 5.4, but for three speakers.

![How the Residual Ratio Algorithm Works for Three-Speaker Conversations](image)

**Figure 5.5**: Illustration of the Residual Ratio Algorithm for Three-Speaker. Original Speech (top panel), remaining speech after first elimination round (middle panel), remaining speech after second elimination round (bottom panel).

In the above figure, the algorithm is shown to have successfully eliminated all the first speaker’s segments in the first round but erroneously removes some segments from Speakers 2 and 3. In the second round, the reference is from Speaker 3 and all Speaker 3’s segments were correctly matched and removed. However, some segments from Speaker 2 were also incorrectly removed. Notwithstanding these errors, the ratio of the number of residual segments to the total number of segments was still greater for the three-speaker conversation than for the two-speaker conversation in the experiments shown above.

**Figures 5.6 and 5.7** provide a better picture of the concept of Residual Ratios. A whole circle (pie) represents the entire conversation and the pieces shown in each panel represents the fraction of the conversation that survived the corresponding elimination process. The figures were obtained from the same files used for Figures 5.4 and 5.5.
Figure 5.6: Illustration of the Residual Ratio Algorithm for a Two-Speaker Conversation. Representation of Original Speech (top-left figure), fraction of speech remaining speech after first elimination round (top-right figure), fraction of remaining speech after second elimination process (bottom figure).

Figure 5.7: Illustration of the Residual Ratio Algorithm for a Three-Speaker Conversation. Representation of Original Speech (top-left figure), fraction of speech remaining speech after first elimination round (top-right figure), fraction of remaining speech after second elimination process (bottom figure).
5.3.2. Next Steps

In determining the speaker count based on the Residual Ratios computed, two approaches are proposed. The first is a tree-type classification procedure where the Residual Ratio is observed for each round, and if it is below a certain threshold (which can be obtained by observing residual ratios for a significant number of multi-speaker telephone conversations) for the $k^{th}$ round, the speaker count is considered equal to $k$, and the Residual Ratios for other rounds are not considered. This method is referred to as the Stopped Residual Ratio (SRR) approach. An alternate approach involves determining the speaker count based on the sum of the Residual Ratios for all K-1 rounds. This method is referred to as the Added Residual Ratio (ARR) approach. The higher the ARR, the higher the speaker count is expected to be. These two methods will be investigated in this research using all the distances and features presented in this proposal, as well as a linear combination of the distances, which is proposed in the next chapter.
CHAPTER 6

PROPOSED WORK: CORRELATION ANALYSIS AND FUSION OF DISTANCE MEASURES

6.1. Introduction

In the formulation of the distances (presented in Chapter 2), it can be observed that each distance possesses a unique property that separates it from all other distances. This point is also supported in [Iyer et al., 2001d], where all the distances are observed to yield different performances depending on the data length and feature considered. Another observation made from [Iyer et al., 2001d] is the fact that no one distance is perfect in all conditions for all purposes, even though the Mahanalobis distance appears to outperform all the other distance measures on average. Nonetheless, some of the measures do have very similar characteristics, and could be substituted for one another without major repercussions. In this chapter, the relationships between distances are examined in order to determine and exploit their complimentary information. Some fusion techniques are also introduced. Finally, the derivation of an optimal function by which the distances can be combined to yield the greatest amount of inter-speaker variability and the least amount of intra-speaker variability is proposed. Some preliminary results are also presented.
6.2. Correlation Analysis

A common method for examining the relationship between two features is to generate a scatter plot with one feature plotted against the other. The more linear the data points appear, the greater the linear relationship between both features. On the other hand, the more ball-shaped (i.e. random) the data points are, the more uncorrelated both features are with each other. When scatter plots of each distance are plotted in a matrix form arrangement with each individual plot being small enough to allow for simultaneous viewing, the resulting figure is known as a draftsman display [Manly, 1994].

Figures 6.1 shows the draftsman display of the Mahalanobis distances, the T-Square statistics, the KL distances, the Bhattacharyya distances and Levene’s test values obtained using 14th order LPCCs. The distances (and all intra-speaker and inter-speaker distances mentioned henceforth in this proposal) were computed between speaker homogeneous models formed as described in Section 4.3, with $N = 5$. 1000 intra-speaker and 1000 inter-speaker distances were computed using speech from the HTIMIT database. The x- and y- axes labels represent the distance measure on each plot. The intra-speaker distances are plotted in black circles while the inter-speaker distances are plotted in grey crosses. Note that the draftsman display is very similar to the covariance matrix with the diagonal being 1.
Figure 6.1: Draftsman display of distances obtained using 14th order LPCCs as features. 1000 intra-speaker and 1000 inter-speaker distances were computed using speech files from the HTMIT database.

From Figure 6.1, it is observed that a nonlinear relationship exists between the T-Square and Mahalanobis distances. The KL and Bhattacharyya distances are also somewhat linearly correlated, and this could be understood from the fact that they both belong to the same class of likelihood based distances, as explained in Section 2.2, and were both derived from the same root, i.e. equation (2.11). With the exception of these cases, all other distance relationships depict a considerable amount of uncorrelation, which can be exploited by combining the information from these distances. Note that, in spite of Levene’s test being computed from the T-square statistics, both distance measures appear to be significantly uncorrelated to each other.

The correlation between distances could also be analyzed by computing the mutual information between them [Ellis and Blimes, 2000], [Torkkola, 2000]. Mutual information between two random variables can be defined as a measure of the amount of information each of the random variables is able to reveal about the other. The greater the mutual information between two distance measures, the higher their correlation. The mutual information of between two random variables, C and Y, is given by:
The mutual information between two random variables $C$ and $Y$ is given by the formula:

$$I(C,Y) = \sum_c \int_y p(c,y) \log \frac{p(c,y)}{p(c)p(y)} \, dy$$

(6.1)

Where $p(c,y)$ is the joint probability mass function of $C$ and $Y$, and $p(c)$ and $p(y)$ are the marginal probability mass functions. Now, if both distributions are completely uncorrelated, their mutual information, $I[C;Y]$, will equal zero.

Fusion of the distance measures such that the mutual information between them is increases is proposed in this research. In the next section, levels of fusion are discussed, while some techniques for fusing the distances are introduced and proposed in the following section.

### 6.3. Fusion

A careful selection of the processing stage in which fusion is performed is essential as it determines the effectiveness of the fusion process. The three basic levels of fusion, data-level, feature-level and decision-level [Hall, 1992] are discussed below.

**Data-level fusion:** this involves fusion of data before any processing has been performed. Information is not lost in this case, as the data is raw. Features are extracted from the fused data for further classification processing. Fusion of data from various images observed on a pixel basis is one example of data-level fusion [Linn and Hall, 1991]. This method of fusion, although very efficient in multi-sensor data fusion systems, is not appropriate for speech processing as it is computationally expensive and highly redundant.

**Feature level fusion:** this is a very common fusion method used mostly in multi-data systems. It involves extracting various features (used for the same classification) from the observed data and then fusing them. When combined, these features are expected to yield better classification accuracy than each feature would individually. It is very important that the features selected to be fused are uncorrelated, otherwise the fusion
process will be redundant and might yield more errors. Another key requirement in feature-level fusion is that a sufficient amount of training data is available for training the fusion procedure. Independent component analysis and nonlinear estimation have previously been applied for fusion of speech processing features [Smolenski et al., 2002b].

**Decision level fusion**: this involves fusion at the highest processing stage. In this method, the outputs of the various classification techniques are fused, in other word, the class identity is separately pre-determined by each classifier, and the final identity is obtained based on these identities. Various techniques exist for decision level fusion, including voting [Kitler, 1998], Bayesian classification [Smolenski and Yantorno, 2003], Consensus method [Benediktsson and Swain, 1992], [Altincay and Demirekler, 2000] and Dempster-Shafer method [Shafer, 1990], [Shafer and Pearls, 1990], [Shafer, 1976], [Dempster, 1968].

A fusion level which falls between feature-level and decision-level fusion, referred to as “measure level fusion”, is proposed. In this case, the distances computed between the features will be fused and used in making decisions on whether or not the models being compared belong to the same speaker. A proposed decision level fusion technique will also be discussed. These methods are discussed in the next section of the proposal.

**6.4. Optimized Fusion of Distances**

In order to exploit the uncorrelation among the distance measures, a new distance measure, which is a combination of distances measures which yields the least separation between intra-speaker, and the most separation between inter-speaker speech is proposed. The goal is to minimize the variance of each class and maximize the difference between their means. This can be achieved by maximizing the t-test, for which the intra-class variances affect the denominator and the separation of the means affect the numerator (Equation (3.15)).
An optimal linear combination of the distances can be computed by solving the equation:

$$T_{\text{max}} = a^T X$$

(6.2)

Where $X$ is a vector consisting of the distance measure values and $a$ is a vector of the weights assigned to each measure value and $T_{\text{max}}$ is the new t-test maximizing distance. Since the lengths of the two classes to be compared (intra-speaker and inter-speaker distances) are always equal (and equal to the number of experiments run), the t-test maximizing cost function can be expressed as:

$$T(a) = \frac{\mu_1 - \mu_2}{\sigma_1^2 - \sigma_2^2}$$

(6.3)

Where $T(a)$ is a representation of the maximum T-value between the classes, $\mu_1$ and $\mu_2$ are the mean values of the two classes, and $\sigma_1^2$ and $\sigma_2^2$ are their variances. The weights, $a$, which would yield the desired t-distances is then computed as:

$$a = \frac{k}{\lambda_1} P^{-1} (\mu_1 - \mu_2)$$

(6.4)

$$\text{Where } k = \frac{\lambda_1}{\| P^{-1} (\mu_1 - \mu_2) \|^2}$$

(6.5)

And $P$ is the sum of the covariance matrices of the two classes. Note that $\lambda_1$ is the maximum eigenvalue obtained by solving the generalized eigenvalue problem:

$$P^{-1}Qa = \lambda_1 a$$

(6.6)

where $Q$ is the square of the distance between the mean vectors of the two classes [Stark and Woods, 2002]. The proposed distance is referred to as the $T_{\text{max}}$ distance.

Figure 6.2 illustrates the effect of combining all distances using the above technique. The $T_{\text{max}}$ intra- and inter-speaker distances were computed between the two classes using Equation (6.2), and their distributions are shown in the top row of the figure. The distributions of the single distance measures are also shown (in
rows below) for comparison. For quantification purposes, the t-test was also conducted between the classes and the values are given on the title of each distribution. The LPCCs were used as features.

**Figure 6.2** shows that the $T_{max}$ distance yields the highest separation between the two classes for both sets of features.

**Figure 6.2:** Illustration of the effect of the weighted combination of distances

In addition to examining the effect of combining the distances, it is also important to study the contribution made by each of the distances in maximizing the class separation. This is performed by increasing the
number of distance measures used in the computation of the $T_{\text{max}}$ distance, starting from the measure with the highest T-value and proceeding in decreasing order of T-value until all the distance measures have been employed, and observing the T-values between the intra-speaker and inter-speaker classes. This is illustrated in Figures 6.3. In order for the study to be effective, it was also necessary to ensure that the next distance measure being added was that which would yield the highest T-value when combined with those already included in the combination. This was achieved by a dynamic programming approach which involved testing all the distances each time before selecting the next one to be included.

![Figure 6.3](image)

**Figure 6.3:** Illustration of the effect of increasing the number of distances – LPCC features used

**Figure 6.3** shows that with the LPCCs, Levene’s test is the most uncorrelated with the Mahalanobis distance, and provides the most significant increase in separation when combined. The combination of other distances has very little effect on the improvement of the distance measure. Note that using any other
distance but Levene’s test with the LPCCs would result in a much less increase in the T-value of the combination.

6.5. Next Steps

It has just been shown that the distance measurements can be enhanced simply by performing a linear combination of distances. There is much room for increase in the amount of separability between the two classes, however, and this will be investigated in this research by formulating other maximization functions, such as attempting to maximize the univariate KL distance or the (simpler) Euclidean distance between the distance measures.

A Decision level fusion technique is also proposed in this research, whereby weights will be assigned to each distance measure based on the T-value between their intra-speaker and inter-speaker values computed using a significantly large dataset. The inter-class T-values obtained for each distance could be normalized such that they sum up to 1, and then, given the output of the likelihood ratio testing (that is, the decision made by each distance on whether or not the models compared are from the same speaker), a final class decision could be made. For instance, for the LPCCs, a weight (denoted by \( \omega_i \) \( i = 1, 2, 3, 4, 5 \) representing the Mahalanobis, T-Square, KL, Bhattacharyya and Levene’s distances respectively) could be assigned to the each distance using the equation:

\[
\omega_i = \frac{T_i}{\sum_i T_i} \quad (6.7)
\]

where \( T_i \) represents the T-value corresponding to each distance.

The application systems proposed in Chapter Five will be enhanced using the fusion techniques introduced in this chapter.
CHAPTER 7
SUMMARY OF PROPOSED WORK

Further work on this research can be summarized as follows:

1. In this proposal, the LPCCs were analyzed in the various speaker discrimination methods introduced. In the same way, analysis of the MFCCs and the delta and delta-delta coefficients, as well as various combinations of all the features, will be performed as outlined in Section 3.3. The speaker discrimination performances of the features and their combinations will also be evaluated using the T-test, as discussed in Section 3.3, as well as other tests of dissimilarity between intra-speaker and inter-speaker distance measurements.

2. Feature combinations will be enhanced using Principal Component Analysis.

3. The Proposed Restrained-Relative Minimum Distance Approach to speaker indexing will be completely developed and analyzed using the various distances, features and feature combinations discussed in this proposal as outlined in Subsection 5.2.2.

4. The stopped Residual Ratio and Added Residual Ratio techniques will be employed in the Generalized Residual Ratio Speaker Count System as discussed in Subsection 5.3.2.

5. Linear and nonlinear methods for the formation of an optimized combination of the distance measurements so as to yield maximum inter-class and minimum intra-class variability will be investigated – in extension to the work proposed in Section 6.4.

6. The decision-based fusion technique proposed in Section 6.5 will be fully developed and analyzed.
7. A finalized telephone conversations-based speaker recognition system will be developed based on all the techniques introduced in this research, and evaluated using telephone conversations databases such as SWITCHBOARD.


