MODEL FORMATION AND CLASSIFICATION TECHNIQUES FOR
CONVERSATIONS-BASED SPEAKER DISCRIMINATION

A Dissertation

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
May 2007

Advisor: Dr. Robert E. Yantorno
This dissertation is dedicated to God the Father, the Son and the Holy Ghost, without whom this would have been impossible
Wisdom is the principal thing; therefore get wisdom: and with all thy getting get understanding
- Proverbs 4:7
ACKNOWLEDGEMENT

To the greatest teacher in the world, and the one who has made the most impact in my life, Dr. Robert E. Yantorno.

To my best friend and the love of my life, Dr. Jude C. Abanulo

To Dr. Brian Butz, Dr. John Helferty, Dr. Saroj Biswas and Dr. Henry Seendaula

To my dissertation committee members, Dr. Iyad Obeid and Dr. Dennis Silage, and to Dr. Rena Krakow

To my friend, Dr. Ananth Iyer and to Poornima

To Abdoul Fall, Joe Fitchgrund, Angela Linse, Marge Pippet and Ralph Oyini; and to the members of the Speech Processing Lab and the faculty of the electrical engineering department

To engineering administrators, Tamika Butler, Carol Dahlberg, Yvette Gibson and Cheryl Sharp, and to Louise, daytime janitress for the engineering building

To the Temple students who volunteered as participants in the Temple Conversations Database

To Temple

To the Air Force Research Labs at Rome – Financial supporters of most of the research

To my parents in-law (to be), Scholastica & Cyril Abanulo; and my siblings in-law (to be), Linda & Ike Nnene, Antoinette Abanulo, Annette and Azuma Kalu, Chidi and Ugochukwu Abanulo

To my parents, Ugo & Joseph Ofoegbu; my siblings, Amaka & Humphrey Onyendi, Nene, Obinna and Chibuzor Ofoegbu; and my grandmother, Cordelia Osuji

To God

Thank you.
MODEL FORMATION AND CLASSIFICATION TECHNIQUES FOR CONVERSATIONS-BASED SPEAKER DISCRIMINATION

Uchechukwu O. Ofoegbu
Doctor of Philosophy
Temple University, 2007
Doctoral Advisory Committee Chair: Robert E. Yantorno, Ph.D.

ABSTRACT

Speaker recognition is the process of recognizing speakers from their voices. Some of the most common speaker recognition applications are speaker identification which involves determining which speaker, amongst a group of registered speakers, corresponds to a given test utterance, and speaker verification systems, which involves either accepting or rejecting the identity claim of a speaker based on comparisons of a test utterance with previously enrolled utterances of the presumed speaker. A general requirement for these systems is that all participating speakers are known a priori and the system is trained with information from the voices of these speakers. Also, individual utterances are obtained from the participating speakers; therefore, a large amount of data is available for training the system, as well as for evaluating its performance.

A more recent application of speaker recognition involves differentiating between speakers in conversational data. This process is unsupervised (performed without a priori data about participants), with only short lengths of speaker utterances available. Current conversations-based speaker recognition systems operate by partitioning the speech data into segments of equal lengths and using features from each segment to represent the speakers. The problems with this method are as follows: 1) speaker change points are unknown. 2) Not all classes of speech are useful in characterizing speakers, and, since only short homogeneous speaker utterances are available, it is possible that an information...
rich segment be compared with an information deficit segment resulting in misclassifications. 3) Some portions of conversational data consist of overlapped speech from two different speakers, and cannot be effectively used to represent a single speaker. As a result of these deficiencies, state-of-the art conversations-based recognition systems report errors that range from 11%-40%.

This research addresses the above mentioned problems by 1) selectively creating data models using a combination of features from enhanced portions of short speaker utterances, 2) determining and implementing the best set of distance measures which will yield the minimum same-speaker and maximum different-speaker separation for conversations, and 3) developing a conversations-based speaker differentiation technique which takes into account the problems of short utterance lengths, co-channel speech, and lack of a priori information. A comprehensive conversations-based system, which can effectively differentiate between up to four different speakers in a conversation, was developed and tested on two different standard speech databases. An accuracy rate of over 90% has been obtained with the system.
# TABLE OF CONTENTS

ABSTRACT ................................................................................................................................. v

Chapter 1: INTRODUCTION .............................................................................................. 1

1.1 An Overview of Speaker Discrimination ................................................................. 1

1.1.1. Types of Speaker Discrimination ...................................................................... 1

1.1.2. Applications of Speaker Discrimination ......................................................... 2

1.2. Distances for Speaker Discrimination ...................................................................... 6

1.3. Features for Speaker Discrimination ...................................................................... 8

1.4. Application of Fusion in Speaker Discrimination ............................................... 10

1.5. Research Goal: The Challenge of Telephone Conversations ......................... 11

1.6. Scope of Research ................................................................................................. 13

1.7. Dissertation Outline ............................................................................................... 15

Chapter 2: BACKGROUND .............................................................................................. 16

2.1. Introduction ............................................................................................................ 16

2.2. Features .................................................................................................................. 16

2.2.1. Cepstral Analysis .............................................................................................. 17

2.2.2. Linear Predictive Cepstral Coefficients ........................................................... 19

2.2.3. Mel-Scale Frequency Cepstral Coefficients .................................................... 20

2.2.4. Delta Cepstral Coefficients .............................................................................. 23

2.3. Distance Measures ............................................................................................... 24

2.3.1. Mahanalobis Distance ...................................................................................... 25

2.3.2. Hotelling’s T-Square Statistics ........................................................................ 26

2.3.3. Kullback Leibler Distance ............................................................................... 27

2.3.4. Bhattacharyya Distance ................................................................................... 28

2.3.5. Levene’s Test .................................................................................................. 29

2.4. Speech Classification and Extraction of Voiced Speech .................................... 29

2.4.1. Speech Classification ....................................................................................... 29
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.4.2. Voiced-Unvoiced Classifiers</td>
<td>30</td>
</tr>
<tr>
<td>2.5. Usable Speech Detection</td>
<td>32</td>
</tr>
<tr>
<td>2.5.1. Target to Interferer Ratio (TIR) Based Usable Speech</td>
<td>33</td>
</tr>
<tr>
<td>2.5.2. Application Based Usable Speech</td>
<td>33</td>
</tr>
<tr>
<td>2.6. Fusion</td>
<td>34</td>
</tr>
<tr>
<td>2.7. Summary of the Chapter</td>
<td>36</td>
</tr>
<tr>
<td>Chapter 3: ANALYSIS OF FEATURES FOR SPEAKER DISCRIMINATION</td>
<td>37</td>
</tr>
<tr>
<td>3.1. Introduction</td>
<td>37</td>
</tr>
<tr>
<td>3.2. Analysis of Distance Measures</td>
<td>37</td>
</tr>
<tr>
<td>3.2.1. Procedural Set-up</td>
<td>37</td>
</tr>
<tr>
<td>3.2.2. Results and Observations</td>
<td>40</td>
</tr>
<tr>
<td>3.3. Analysis of Features</td>
<td>45</td>
</tr>
<tr>
<td>3.3.1. Procedural Set-Up</td>
<td>46</td>
</tr>
<tr>
<td>3.3.2. Results and Observations</td>
<td>46</td>
</tr>
<tr>
<td>3.3.3. Feature Combination</td>
<td>49</td>
</tr>
<tr>
<td>3.4. Summary of the Chapter</td>
<td>57</td>
</tr>
<tr>
<td>Chapter 4: NOVEL SPEAKER DISCRIMINATION METHODS AND TECHNIQUES</td>
<td>58</td>
</tr>
<tr>
<td>4.1. Introduction</td>
<td>58</td>
</tr>
<tr>
<td>4.2. Formation and Analysis of Data Models</td>
<td>58</td>
</tr>
<tr>
<td>4.3. Unsupervised Speaker Indexing System</td>
<td>69</td>
</tr>
<tr>
<td>4.4. Speaker Count System</td>
<td>74</td>
</tr>
<tr>
<td>4.5. Speaker Count-Indexing System</td>
<td>79</td>
</tr>
<tr>
<td>4.6. Fusion of Distance Measures</td>
<td>80</td>
</tr>
<tr>
<td>4.6.1. Correlation Analysis</td>
<td>81</td>
</tr>
<tr>
<td>4.6.2. Optimized T Distance</td>
<td>83</td>
</tr>
<tr>
<td>4.6.3. Decision-Level Fusion</td>
<td>90</td>
</tr>
<tr>
<td>4.7. Summary of the Chapter</td>
<td>91</td>
</tr>
<tr>
<td>Chapter 5: EVALUATION</td>
<td>92</td>
</tr>
<tr>
<td>5.1. Introduction</td>
<td>92</td>
</tr>
<tr>
<td>5.2. Databases</td>
<td>92</td>
</tr>
<tr>
<td>5.2.1. Standard Databases</td>
<td>93</td>
</tr>
<tr>
<td>5.2.2. Temple Conversations Database</td>
<td>94</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
</tr>
<tr>
<td>--------</td>
<td>---------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>2-1.</td>
<td>Illustration of cepstral analysis. The circle represents the desired vocal tract</td>
</tr>
<tr>
<td></td>
<td>information obtained by liftering.</td>
</tr>
<tr>
<td>2-2.</td>
<td>Mel-scale Frequency Function</td>
</tr>
<tr>
<td>3-1.</td>
<td>Block diagram showing the intra-speaker distance computation procedure.</td>
</tr>
<tr>
<td>3-2.</td>
<td>Block diagram showing the inter-speaker, same utterance distance computation procedure.</td>
</tr>
<tr>
<td>3-3.</td>
<td>Block diagram showing the inter-speaker, different utterance distance computation procedure.</td>
</tr>
<tr>
<td>3-4.</td>
<td>Estimated Probability Density Functions for Intra-Speaker (SSDU), inter-speaker same</td>
</tr>
<tr>
<td></td>
<td>utterance (DSSU) and inter-speaker different utterances (DSDU) Mahalanobis distances.</td>
</tr>
<tr>
<td>3-5.</td>
<td>Estimated Probability Density Functions for Intra-Speaker (SSDU), inter-speaker same</td>
</tr>
<tr>
<td></td>
<td>utterance (DSSU) and inter-speaker different utterances (DSDU) T-Square statistics.</td>
</tr>
<tr>
<td>3-6.</td>
<td>Estimated Probability Density Functions for Intra-Speaker (SSDU), inter-speaker same</td>
</tr>
<tr>
<td></td>
<td>utterance (DSSU) and inter-speaker different utterances (DSDU) Kullback Leibler distances.</td>
</tr>
<tr>
<td>3-7.</td>
<td>Estimated Probability Density Functions for intra-speaker (SSDU), inter-speaker same</td>
</tr>
<tr>
<td></td>
<td>utterance (DSSU) and inter-speaker different utterances (DSDU) Bhattacharyya distances.</td>
</tr>
<tr>
<td>3-8.</td>
<td>Estimated Probability Density Functions for intra-speaker (SSDU), inter-speaker same</td>
</tr>
<tr>
<td></td>
<td>utterance (DSSU) and inter-speaker different utterances (DSDU) Levene’s Test values.</td>
</tr>
<tr>
<td>3-9.</td>
<td>Histograms and Probability Density Functions for intra-speaker and inter-speaker</td>
</tr>
<tr>
<td></td>
<td>Mahalanobis distance values using LPCCs and MFCCs as features.</td>
</tr>
</tbody>
</table>
3-10. Histograms and Probability Density Functions for intra-speaker and inter-speaker T-Square statistic values using LPCCs and MFCCs as features. ..................................................44

3-11. Histograms and Probability Density Functions for intra-speaker and inter-speaker Kullback Leibler distance values using LPCCs and MFCCs as features. ..........................44

3-12. Histograms and Probability Density Functions for intra-speaker and inter-speaker Bhattacharyya distance values using LPCCs and MFCCs as features. .................................45

3-13. Histograms and Probability Density Functions for intra-speaker and inter-speaker Levene’s distance values using LPCCs and MFCCs as features. .........................................45

3-14. T-values for the various feature combinations: feature set index 1) 22nd order LPCCs, 2) 22nd order MFCCs, 3) 11th order LPCCs and 11th order MFCCs, 4) 22nd order LPCCs and their delta, 5) 22nd order LPCCs and their deltas, 6) 11th order LPCCs and 11th order MFCCs and their corresponding deltas, 7) 22nd order LPCCs and their delta and delta-deltas, 8) 22nd order LPCCs and their deltas and delta-deltas, 9) 11th order LPCCs and 11th order MFCCs and their corresponding deltas and delta-deltas. ........................................................................................................ ..........48

3-15. T-values obtained by increasing the order of the LPCC coefficients from 3 to 30 ...............50

3-16. T-values obtained by increasing the order of the MFCC coefficients from 3 to 30. ............51

4-1. Formation of speaker homogeneous segments from voiced speech segments (v); u = unvoiced speech segments, s = silence portions. ..................................................................................54

4-2. Mahalanobis distances obtained using homogenous speaker models formed from 20 voiced phonemes. ..................................................................................................................55

4-3. T-Square distances obtained using homogenous speaker models formed from 20 voiced phonemes. ..................................................................................................................55

4-4. Kullback-Leibler distances obtained using homogenous speaker models formed from 20 voiced phonemes. .................................................................56

4-5. Bhattacharyya distances obtained using homogenous speaker models formed from 20 voiced phonemes. .................................................................56

4-6. Levene’s distances obtained using homogenous speaker models formed from 20 voiced phonemes. ..................................................................................57

4-7. Comparison of Mahalanobis Distances for different data sizes. X-axis represents the number of voiced phonemes used to form each model. .................................58

4-8. Mahalanobis distances obtained using homogenous speaker models formed from 5 voiced phonemes. ..................................................................................59
4.9. T-Square distances obtained using homogenous speaker models formed from 5 voiced phonemes .......................................................... 59

4.10. Kullback-Leibler distances obtained using homogenous speaker models formed from 5 voiced phonemes ........................................... 60

4.11. Bhattacharyya distances obtained using homogenous speaker models formed from 5 voiced phonemes .............................................. 60

4.12. Levene’s distances obtained using homogenous speaker models formed from 5 voiced phonemes ......................................................... 61

4.13. Voiced-phoneme based T-values obtained by increasing the order of the LPCC coefficients from 3 to 30 .................................................... 62

4.14. Voiced-phoneme based T-values obtained by increasing the order of the MFCC coefficients from 3 to 30 .................................................... 62

4.15. Illustration of Relative Distance Conditions for Speaker Model Matching ............................................................. 66

4.16. Average percent indexing accuracy with respect to number of voiced phonemes used to form models .......................................................... 68

4.17. 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) ........................................ 70

4.18. 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) ........................................ 71

4.19. Illustration of the Residual Ratio Algorithm for Two-Speaker. Representation of Original Speech, fraction of speech remaining speech after first elimination round, fraction of remaining speech after second elimination process (bottom figure) ........................................ 72

4.20. Illustration of the Residual Ratio Algorithm for Two-Speaker. Representation of Original Speech, fraction of speech remaining speech after first elimination round, fraction of remaining speech after second elimination process (bottom figure) ........................................ 72

4.21. Added Residual Ratio distribution functions for 1000 artificially generated conversations each for 1-4 speaker cases ........................................ 73

4.22. 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 ........................................ 76

4.23. Draftsman display of distances obtained using 14th order MFCCs as features. 1000 intra-speaker and 1000 inter-speaker distances were computed using speech files from the HTMIT database ........................................ 77
4.24. Illustration of the effect of the weighted combination of distances – LPCCs used as features. Top panel shows the maximized T distance. ..........................................................79

4.25. Illustration of the effect of the weighted combination of distances – MFCCs used as features. Top panel shows the maximized T distance. ..........................................................80

4.26. Illustration of the effect of the increasing the number of distances – LPCC features used. ..................................................................................................................81

4.27. Illustration of the effect of the increasing the number of distances – MFCC features used. ..................................................................................................................81

5.1a. Average percent indexing accuracy with respect relative distance threshold. Indexing was performed using the Mahalanobis distance with the RRMD Technique. Features used – 12th order LPCC. ..................................................................................90

5.1b. Average percent indexing accuracy with respect relative distance threshold. Indexing was performed using the Mahalanobis distance with the RRMD Technique. Features used – 22nd order MFCC. ..................................................................................90

5.2a. Average percent indexing accuracy with respect relative distance threshold. Indexing was performed using the T-Square distance with the RRMD Technique. Features used – 14th order LPCC. ..................................................................................91

5.2b. Average percent indexing accuracy with respect relative distance threshold. Indexing was performed using the T-Square distance with the RRMD Technique. Features used – 28th order MFCC. ..................................................................................91

5.3a. Average percent indexing accuracy with respect relative distance threshold. Indexing was performed using the Bhattacharyya distance with the RRMD Technique. Features used – 12th order LPCC. ..................................................................................92

5.3b. Average percent indexing accuracy with respect relative distance threshold. Indexing was performed using the Bhattacharyya distance with the RRMD Technique. Features used – 22nd order LPCC. ..................................................................................92

5.4a. Speaker count accuracy for each distance. Results obtained using the HTIMIT database and LPCCs as features. ..................................................................................94

5.4b. Speaker count accuracy for each distance. Results obtained using the HTIMIT database and MFCCs as features. ..................................................................................95

5.5a. Speaker count accuracy comparison for distance fusion techniques. Results obtained using the HTIMIT database and LPCCs as features. ............................................................................96

5.5b. Speaker count accuracy comparison for distance fusion techniques. Results obtained using the HTIMIT database and MFCCs as features. ............................................................................96

5.6a. Speaker count accuracy for each distance. Results obtained using the Temple Conversations Database and LPCCs as features. ............................................................................97
5.6b. Speaker count accuracy for each distance. Results obtained using the Temple Conversations Database and MFCCs as features. ................................................................. 98

5.7a. Speaker count accuracy comparison for distance combination techniques. Results obtained on the Temple Conversations Database using LPCCs as features......................... 98

5.7b. Speaker count accuracy comparison for distance combination techniques. Results obtained on the Temple Conversations Database using MFCCs as features............................... 99

5.8. Cross evaluation of the speaker count system using the HTIMIT and Temple Conversations (TC) databases. .................................................................................... 100

5.9. Indexing accuracy results obtained for both databases. The black solid bars represent the accuracy results for the HTIMIT while the black and grey checkered bar represents the result for the new database. ................................................................. 101
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1 Outline of Feature Combinations for Speaker Recognition</td>
<td>46</td>
</tr>
<tr>
<td>6-1. Outline of Optimal Number of Coefficients for Voiced-Phoneme Based Feature Computation</td>
<td>63</td>
</tr>
</tbody>
</table>
# LIST OF ACRONYMS

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARR</td>
<td>Added Residual Ratio</td>
</tr>
<tr>
<td>DBC</td>
<td>Decision-Based Combination</td>
</tr>
<tr>
<td>DCT</td>
<td>Discrete Cosine Transform</td>
</tr>
<tr>
<td>DLR</td>
<td>Distance Likelihood Ratio</td>
</tr>
<tr>
<td>DSDU</td>
<td>Different Speaker Different Utterance</td>
</tr>
<tr>
<td>DSSU</td>
<td>Different Speaker Same Utterance</td>
</tr>
<tr>
<td>FIR</td>
<td>Finite Impulse Response</td>
</tr>
<tr>
<td>BIC</td>
<td>Bayesian Information Criterion</td>
</tr>
<tr>
<td>GLR</td>
<td>Gaussian-Likelihood-Ratio</td>
</tr>
<tr>
<td>GMM</td>
<td>Gaussian Mixture Model</td>
</tr>
<tr>
<td>GRRA</td>
<td>Generalized Residual Ratio Algorithm</td>
</tr>
<tr>
<td>IDFT</td>
<td>Inverse Discrete Fourier Transform</td>
</tr>
<tr>
<td>IFT</td>
<td>Inverse Fourier Transform</td>
</tr>
<tr>
<td>KL</td>
<td>Kullback-Leibler</td>
</tr>
<tr>
<td>LAR</td>
<td>Log Area Ratio</td>
</tr>
<tr>
<td>LPC</td>
<td>Linear Predictive Coding</td>
</tr>
<tr>
<td>LPCC</td>
<td>Mel Frequency Cepstral Coefficient</td>
</tr>
<tr>
<td>LSP</td>
<td>Line Spectral Pair</td>
</tr>
<tr>
<td>Acronym</td>
<td>Definition</td>
</tr>
<tr>
<td>---------</td>
<td>------------</td>
</tr>
<tr>
<td>MD</td>
<td>Minimum Distance</td>
</tr>
<tr>
<td>MFCC</td>
<td>Linear Predictive Cepstral Coefficient</td>
</tr>
<tr>
<td>OTD</td>
<td>Optimized T Distance</td>
</tr>
<tr>
<td>PCA</td>
<td>Principal Component Analysis</td>
</tr>
<tr>
<td>PLP</td>
<td>Perceptual Linear Prediction</td>
</tr>
<tr>
<td>RMS</td>
<td>Root Mean Square</td>
</tr>
<tr>
<td>RRA</td>
<td>Residual Ratio Algorithm</td>
</tr>
<tr>
<td>RRMD</td>
<td>Restrained-Relative Minimum Distance</td>
</tr>
<tr>
<td>SID</td>
<td>Speaker Identification</td>
</tr>
<tr>
<td>SNR</td>
<td>Signal-to-Noise Ratio</td>
</tr>
<tr>
<td>SSDU</td>
<td>Same Speaker Different Utterances</td>
</tr>
<tr>
<td>STFT</td>
<td>Short Time Fourier Transform</td>
</tr>
<tr>
<td>SV</td>
<td>Speaker Verification</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>TIR</td>
<td>Target-to-Interferer-Ratio</td>
</tr>
<tr>
<td>UBM</td>
<td>Universal Background Model</td>
</tr>
<tr>
<td>WDBC</td>
<td>Weighted Decision-Based Combination</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

1.1 An Overview of Speaker Discrimination

Speaker discrimination, which is sometimes referred to as 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 stage and the testing stage. 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
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 from 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
Speaker Indexing: This 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 [Ferganil 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 can 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 they 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 Gopalakrishnan, 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 considerably enhance 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. 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 Gray and Markel (1976), 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 were 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 Mahalanobis distance, KL distance, T-Square statistic, Euclidean distance, Hellinger distance, Bhattacharyya distance, 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. Due to the nonstationary nature of speech which causes its characteristics to change after short durations, the measurements are taken on short frames of speech. 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. Cepstral features, on the other hand 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. This is achieved 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 to convert 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 derived from LPC coefficients, and the Mel-Scale Frequency Cepstral Coefficients (MFCCs), which are based on auditory perception of speech signals.
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.

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 somewhat uncorrelated to one another. As a result, in the past decade, some techniques have been developed 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. Each of these approaches yielded a significant increase in the performance of the corresponding speaker recognition system.

Fusion of generative and discriminative classifiers (generative being those that entail learning a model of the joint probability of the inputs and making predictions based on Bayes probability rules, while discriminative classifiers are 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 they combined the UBM-GMM (Universal Background Models – Gaussian Mixture Models) classifier with SVM for speaker detection. Again, speaker recognition systems were shown to perform better when these fusion techniques were implemented.

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 differentiate between speakers participating in conversations. Discriminating speakers participating in conversations is a more complex task than other speaker recognition endeavors for the following reasons:
1. **Speakers change rapidly**: during 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 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 conversations, the contrary is the case [Iyer et al., 2006d].

2. **Lack of a priori information**: Unlike for SID, SV and speaker indexing of broadcast or conference data, speech data from speakers participating in telephone conversations is usually unavailable, so the system cannot be trained with any speaker information. Moreover, for speaker-indexing, the number of speakers recorded is sometimes known to the system. This is not always the case. For instance, when handling telephone conversations such as those made from prisons, the number of participants is never known. This lack of information poses a challenge for speaker recognition systems which have to deal with conversations.

3. **Adverse conditions**: Like with any speech produced in uncontrolled environments, working with conversational 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, 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. 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, is performed in this research. Examination of features and features combinations will also be addressed in this research, as well a technique for enhancing features for desirable performances even in unfavorable conditions.

1.6. Scope of Research

This research is relevant 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. 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 finger prints 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 services 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 examined 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. Dissertation Outline

This dissertation is organized as follows: In Chapter Two, background information about the techniques used in this dissertation is presented. This includes a presentation of features and distance measures for speaker discrimination features and a discussion about speech segmentation, usable speech detection, and fusion. Investigation of the speaker differentiation capabilities of the features, using the different distance measures, is presented in Chapter Three. In Chapter Four, novel techniques for speaker discrimination of conversational data are introduced. In this Chapter, a detailed description of process of forming data models to represent speakers is given. The development of techniques for specific speaker discrimination applications is presented, and correlation analysis and fusion of the distance measures are investigated. Experimental Results are presented in Chapter Five. Conclusion and some possible future directions for the research are given in Chapter Six.
CHAPTER 2
BACKGROUND

2.1. Introduction

In this chapter, background information about the techniques applied in this research is presented. Cepstral features, which are the most commonly used features for speaker discrimination, are discussed in detail. Some distance measures which have been applied successfully in speaker recognition techniques are also presented. In this research, methods for enhancing data are investigated. These methods involve the automatic segmentation of speech and the extraction of voiced phonemes from the segmented data. As a result, currently existing approaches for data segmentation and usable speech detection are also presented in this chapter. Finally, because this research deals with the development and investigation of distance combination techniques, a brief overview of the concept of information fusion is given.

2.2. Features

As discussed in Chapter One, several high and low level features have been successfully 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 they are low level features and therefore 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 section

2.2.1. 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 respectively 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:

$$S(e^{i\theta}) = H(e^{i\theta})E(e^{i\theta})$$

(2.2-1)

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 (2.2-1) above, the following equation is obtained:

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

(2.2-2)

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 2.1 below, which is an adaptation from [Deller et al., 2000].

Figure 2.1: Illustration of cepstral analysis. The circle represents the desired vocal tract information obtained by liftering. (Adapted from [Deller et al., 2000]).

Figure 2.1 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 easily separated.
The derivations of the LPCCs and MFCCs are based on this analysis, even though their computations do not directly follow the computational steps explained above. The computations of the features are explained in the following two sub-sections.

2.2.2. Linear Predictive Cepstral Coefficients

The LPCCs are derived from a combination of cepstral analysis (illustrated in Figure 2.1) 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:

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

(2.2-3)

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:

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

(2.2-4)

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\textsuperscript{th} order filter using information from the (i-1)\textsuperscript{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 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:

\[
c_0 = \ln E^2
\]  \hspace{1cm} (2.2-5)

\[
c_m = a_m + \frac{1}{m} \sum_{k=1}^{m-1} \left[ -(m-k) a_k c_{(m-k)} \right] \quad 1 \leq m \leq p
\]  \hspace{1cm} (2.2-6)

\[
c_m = \sum_{k=1}^{m-1} \left[ \frac{- (m-k)}{m} a_k c_{(m-k)} \right], \quad p \leq m \leq N
\]  \hspace{1cm} (2.2-7)

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

2.2.3. Mel-Scale Frequency Cepstral Coefficients

Using MFCCs is motivated by the idea of building a model based on how the human auditory system analyzes speech. This is standard in most speech processing applications. The Mel frequency scale, depicted in Figure 2.2, 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 STFT of frames (or windows) of the speech signal are computed as:

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

(2.2-8)

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:

\[
\text{mel}(f) = 2595 \log_{10}(1 + f/700)
\]  

(2.2-9)

This function is illustrated in Figure 2.2 for the frequency range of 0-10KHz.
Step 3: The energy of the output of each Mel-Scale filter is then computed using Equation (2.2-10), 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.

$$E(n,i) = \frac{1}{\lambda_i} \sum_{k=A_i}^{B_i} |\eta_i(\omega_k)F(n,\omega_k)|^2$$  \hspace{1cm} (2.2-10)

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} (2.2-11)
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 of the function, 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:

\[
C(n,m) = \frac{1}{R} \sum_{i=0}^{R-1} \ln\{E(n,i)\} \cos(2\pi im / R)
\]

(2.2-12)

where \(R\) is the number of filters.

2.2.4. 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]:
\[ \Delta c_{m,n} = \frac{\sum_{k=-P}^{P} k c_{m,n+k}}{\sum_{k=-P}^{P} |k|} \]  
(2.2-13)

\[ \Delta\Delta c_{m,n} = \frac{\sum_{k=-P}^{P} k^2 c_{m,n+k}}{\sum_{k=-P}^{P} k^2} \]  
(2.2-14)

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.

2.3. Distance Measures

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 these distances where 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 section, the distance measures will be presented in detail. The following notations will be applied. The random variables:

\[
X = [X_1, X_2, \ldots, X_p] \tag{2.3-1}
\]

\[
Y = [Y_1, Y_2, \ldots, Y_p] \tag{2.3-2}
\]
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, \(f_x(X)\) and \(f_y(Y)\) which follow the multivariate Gaussian distribution:

\[
f_x(x) = \frac{1}{(2\pi)^{n_x/2} |\Sigma_x|^{1/2}} \exp \left\{ -\frac{1}{2} (x - \mu_x)^T \Sigma_x^{-1} (x - \mu_x) \right\}. \tag{2.3-3}
\]

In order to be considered a valid distance metric, each distance measure, \(Q(X,Y)\), between two random variables, \(X\) and \(Y\), 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 satisfy the following properties:

\[
Q(X,Y) \geq 0, \tag{2.3-4}
\]

\[
Q(X, Y) = 0 \text{ iff } X = Y, \tag{2.3-5}
\]

\[
Q(X, Y) = Q(Y, X), \tag{2.3-6}
\]

\[
Q(X, Y) \leq Q(X, Z) + Q(Z,Y) \tag{2.3-7}
\]

### 2.3.1. Mahanalobis Distance

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

\[
Q_{MAHA}(X,Y) = (\mu_x - \mu_y)^T \Sigma^{-1} (\mu_x - \mu_y) \tag{2.3-8}
\]

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

### 2.3.2. 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:

\[
Q_{T SQ}(X,Y) = \frac{n_x n_y}{n_x + n_y} \sum_{i=1}^{p} \sum_{k=1}^{p} \left( \sum_{j=1}^{n_x} x_{ij} - \mu_{yi} \right) \left( \sum_{j=1}^{n_y} y_{kj} - \mu_{yk} \right) \tag{2.3-9}
\]

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

\[
C = \frac{(n_x - 1) \Sigma_x + (n_y - 1) \Sigma_y}{n_x + n_y - 2} \tag{2.3-10}
\]

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. Large values of the T-Square statistic indicate more separation between
the feature sets being compared.
2.3.3. 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:

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

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:

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

\[ g(x) = x \]  

resulting in the expression:

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

where \( S \) denotes the probability space spanned by the feature vector sets compared. When the random variables are assumed to be Gaussian (Equation (2.3-3)), the expression of the Kullback Leibler (KL) distance becomes:

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

(2.3-15)
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.

### 2.3.4. 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 Equations (2.3-12 and 2.3-13) are expressed as:

\[
\begin{align*}
  f(x) &= -\sqrt{x}, \text{ and} \\
  g(x) &= -\log(x)
\end{align*}
\]

The general form of the Bhattacharyya distance is given by:

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

where

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

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:
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 the use of 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 $d_{\text{LEVENE}}$, 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.

2.4. Speech Classification and Extraction of Voiced Speech

2.4.1. 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 can 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 signal is generally used in speaker recognition systems, without any removal of unvoiced speech, since removal result 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].

2.4.2. 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 will 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 coefficients, 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
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. 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].
2.5. Usable Speech Detection

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 that has been 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 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].

2.5.1. 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].

2.5.2. 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 Chapter Four, which deals with the development of application systems involving conversational data.

2.6. Fusion

Fusion of the distance measures such that the mutual information between them is increased is investigated in this research. In the next section, a brief introduction to the concept of fusion is given, while some techniques for fusing the distances are introduced later in Chapter Four.
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 to be used for training with 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 words, the class identity is separately predetermined 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].
2.7. Summary of the Chapter

In this chapter, background information of existing techniques applied in this research was given. In subsequent chapters, the performances of these techniques are analyzed, and the various ways in which they are implemented in the newly introduced speaker discrimination technology are presented.
CHAPTER 3
ANALYSIS OF FEATURES FOR SPEAKER DISCRIMINATION

3.1. Introduction

In the previous chapter, some common speaker discrimination features were introduced along with distance measurements for multivariate distributions. In this chapter, the performances of these distance measures and features in distinguishing speakers are investigated. The combination of several features and the use varying orders of the features, in order to increase speaker differentiability, are examined as well.

3.2. Analysis of Distance Measures

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 the distances are obtained from the data and analyzed.

3.2.1. 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 lengths, 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 conversations. Twelfth (12th) order LPCCs 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) 12th 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 3.1**.

![Figure 3.1: Block diagram showing the intra-speaker distance computation procedure](image-url)
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) 12th 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 3.2.

![Figure 3.2](image.png)

**Figure 3.2:** 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) 12<sup>th</sup> 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 3.3**.

![Figure 3.3: Block diagram showing the inter-speaker, different utterance distance computation procedure](image)

### 3.2.2. Results and Observations

The probability density functions (pdfs) of the distances obtained using the procedures illustrated in **Figures 3.1 – 3.3** were estimated based on their histograms, and the pdfs for the three cases discussed above were plotted for the five different distances discussed in Chapter Two.
Figure 3.4 shows the histograms and estimated Mahalanobis distance pdfs for the intra-speaker distances, inter-speaker, same utterance distances and inter-speaker different utterance distances. The intra-speaker distances are defined as 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.

The histograms of the Mahalanobis distances for the intra-speaker distances; inter-speaker, same utterance distances; and inter-speaker different utterance distances were observed to fit a Gaussian distribution with parameters \( \{\mu, \sigma\} \) expressed as:

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

(3.2-1)

Figure 3.4: Estimated Probability Density Functions for Intra-Speaker (SSDU), inter-speaker same utterance (DSSU) and inter-speaker different utterances (DSDU) Mahalanobis distances.
From **Figure 3.4**, it can be observed that intra-speaker (SS) and inter-speaker distances (DS) could be separated. 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 distinguishable, indicating that the Mahalanobis distance is a reliable speaker discriminating distance measure. The inter-speaker different utterance case (DSDU), 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 this distance for this research; nevertheless, further analysis of the data suggests that the speaker discriminative capability of the distance can be enhanced by combining it with distances with which it is uncorrelated. This will be discussed in subsequent chapters of this proposal.

The estimated pdfs for the T-Square distances are shown in **Figure 3.5**. The distances were observed from their histograms to follow the Gamma distribution expressed as:

$$
\gamma = f(x \mid a, b) = \frac{1}{b^a \Gamma(a)} x^{a-1} e^{x/b} 
$$

(3.2-2)

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

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

(3.2-3)

$\mu$ and $\sigma$ are the mean and standard deviation of the distances.
The KL and Bhattacharyya distance histograms were observed to follow a Gamma distribution, and the histograms and pdf estimates computed using Equation (3.2-2) and (3.3-3) are shown in Figures 3.6 and 3.7.

Figure 3.5: Estimated Probability Density Functions for Intra-Speaker (SSDU), inter-speaker same utterance (DSSU) and inter-speaker different utterances (DSDU) T-Square statistics.

Figure 3.6: Estimated Probability Density Functions for Intra-Speaker (SSDU), inter-speaker same utterance (DSSU) and inter-speaker different utterances (DSDU) Kullback Leibler distances.
For the KL and Bhattacharyya distances, both the intra-speaker distance and inter-speaker same speaker distance pdfs are practically indistinguishable as can be noted from Figure 3.6 and 3.7. 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 is, “of what good are these distances, then?” As will be illustrated, 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 for 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 (3.2-2) and (3.3-3). Note that this similarity is inevitable seeing that the computation of the Levene’s Test relies heavily
upon the T-Square statistics. The histograms and estimated pdfs of the Levene’s Test values are shown in Figure 3.8 below.

Figure 3.8: Estimated Probability Density Functions for Intra-Speaker (SSDU), inter-speaker same utterance (DSSU) and inter-speaker different utterances (DSDU) Levene’s Test values.

Figure 3.8 above clearly suggests that Levene’s Test possesses hardly any ability to distinguish between speakers whether the speech uttered by both speakers is 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.

3.3. Analysis of Features

In the previous section, the distance measures were analyzed using the 12th LPCCs as features. This feature set was chosen for the initial investigation since they are used traditionally in speaker
recognition system. In this section, the focus is on the features. In this case, the same analysis is performed using different features/feature combinations.

### 3.3.1. Procedural Set-Up

The initial investigation was to perform a comparison of LPCCs and MFCCs. The order of both features at 22. In this case, only same-speaker (intra-speaker) and different speaker (inter-speaker) distances were observed. The inter-speaker same utterance distances were not considered in this section because such comparisons are not applicable to this particular research, which involves conversational speech. The distances were computed using the same procedure outlined in Section 3.2 for each set of features.

### 3.3.2. Results and Observations

In Figures 3.9 – 3.13, the intra- and inter-speaker distances are compared for the Mahalanobis distance, T-Square statistics, KL distance, Bhattacharya distance and Levene’s test. The histograms and estimated probability density functions (computed as explained in Section 3.2) of each of the distances are shown.
Figure 3.9: Histograms and Probability Density Functions for Intra-Speaker and inter-speaker Mahalanobis distance values using LPCCs and MFCCs as features.

Figure 3.10: Histograms and Probability Density Functions for Intra-Speaker and inter-speaker T-Square statistic values using LPCCs and MFCCs as features.
Figure 3.11: Histograms and Probability Density Functions for Intra-Speaker and inter-speaker KL distance values using LPCCs and MFCCs as features.

Figure 3.12: Histograms and Probability Density Functions for Intra-Speaker and inter-speaker Bhattacharya distance values using LPCCs and MFCCs as features.
Figure 3.13: Histograms and Probability Density Functions for Intra-Speaker and inter-speaker Levene’s test values using LPCCs and MFCCs as features.

For all the distance measures, the MFCCs are shown to perform comparably with LPCCs in speaker discrimination, and even result in better separability of speakers with the KL distance, the Bhattacharya distance and Levene’s test.

3.3.3. Feature Combination

In this section, different combination of the LPCCs and MFCCs and their delta and delta-delta coefficients will be compared. This comparison procedure is carried in order to determine:

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.
The above mentioned inquiries are carried out by selectively combing the coefficients as outlined in Table 3.1.

**Table 3.1: Outline of Feature Combinations for Speaker Recognition**

<table>
<thead>
<tr>
<th></th>
<th>LPCC</th>
<th>MFCC</th>
<th>∆LPCC</th>
<th>∆MFCC</th>
<th>∆∆LPCC</th>
<th>∆∆MFCC</th>
<th>Total Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>22</td>
</tr>
<tr>
<td>2</td>
<td>22</td>
<td>11</td>
<td>11</td>
<td></td>
<td></td>
<td></td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>22</td>
<td>44</td>
<td>66</td>
<td>66</td>
<td>66</td>
<td></td>
</tr>
</tbody>
</table>

The first two sets of features are the 22\textsuperscript{nd} order LPCCs and MFCCs, respectively. The third set is a combination of 11\textsuperscript{th} order LPCCs and MFCCs – the order was chosen such that, for fair comparison, the total number of features would remain the same as the first two. The fourth feature set consists of a combination of 22\textsuperscript{nd} order LPCCs and their deltas, while the fifth consists of the same set as the MFCCs. Again, a combination of 11\textsuperscript{th} order MFCCs and LPCCs and their deltas are the sixth feature set, for fair comparison. The structure of the seventh to ninth feature sets is similar to those their predecessors, with the only difference being the inclusion of the delta-delta coefficients.

In subsequent analyses of the speaker recognition distances and features, inferences are made from the histograms and estimated 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 would be helpful. In this section, the T-test, which is a measure of similarity between two sample means, is applied in quantifying the
separation between the intra-speaker and inter-speaker distributions. 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 approximated using the Gaussian distribution, as explained earlier, and as could be observed from Figures 3.9 – 3.13, 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:

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

(3.3-1)

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

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

(3.3-2)
A higher value of T indicates greater separation between the classes being compared.

**Figure 3.14** shows the T-values for the various feature combinations outlined in **Table 3.1**, for all distances.

![Graph](image)

**Figure 3.14:** T-values for the various feature combinations: feature set index 1) 22nd order LPCCs, 2) 22nd order MFCCs, 3) 11th order LPCCs and 11th order MFCCs, 4) 22nd order LPCCs and their delta, 5) 22nd order LPCCs and their deltas, 6) 11th order LPCCs and 11th order MFCCs and their corresponding deltas, 7) 22nd order LPCCs and their deltas and delta-deltas, 8) 22nd order LPCCs and their deltas and delta-deltas, 9) 11th order LPCCs and 11th order MFCCs and their corresponding deltas and delta-deltas.

The most striking observation to be made from **Figure 3.14** above is the fact that the Mahalanobis distance and T-Square statistics are shown to have the best speaker discrimination performance out of all distances, which agrees with inferences that could be made from the distributions (**Figures 3.9-3.13**). Combining the LPCCs with the MFCCs is
clearly shown to make no notable difference in speaker differentiation. Also, the delta and delta-delta coefficients are shown to have an adverse affect on the feature sets. This is an unexpected but interesting observation, which appears to suggest that the variability captured by delta coefficients are more speech related than speaker related. Further studies on these results are beyond the scope of this research and left as topics for further research. The MFCCs appear to yield better performance than the LPCCs in all cases.

The process of obtaining the best features by feature combination could be enhanced by the use of PCA, which involves the simplification of a feature-set by reducing the number of features to the minimum necessary for accurate classification. The chosen features amongst the set are referred to as the principal components, and they are obtained by diagonalizing the sample covariance matrix $\Sigma$ of the feature vectors of a set of observable data (Jackson 1991). The diagonalization process is given by:

$$D = U\Sigma U^{-1}$$

where $U$ is the matrix of which the eigenvectors of $\Sigma$ are column vectors and $D$ is a diagonal matrix containing the corresponding eigenvalues of $\Sigma$. It can be observed that when the order of the columns of $\Sigma$ is changes, so is the order of the eigenvalues along the diagonal of $D$. The trace of $D$ is the total variance of the observed data, therefore, by arranging the columns of $U$ in order that the eigenvalues of $D$ are in descending order from the first row to the last, the features which provide the most significant contribution to the variability of the feature-set can be determined. The principal components of the feature-set can also be obtained by transforming each feature vector $x$ by $U^{-1}$.

It must be noted that although principal components are decorrelated features, they are not always statistically independent except in the case of Gaussian random variables where decorrelation is synonymous with independence (Roberts and Everson, 2001). Also, the
nonsingular linear transformation of a set of Gaussian variables results in Gaussian variables. If the transformation of the original variables yielded a resulting diagonal covariance matrix, then the resulting multivariate normal distribution are independent.

Since combining the different feature types – MFCC, LPCC and their deltas – did not yield the expected increase in performance of the feature sets, PCS was performed on the individual cepstral coefficients. It was observed that performing PCA on the coefficients was analogous to simply increasing the number of coefficients obtained during the cepstral feature computation process, i.e., an increase in the number of coefficients (features) was expected to initial yield an increase in the feature’s performance, up to a certain value, and then the addition of more features would either result in no significant change in performance, i.e., a T-value increase of less than 0.5, or a decrease in performance due to feature redundancy.

**Figure 3.15** shows the T-values for obtained for the LPCC features, the order being varied from 3 to 30.
From **Figure 3.15**, it can be observed that for the Mahalanobis and T-Square distance measures, the performance of the feature-set increases noticeably with increase in LPCC order up to the 12th order, and then the rate of increase begins to decrease until the performance is almost constant (or even decreasing as seen for the Mahalanobis distance at the 24th order).

For Levene’s test, the feature-set’s performance appears to follow the same pattern as the T-Square until about the 12th order, during which the performance continues to increase at the same rate (as opposed to a slower rate). For the KL distance, the feature performance also increases with increase in order up to the 12th order; however, after the 12th order, the performance begins to decrease almost at the same pace at which it was increasing. An interesting observation can be made with the Bhattacharyya distance, for which the performance of the features is observed to decrease with increase in order up to the 12th order, and then remain constant for a few orders, and then begin to increase gradually. This is quite...
abnormal, and could possibly be a computational error (although the same result was obtained when the experiment was repeated several times). Note that the 12\textsuperscript{th} order is the key point of change for all distance measures, suggesting that this is probably the ‘optimum’ order for LPCC computations, except in the case of the Bhattacharyya distance, where the 3\textsuperscript{rd} order appears to suffice.

**Figure 3.16** shows the T-values for obtained for the MFCC features, the order being varied from 3 to 30.

![Analyses of the Effects of Increasing the Size of the Feature Set (MFCC)](image)

**Figure 3.16**: T-values obtained by increasing the order of the MFCC coefficients from 3 to 30.

From **Figure 3.16**, it can be observed that for the Mahalanobis and T-Square distance measures, the performance of the feature-set increases noticeably with increase in MFCC
order up to the 18th order, and then the rate of increase begins to decrease until the performance is almost constant (or even decreasing as seen for the Mahalanobis distance at the 24th order). For Levene’s test, the feature-set’s performance increases until the 24th order and then becomes constant. For the KL distance, the feature performance also increases with increase in order up to the 18th order, after which the performance begins to decrease almost at the same pace at which it was increasing. Again, the Bhattacharyya distance exhibits an anomalous behavior, as it increases up to the 6th order, decreases up to the 9th, increases again to the 12th and then remains constant. In the case of MFCCs, the 18th order is shown to be the key point of change for most of the distance measures.

It is worthwhile to note that, for all distance measures, the MFCC and LPCC features exhibit very similar trends, the only difference being the order at which the performance change occurs. This analysis provides one with the ability to effectively select the order for each feature type for each of the distance measures being considered.

3.4. Summary of the Chapter

The basic performances of features and distances in distinguishing speakers were presented in this Chapter. Also, analysis of these features was performed by varying the feature sizes and also combining several features. It was observed that hardly any improvement was obtained by combining more than one type of feature. Also, surprisingly, inclusion of delta cepstral coefficients seemed to decrease the speaker differentiation performances of the cepstral features. The appropriate number of coefficients for each distance measure was determined in this chapter. It must be noted that in this chapter, data enhancements issues for the improvement of feature performances were not addressed. The topic will be covered in the next chapter.
CHAPTER 4
NOVEL SPEAKER DISCRIMINATION METHODS AND TECHNIQUES

4.1. Introduction

In this chapter, the development of new techniques for speaker discrimination is presented. A new method for the formulation of data “models” for differentiating speakers participating in conversations is introduced. Novel speaker indexing and speaker count systems for conversational data are also presented in detail in the chapter. These systems are based on the distance computation and features discussed in preceding chapters. The effect of eliminating “unusable” speech from speech signals is also be investigated.

4.2. Formation and Analysis of Data Models

In conversational data, 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. Voiced phonemes are used because it is quite improbable for two speakers’ speech to be contained in the same voiced phoneme. Moreover, most of the important information about a speaker is contained in voiced speech, as compared to unvoiced which is more noise-like in nature. Speaker homogeneous utterances are usually expected to consist of more than one voiced phoneme. In this research a voiced phoneme is defined as a concatenation of successive 30 millisecond frames of speech
which are labeled as voiced using a speech segmentation system (as discussed in Section 2.4). The phonemes that fall in this category are usually high-energy, well-structured speech sounds such as vowels and voiced fricatives. Models are formed from speech segments created by concatenating $N$ consecutive voiced phonemes, with $N > 2$, as illustrated in Figure 4.1 (for $N = 3$).

Figure 4.1: Formation of speaker homogeneous segments from voiced speech segments (v); u = unvoiced speech segments, s = silence portions.

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 performance. Nevertheless, as mentioned earlier, some applications require that short data lengths be used in order to avoid overlapping speakers in the same model [Ofoegbu et al., 2006b].

Figures 4.2 – 4.6 show the intra-speaker and inter-speaker distributions for $N$ (Number of voiced phonemes) = 20 using all the distance measures studied in this research. In this case 22rd order LPCCs and MFCCs were computed. Each homogeneous speaker model was obtained by concatenating 20
consecutive voiced phonemes from the same speaker. 1000 intra-speaker and inter-speaker distances were computed.

Figure 4.2: Mahalanobis distances obtained using homogenous speaker models formed from 20 voiced phonemes.

Figure 4.3: T-Square distances obtained using homogenous speaker models formed from 20 voiced phonemes.
Figure 4.4: KL distances obtained using homogenous speaker models formed from 20 voiced phonemes.

Figure 4.5: Bhattacharyya distances obtained using homogenous speaker models formed from 20 voiced phonemes.
Figure 4.6: Levene distances obtained using homogenous speaker models formed from 20 voiced phonemes.

A clear separation can be observed between all intra-speaker and inter-speaker distances for $N = 20$; most especially with the Mahalanobis distances, which show almost no overlap at all between the intra- and inter-speaker classes. 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. In other words, the least number of segments with adequate separation is desired.
Figure 4.7 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) had an average length of 200 milliseconds. The basic idea is to determine the least number of phonemes that will yield no overlap between the standard deviations. From Figure 4.7, it is observed 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.

![Speaker Differentiation with Respect to Data Size](image)

**Figure 4.7:** Comparison of Mahalanobis Distances for different data sizes. X-axis represents the number of voiced phonemes used to form each model.

Figures 4.8 – 4.12 shows the intra-speaker and inter-speaker distributions for $N = 5$ using the same parameters as in Figure 4.2-4.6. In this case, the separation between the intra-speaker and inter-speaker distances is shown to have decreased from what was obtained for $N=20$, as expected, however, a considerable amount of separation does still exist.
Figure 4.8: Mahalanobis distances obtained using homogenous speaker models formed from 5 voiced phonemes.

Figure 4.9: T-Square distances obtained using homogenous speaker models formed from 5 voiced phonemes.
Figure 4.10: KL distances obtained using homogenous speaker models formed from 5 voiced phonemes.

Figure 4.11: Bhattacharyya distances obtained using homogenous speaker models formed from 5 voiced phonemes.
The above figures show that the Mahalanobis distance is superior to all other distance measures, with Levene’s distance providing the least separation. In Chapter Three, the effect of increasing the number of coefficients for each feature was investigated using 5-second utterances, and without the new model formation technique (the entire data was used). Five voiced phonemes is contained in much less than 5-seconds, as it would be wrong to assume that, in conversations, each speaker will always speak for up to five seconds before the another speaker speaks. Figures 4.13 and 4.14 show the T-values obtained by varying the order for the LPCC and MFCC features, respectively, from 3 to 30. Speaker models were formed using the new 5-voiced-phoneme based model formation technique.
Analyses of the Effects of Increasing the Size of the Feature Set  
(Voiced - Segment Based - LPCC)

Figure 4.13: Voiced-phoneme based T-values obtained by increasing the order of the LPCC coefficients from 3 to 30.

Analyses of the Effects of Increasing the Size of the Feature Set  
(Voiced - Segment Based - MFCC)
Figure 4.14: Voiced-phoneme based T-values obtained by increasing the order of the MFCC coefficients from 3 to 30.

From Figures 4.13 and 4.14, it is observed that, even though the amount of data has been reduced (as compared with Figures 3.9 and 3.10 where entire utterances were used), slight improvement in the t-test values for the intra- and inter-speaker distances are observed. For the LPCCs, the 12th order is shown to be ideal for the Mahalanobis and Bhattacharyya distances. The ideal LPCC orders for the T-Square, KL and Levene’s distances are 15th, 9th and 30th, respectively. For the MFCCs, the 30th order is apparently ideal for the T-square and Levene’s distances. The 21st is ideal for the Mahalanobis, the 15th for KL, and the 24th for the Bhattacharyya distance. These ideal orders along with their corresponding T-statistics are outlined in Table 4.1.

Table 4.1: Outline of Optimal Number of Coefficients for Voiced-Phoneme Based Feature Computation.

<table>
<thead>
<tr>
<th>Distances</th>
<th>LPCC</th>
<th>MFCC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ideal # of Coefficients</td>
<td>Intra/Inter Speaker T-Statistic</td>
</tr>
<tr>
<td>Mahalanobis</td>
<td>12</td>
<td>25</td>
</tr>
<tr>
<td>T-Square</td>
<td>15</td>
<td>20.25</td>
</tr>
<tr>
<td>KL</td>
<td>9</td>
<td>14.8</td>
</tr>
<tr>
<td>Bhattacharyya</td>
<td>12</td>
<td>18.5</td>
</tr>
<tr>
<td>Levene</td>
<td>30</td>
<td>13</td>
</tr>
</tbody>
</table>

Note that the optimal number of coefficients depends on the distance measure used. Henceforth in this research, the number of coefficients for the LPCC features was set at 12 for all distance measures (except Levene’s, where the 30th order was used because there was a great difference between the T-statistic for the 12th and 30th order - Figures 4.13 and 4.14). For the MFCCs, the order was set at 22nd (except for KL, where the 15h order was used for the same reason as with Levene’s in the case of the LPCCs). This was done for computational ease/uniformity. However, in the actual evaluation of the systems, the optimum number of coefficients was used.
4.3. 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.

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 (3.2-3) for the Gamma case. 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 – about 1000 comparisons – 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 otherwise they are assured to be from different speakers. A Distance Likelihood Ratio (DLR) can thus be defined as:

$$DLR = \frac{f(x|\alpha_1, \beta_1)}{f(x|\alpha_2, \beta_2)} \quad (4.3-1)$$

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 model 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:

$$D_{rel} = d_{\text{min}} - d_{\text{max}} \quad (4.3-2)$$

If the value of $D_{rel}$ 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 4.15.
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](image)

**Figure 4.15**: 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. 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.

Initial experiments were performed using the SWITCHBOARD telephone conversation database [Godfrey et al., 1992] in order to determine an appropriate value for $N$, the number of voiced phonemes used for each speaker homogeneous model. 245 conversations were used, each with an average duration of about 400 seconds. In this case, the RRMD indexing technique was not applied. Instead, the two maximally separated models were selected as references, and the remaining were matched to the closest reference. Evaluation was performed using transcriptions provided by the
Mississippi Switchboard Transcription Project [Hamaker et al., 1998]. The percent accuracy values were determined as follows:

- An index of 1 or 2, representing each of the two speakers in the conversation, was assigned to each speech sample based on the ground truth transcriptions.

- An index of 1 or 2 was assigned to each speech sample based on Minimum Distance (MD) matching (not RRMD) of the reference models.

- Two accuracy computations were obtained as follows: let $L_T$ be the total number of ground truth samples, and let $L_{xy}$ be the number of samples with ground truth index = $x$ and MD index = $y$.

\[
Accuracy_1 = \frac{L_{11} + L_{22}}{L_T} \quad (4.3-3)
\]

\[
Accuracy_2 = \frac{L_{21} + L_{12}}{L_T} \quad (4.3-4)
\]

The maximum of the two accuracy values was taken as the correct accuracy, expressed as:

\[
Accuray = \text{Max}(Accuracy_1, Accuracy_2) \times 100\% \quad (4.3-5)
\]

Figure 4.16 shows the percent accuracy (of all 245 conversations) of the basic MD indexing procedure with respect to $N$, the number of phonemes used, for the T-Square statistics. Note that, as expected, the accuracy increases as $N$ increases, up to a certain value and then begins to decrease, most likely due to the fact that models are being formed using phonemes of different speakers. From Figure 4.16, the optimum value for $N$ is observed to be 5. Note that a similar trend was observed for all other distances, with very slight deviations in the number of phonemes that yielded the maximum accuracy percentage.
Using the optimal value of $N$, determined from the process described above, speaker indexing was then performed using the RRMD technique, the Drel threshold, which controls the amount of undecided speech samples, was varied for each distance based on its inter-class distributions. Two types of error were determined, one being the indexing error and the other 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. These experiments and results are discussed in further detail in Chapter Five.

The proposed RRMD speaker indexing system was designed with the assumption that the number of speakers participating in the conversation is known \textit{a priori}, however, it is possible to encounter situations in which more than two speakers are present in the conversation, and the number of speakers is unknown. In determining the number of speakers in such cases, a speaker count system could 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 conversational data could be very challenging especially when no information about the

\textbf{Figure 4.16:} Average percent indexing accuracy with respect to number of voiced phonemes used to form models.
conversation is known. In the next section, a novel technique for determining the number of speakers in a given conversation is presented. This method is then extended to a speaker-count/indexing system for multiple speakers.

4.4. 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.

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 \textit{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 \textbf{Figure 4.17}. 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.

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{residual-ratio-algorithm.png}
\caption{How the Residual Ratio Algorithm Works for Two-Speaker Conversations}
\end{figure}

\textbf{Figure 4.17:} 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 4.18, which shows the same process as in Figure 4.16, but for three speakers.

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 4.19 and 4.20 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 4.17 and 4.18.

Figure 4.19: 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 4.20: 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).

In determining the speaker count based on the Residual Ratios computed, a classification approach which involves determining the speaker count based on the sum of the Residual Ratios for all K-1 rounds was implemented. 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. Appropriate thresholds for the ARR method were obtained from Figure 4.21, which shows the estimated distributions of ARR values using the training conversations described above.
Figure 4.21: Added Residual Ratio distribution functions for 1000 artificially generated conversations each for 1-4 speaker cases

From Figure 4.21, it can be inferred that the speaker count accuracy will decrease as the number of speakers in the conversation increases, because a relatively small separation between the ARR distributions for three and four speakers is observed.

This speaker count system was evaluated on the HTIMIT database as well as a newly created conversations database, and results are presented in Chapter Five.

4.5. Speaker Count-Indexing System

The speaker count-indexing system (an extension of the GRRA speaker count system) is explained below:
After the number of speakers in the conversation has been determined using the K-1 elimination steps in the GRRA, the conversation is indexed as follows:

- Models that initially matched the valid reference models are given the same index (i.e., considered to be of the same speaker) as the reference models.
- Let the speaker count be $C$, if $C < K$, then each unmatched (residual) model is assigned to the model amongst the first $C$ reference models from which it has minimum distance.
- If $C = K$, then Step iv of the RRA is repeated for the $K^{th}$ round. The models which matched the $K^{th}$ reference are assigned the same index as that reference, and the unmatched models are assigned to the model amongst the first $C-1$ reference models from which it has minimum distance.

This system was also evaluated on the HTIMIT database as well as a newly created conversations database. Results are presented in the next chapter.

### 4.6. Fusion of Distance Measures

In the formulation of the distances (presented in Chapter 2), it was 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 section, the relationships between distances are examined in order to determine and exploit their complimentary information and 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 derived.
4.6.1. 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 the scatter plot of each distance is plotted against the others 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 4.22 and 4.23 show the draftsman display of the Mahalanobis distances, the T-Square statistics, the KL distances, the Bhattacharyya distances and Levene’s test values obtained using $14^{th}$ order LPCCs and $22^{nd}$ order MFCCs, respectively. The distances were computed between speaker homogeneous models formed as described in Section 4.2, 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 4.22: 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 4.22, it is observed that a nonlinear relationship exists between the T-Square and Mahalanobis distances. The KL and Bhattacharya 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 2.3, and were both derived from the same root, i.e. equations (2.3-12 and 2.3-13). 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 uncorrelated to each other.
Figure 4.23: Draftsman display of distances obtained using 14th order MFCCs as features. 1000 intra-speaker and 1000 inter-speaker distances were computed using speech files from the HTMIT database.

Figure 4.23 depicts the same nonlinear relationship between the T-Square and Mahalanobis distances, as in Figure 4.22, revealing that the correlation exists regardless of the feature used. The correlation between the KL and Bhattacharya distances is, however, much less conspicuous with the MFCCs. Again, all other distance relationships show significant uncorrelation.

4.6.2. Optimized T Distance

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, was developed. The goal was to minimize the variance of each class and maximize the difference between their means. This could 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 (3.3-1)).
An optimal linear combination of the distances can be computed by solving the equation:

$$T_{\text{max}} = a^T X$$ \hspace{1cm} (4.6-1)

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 maximized 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:

$$T(a) = \frac{\mu_1 - \mu_2}{\sigma_1^2 - \sigma_2^2}$$ \hspace{1cm} (4.6-2)

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:

$$a = \frac{k}{\lambda_1} P^{-1} (\mu_1 - \mu_2)$$ \hspace{1cm} (4.6-3)

Where $k = \lambda_1 \left\| P^{-1} (\mu_1 - \mu_2) \right\|^{-1}$ \hspace{1cm} (4.6-4)

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:

$$P^{-1} Q a = \lambda_1 a$$ \hspace{1cm} (4.6-5)

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

Figures 4.24 and 4.25 illustrate the effect of combining all distances using the above technique (for LPCCs and MFCCs respectively). The $T_{\text{max}}$ intra- and inter-speaker distances were computed between the two classes using Equation (4.6-1), and their distributions are
shown in the top row of the figures. 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.

**Figure 4.24:** Illustration of the effect of the weighted combination of distances – LPCCs used as features. Top panel shows the maximized T distance.
Figure 4.24: Illustration of the effect of the weighted combination of distances – MFCCs used as features. Top panel shows the maximized T distance.

Figures 4.23 and 4.24 show that the $T_{\text{max}}$ distance yields the highest separation between the two classes for both sets of features. This increase, however, is not as much as was anticipated, and is significantly less than those observed within the different distance measures by themselves (the increase in T value is only 1.5). The combinational effect is shown to be higher with the MFCCs, where the T-value increases from about 27 to about 28.3, as opposed to an increase from about 25 to about 25.3 obtained with the LPCCs. Note,
also, that the MFCCs are more effective in separating the classes, as they result in higher T-values than the LPCCs.

In addition to examining the effect of combining the distances, it was also important to study the contribution made by each of the distances in maximizing the class separation. This was 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 4.26 (LPCC) and 4.27 (MFCC). 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 simply testing all the distances each time before selecting the next one to be included.
Figure 4.26: Illustration of the effect of increasing the number of distances – LPCC features used
Figure 4.26 shows that with the LPCCs, Levene’s test is the most uncorrelated with the Mahalanobis distance, and provides an increase of about 0.6 (the most significant increase in separation) in $T$ value when combined. The combination of other distances has very little effect on the improvement of the distance measure. With the MFCCs (Figure 4.27), although Levene’s test is also the most uncorrelated, inclusion of the other distance has about the same effect on the $T_{\text{max}}$ distance – indicating that the distances are more uncorrelated with the MFCCs. Practically, however, the LPCCs are clearly more efficient as they result in higher separation even with one distance, and require fewer distances to attain the highest improvement (in this case). Note that, with the LPCCs, using any other distance but Levene’s test with the LPCCs would result in less increase in the $T$-value of the combination. With the
MFCCs, however, changes in the order of the distances after the Mahalanobis did not significantly affect the combined performance. Note that even though the KL and Levene’s distances were the least effective in terms of the T-statistic between the inter-class distributions, they do offer some complimentary information, and could therefore improve the performance of the systems when combined with other distances. During evaluation of the systems, these two distances are only used in combination with the other distances, and not by themselves.

This fusion technique was applied to the speaker count and speaker count-indexing systems. Results are presented in Chapter Five.

4.6.3. Decision-Level Fusion

A decision level fusion technique was also examined in this research. This involved a voting technique, where the comparison decision made by the majority of distance measures was taken as the final matching decision. For instance, if three of the distances decided that two models were from the same speaker, then those two models were considered a match. A modification to this voting technique was also investigated. In this case, weights were assigned to each distance measure based on the T-value between their intra-speaker and inter-speaker values computed using a significantly large dataset (over 1000 distance comparisons). The inter-class T-values obtained for each distance were normalized such that they summed 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 was made. For instance, for the each feature, a weight (denoted by $\omega_i \{i = 1, 2, 3, 4, 5\}$ representing the Mahalanobis, T-Square, KL, Bhattacharyya and Levene’s distances respectively) were assigned to the each distance using the equation:
\[ \omega_i = \frac{T_i}{\sum_i T_i} \]  \hspace{1cm} (4.6-6)

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

The speaker count and speaker count-indexing systems were enhanced using this fusion approach. Results are presented in the next chapter.

4.7. Summary of the Chapter

In this chapter, new data enhancement and speaker modeling techniques were introduced. Also, three new application-specific speaker discrimination techniques were presented in detail, and brief illustrations of their performances were given. In the next chapter, the results obtained from the evaluation of the proposed speaker discrimination on standard and newly developed speech databases systems are presented and analyzed.
5.1. Introduction

Evaluation of this research effort is presented in this chapter. A detailed description is given of each of the three databases which were used in obtaining results for the speaker discrimination systems developed. Experimental procedures for each of the systems are presented, and results obtained using all distance measures, features, and distance fusion techniques will be shown and discussed. Cross evaluation of the speaker count and speaker count-indexing systems will also be performed for the standard and new databases.

5.2. Databases

The databases used in the evaluation of this research can be categorized into two groups – the standard databases used generally in testing speaker recognition systems and a Temple Conversations Database created specifically for use in discriminating multiple speakers in a conversations.
5.2.1. Standard Databases

1. **HTIMIT**: The HTIMIT corpus [Reynolds, 1997] was designed specifically for speaker recognition of telephone speech. It was created by playing speech data from an already existing speech database, TIMIT [Fisher et al., 1986] through several telephone handsets. It consists of 384 single-speaker utterances – 192 males and 192 females. Each of the 384 speakers was represented by 10 phonetically rich sentences, two of which were repeated for all speakers. The average length of each sentence was about five seconds. A possible concern with the HTIMIT is that the speech files were recorded through a loudspeaker, resulting in the imposition of unwanted frequency response on the signals. The total number of files in the HTIMIT database is 3840. Each file was sampled at 8KHz, and the sampling format was single channel, 16-bit linear.

The HTIMIT corpus was used for initial investigations of distance and feature performances. It was also used as a standard database to obtain all testing parameters such as distance thresholds and pdf parameters. Although the database consists of individual utterances and not conversational data, it was also used for evaluation. Artificial conversations were formed using conversation statistics (such as pdfs of homogeneous speaker utterance lengths, interspeaker pause lengths and intra-speaker pause lengths) from SWITCHBOARD, a telephone-based conversations database [Iyer et al., 2006b].

2. **SWITCHBOARD**: The SWITCHBOARD [Godfrey et al., 1992] database consists of about 2400 two-speaker telephone conversations from 543 speakers (302 males and 241 females). The data was created from speakers from all areas of the United States. The average length of the conversations was about 400 second. The calls were handled by a computer-operated robot which prompted the caller, dialed the callee, presented the discussion topic, and recorded the speech from the two speakers until the end of the conversation. About 70 topics were discussed. The following constraints were made on the conversations: (1) no two speakers
participated in the same conversation more than once and (2) no participant discussed any given topic more than once.

Conversations were selected at random from the SWITCHBOARD and used for evaluation of the proposed speaker discrimination systems. Groundtruth labels indicating each speaker’s utterances as well as pauses and portions of overlap were provided by the Mississippi Switchboard Transcription Project [Hamaker et al., 1998]. Analysis of the groundtruth labels was performed, which showed that mode speaker utterance length for conversations was about 1 second [Iyer et al., 2006b].

5.2.2. Temple Conversations Database

Due to the unavailability of conversational data with more than two speakers, a new conversation database was created for evaluating the speaker count and speaker count indexing systems. This database consists of 40 (10 each) one to four speaker monologues/conversations.

Recording Setting: All conversations were recorded in the speech processing laboratory of Temple University College of Engineering. Four recording sessions were held within a span of two weeks, and participants were asked to either hold a discussion on a topic that was assigned about three days prior to the recording session, or act as characters from TV series for which they were given scripts about three days prior to recording.

Participants: The database consists of speech from 26 participants (10 females and 16 males). All participants were Temple University students, and they ranged in age from 19 – 35, the average age being about 22. Participants were also from very diverse backgrounds. About 4 of them hailed from India, 4 of them were from Africa, 1 of them was Chinese, 4 were African American, and 10 of them were Caucasian American. The average length of the
conversations was 1 minute. Each participant was recorded in at least two of the one to four speaker databases described below.

**One-Speaker Database:** This consists of 10 different monologues recited by 5 speakers (3 male and 2 female). There were two monologues for each speaker. The monologues were obtained from scripts from TV series and sent to participants three days prior to recording. The average length of the one-speaker files was 1 minute.

**Two-Speaker Database:** This consists of 10 different dialogues from 8 speakers (4 females, 4 males). The database is comprised of two male-male, two female-female and six male-female conversations. Nine of these conversations were from dialogues from TV series, while one was based on a discussion topic given to the participants three days prior to the recording sessions.

**Three-Speaker Database:** This consists of 10 different conversations from 17 speakers (7 females, 10 males). The database is comprised of one all-male conversation, one all female conversation, two conversations with 2 males and one female, and six conversations with 2 females and 1 male participant. Nine of these conversations were from TV series, while one was based on a discussion topic given to the participants three days prior to the recording sessions.

**Four-Speaker Database:** This consists of 10 different conversations from 12 speakers (5 females, 7 males). The conversations were comprised of four conversations with 3 males and 1 female, and six conversations with two males and two females.
Speech Quality, Formatting and Specifications: Each speech file was sampled at 16 KHz on a single channel with 16-bit resolution. The laboratory consisted of several computers and air conditioning/heating systems, and could therefore be considered an uncontrolled environment. Moreover, the wireless microphone with which the recording was performed was placed about 15 feet away from the recording station. The average signal-to-noise ratio (SNR) of each of the conversations recorded in the main laboratory was about 10dB, which is considered relatively poor for automatic speech processing purposes.

In spite of the poor quality of the signals, the Temple Conversations Database was used in evaluating the speaker count and speaker count-indexing systems. The same testing parameters used for the HTIMIT were used for this database. This was done to ensure that the evaluation procedure was data independent.

5.3. Experiments and Results

Experimental procedures and results obtained from evaluating the each system using the databases discussed in the previous section are presented in this section.

5.3.1. Unsupervised Speaker Indexing System

The intra-speaker parameters $\alpha_1$ and $\beta_1$, and the inter-speaker parameters, $\alpha_2$ and $\beta_2$, used in the DLR tests were obtained from 1000 (each) model-based intra- and inter-speaker distances computed using the HTMIT database (for each distance measure). Models were formed for $N = 5$, based on the accuracy plot shown in Figure 4.16. Speaker indexing was then performed on 100 five-minute conversations selected at random from SWITCHBOARD. The RRMD technique was used, with the $D_{rel}$ threshold (which controls the amount of undecided speech
samples) varying between different ranges for different distance measures depending on the means of their intra- and inter-speaker distributions.

Two types of error were determined, one being the indexing error, computed as \( I_{err} = 100 - Accuracy \), determined using equation (4.3-5). The other error was 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, \( L_T \). In other words, Let \( N_u \) be the number of samples declared undecided/unusable by the proposed technique, and let \( N_c \) be the number of samples amongst them that were labeled as co-channel data from the ground truth transcriptions. The ‘undecided error’ was computed as:

\[
U_{Err} = \frac{N_u - N_c}{L_T} \times 100\% 
\]

(Figure 5.3-1)

Figures 5.1 to 5.3 show the average percent error with respect to varying \( D_{rel} \) thresholds for both types of errors using the three best performing distances, namely: the Mahalanobis distance, the T-Square statistics, and the Bhattacharyya distance, respectively. The ‘a’ figures were obtained using LPCCs as features while the ‘b’ figures were obtained using MFCCs.
Figure 5.1a: Average percent indexing accuracy with respect relative distance threshold. Indexing was performed using the Mahalanobis distance with the RRMD Technique. Features used – 12th order LPCC

Figure 5.1b: Average percent indexing accuracy with respect relative distance threshold. Indexing was performed using the Mahalanobis distance with the RRMD Technique. Features used – 22nd order MFCC
Figure 5.2a: Average percent indexing accuracy with respect relative distance threshold. Indexing was performed using the T-Square distance with the RRMD Technique. Features used – 14th order LPCC

Figure 5.2b: Average percent indexing accuracy with respect relative distance threshold. Indexing was performed using the T-Square distance with the RRMD Technique. Features used – 28th order MFCC
Figure 5.3a: Average percent indexing accuracy with respect to relative distance threshold. Indexing was performed using the Bhattacharyya distance with the RRMD Technique. Features used – 12th order LPCC.

Figure 5.3a: Average percent indexing accuracy with respect to relative distance threshold. Indexing was performed using the Bhattacharyya distance with the RRMD Technique. Features used – 22nd order LPCC.
The trade-off of higher indexing accuracy for loss of ‘usable’ (non-co-channel) data is evident in all the above figures. For the maximum observed threshold, a minimum indexing error is obtained, corresponding to a maximum ‘undecided error’. With the LPCCs, the Mahalanobis distance gives the best performance, with an equal error rate of slightly less than 6%. The Mahalanobis distance also gives the best performance with the MFCCs, yielding an equal error rate of about 6%. The LPCCs outperform the MFCCs for all distances in these experiments. This is quite surprising, considering the fact that the inter-class t-statistics for the MFCCs appear higher than LPCCs in all cases. Note the relatively very poor performance obtained when the Bhattacharyya distance is used with MFCCs as features. This could be predicted from the low t-statistics observed between the same and different speaker distributions (Figure 4.25).

5.3.2. Speaker Count System

Experiments were performed using artificial conversations from the HTIMIT database (constructed as described in the previous chapter), as well as the Temple Conversations Database. Results were obtained using both the LPCCs and MFCCs as features. The Mahalanobis, T-Square and Bhattacharyya distances were used, as well as the three distance fusion techniques. The intra-speaker parameters $\mu_1$ and $\sigma_1$, and the inter-speaker parameters, $\mu_2$ and $\sigma_2$, used in the DLR tests were obtained from 1000 (each) intra- and inter-speaker Distances. A maximum count of $K = 4$ speakers was considered in this research. Each conversation was about 60 seconds in length, and each speaker contributed an approximately equal amount of speech.

The speaker count accuracy was determined in three different ways as described below:

1. One or more speakers: an accurate count was obtained if there was one speaker and the proposed system yielded a speaker count of 1, or if there were two, three or four speakers, and the proposed system yielded a speaker count greater than 1.
2. One, two or more speakers: an accurate count was obtained if there was one or two
speakers and the proposed system yielded a count of one or two, respectively, or if
there were two or three speakers and the proposed system yielded a count greater
than 2.

3. One, two, three or four speakers: an accurate count was considered if the proposed
system yielded the correct number of speakers.

The accuracy rate of the system was obtained as the ratio of the number of correct speaker
counts to the total number of conversations.

Results obtained from each of the databases are present below:

**HTIMIT:**

The speaker count system was tested on 4000 testing conversations having the same statistics
as the data used in obtaining the parameters. conversations of the database. Testing
parameters were the same as those used for the HTIMIT database. **Figures 5.4a** (LPCC) and
5.4 (MFCC) show the accuracy rates for each of the distance measures.
From Figures 5.4a and 5.4b, it can be observed that the performance of the proposed technique diminishes with increase in the complexity of the task (as a measure of the number of speakers to be determined). The Mahalanobis distance yields the best performance while
the Bhattacharyya distance yields the worst. The LPCCs are also observed to outperform the MFCCs.

Figures 5.5a (LPCC) and 5.5b (MFCC) show a comparison of the accuracy rates for the Mahalanobis distance and the combined distances obtained using the three fusion techniques namely: the optimized T distance (OTD), the decision-based combination (DBC), and the weighted decision-based combination (WDBC).

![Speaker Count Accuracy - LPCC](image)

**Figure 5.5a:** Speaker count accuracy comparison for distance fusion techniques. Results obtained using the HTIMIT database and LPCCs as features.
Figure 5.5b: Speaker count accuracy comparison for distance fusion techniques. Results obtained using the HTIMIT database and MFCCs as features.

From Figures 5.5a and 5.5b it can be observed that the optimum T-test distance computation technique (OTD) produces hardly any improvement to the performance of the Mahalanobis distance. The best performing combination technique is the weighted decision-based combination technique, which offers a maximum accuracy increase of 3% from the Mahalanobis distance, for the case of LPCCs when counting up to four speakers.

Temple Conversations Database:
The speaker count system was tested on all 40 conversations of the database. Testing parameters were the same as those used for the HTIMIT database. Figures 5.6a (LPCC) and 5.6b (MFCC) show the accuracy rates for each of the distance measures. While Figures 5.7a (LPCC) and 5.7b (MFCC) show a comparison of the accuracy rates for the Mahalanobis distance and the combined distances obtained using the three fusion techniques.
Figure 5.6a: Speaker count accuracy for each distance. Results obtained using the Temple Conversations Database and LPCCs as features.

Figure 5.6b: Speaker count accuracy for each distance. Results obtained using the Temple Conversations Database and MFCCs as features.

106
Figure 5.7a: Speaker count accuracy comparison for distance combination techniques. Results obtained on the Temple Conversations Database using LPCCs as features.

Figure 5.7b: Speaker count accuracy comparison for distance combination techniques. Results obtained on the Temple Conversations Database using MFCCs as features.
From the above figures, it can be observed that only the T-square distance performs well in the speaker count systems. When the distances are combined, the systems performance is clearly degraded as compared to when the T-square distance is used alone. The performances of both the MFCCs and LPCCs are similar in this case.

Cross evaluation:

From Figures 5.4 to 5.7, it can be observed that the performance of the speaker count system on the HTIMIT database is much higher than on the Temple Conversations Database. This is shown more clearly in Figure 5.8, which shows the best case scenario for each distance. For the HTIMIT, results are shown for the weighted decision-based distance computation and the LPCCs features. For the Temple Conversations Database, results are shown for the T-square distance and the MFCCs.

Figure 5.8: Cross evaluation of the speaker count system using the HTIMIT and Temple Conversations (TC) databases.
It must be noted that even though the results for the Temple database are inferior to those for HTIMIT, accuracies of up to 80% were obtained were obtained using the TCD, and this is considered acceptable due to the amount of noise (10dB SNR) present.

5.3.3. Speaker Count-Indexing System

For the speaker count-indexing system, the same files were used as in the case of the speaker count system. Only the T-square distance and the MFCC features were used for the Temple Conversations Database. For the HTMIT database, the LPCCs were used since they outperformed the MFCCs in all previous evaluations.

Figure 5.9, shows the indexing accuracy results obtained for both databases. The black solid bars represent the accuracy results for the HTMIT while the black and grey checkered bar represents the result for the new database.

![Figure 5.9: Indexing accuracy results obtained for both databases. The black solid bars represent the accuracy results for the HTMIT while the black and grey checkered bar represents the result for the new database.](image-url)
Note that the weighted decision-based combination technique outperforms all other techniques for HTIMIT, and the performance of the system is much better for the HTIMIT database than for the Temple Conversations Database.

5.4. Summary of the Chapter

In this chapter, this research effort was evaluated using two standard speech databases and a newly created conversations database. Results were obtained using different distance measures, distance combination techniques and features. In general, the LPCCs were shown to outperform the MFCCs for both standard databases. Also, the Mahalanobis distance was observed to yield superior performance as compared to other distances. When distance combination techniques were applied, however, the weighted decision-based combination technique outperformed other techniques. It is quite interesting to note that the most complex fusion technique, the optimized T approach, yielded the least improvement to the systems’ performance.

With the Temple Conversations Database, all distances except the T-square distance seemed to produce poor results, with the Mahalanobis distance, which appeared to be the best distance for the standard databases. Also, the distance combination technique yielded poor results for this database. This could be explained by the fact that the weights for the distances were obtained using parameters from the HTIMIT database. This was done so as to ensure that results were not data-dependent. Since the distances did not have the same trend of performances for the new database, the weights failed when used. For the simple voting technique, which used no weights, the performance degradation could be explained by the fact that there were more poor performing distances. It should be noted that the speaker count and speaker count/indexing systems were also tested on an actual conversations database recorded in a controlled environment, which was not discussed for privacy reasons, and
results obtained were comparable to those obtained with the HTIMIT database, with reduction in accuracy ranging from 1%-3%.

Error rates of less than 10% were observed in spite of the lack of information and the unavailability of sufficient data to represent speakers. For the new databases, the system’s performance could be applauded considering the fact that acceptable results were obtained even though all testing parameters were obtained using the standard databases. Moreover, the quality of the speech files used was extremely poor. To some extent, the systems could be considered somewhat robust.
6.1. Summary

The goal of this research effort was to address the following problems encountered when attempting to differentiate between speakers in conversational data:

• No \textit{a priori} information about the speakers
• Limited data size available for representing each speaker
• No knowledge of speaker change points
• Co-channel speech

In this dissertation, a novel technique for the formation of speaker models from the conversational data (since models are not available to the systems \textit{a priori}) was introduced. The models were formed using the LPCCs and MFCCs, which are conventional speaker discrimination features. Experiments were performed to determine the best number of coefficients to use for each of these features. Models were compared using five different distance measures – the Mahalanobis, T-square, Kullback-Leibler, Bhattacharyya and Levene’s distances. The best number of coefficients for each feature was observed to vary with the different distance measures used. In order to improve results obtained in differentiating between speaker models, three distance fusion techniques were formulated. Finally, three novel application-specific speaker discrimination systems were designed namely: an unsupervised two-speaker indexing system, a speaker count system for counting multiple speakers in a
conversation, and a speaker count-indexing system four indexing multiple speakers by first determining the number of speakers in the conversation.

The newly introduced systems were evaluated using three databases. Two of them are standard speaker recognition databases, while the third is a newly created multiple-speaker conversations database created for the purpose of cross evaluation.

For the two-speaker indexing system, a minimum equal error rate of 6% was obtained when it was evaluated using SWITCHBOARD. This was achieved using the Mahalanobis distance and the LPCCs as features. For the speaker count and speaker count-indexing systems, accuracy values of over 90% were obtained for the HTIMIT database. The best performing feature in this case was the LPCCs, and the weighted decisions-based distance combination technique yielded the highest accuracies. With the new database, however, the systems performed poorly except for the T-square distance. With the T-square distance, up to 80% accuracy was obtained, which is reasonably good, although inferior to those obtained using the HTIMIT database. Note that the conversations were recorded in highly noisy settings, and results could be improved if noise reduction technique were to be applied on the signals. This could be considered a possible area of further research.

In this research, conversations-based speaker discrimination systems have been developed, and evaluation of these systems have shown that they could be compared to conventional speaker recognition systems which do not face the same challenges as with conversations-based systems. A direct comparison cannot be made with these conventional systems, since the problems are not the same; however, it might be worthwhile to note that state-of-the-art speaker identification and verification systems, which have been trained using information from all participating speakers yield accuracies of about 97% when tested on standard non-conversation databases. On the other hand, state-of-the-art speaker discrimination systems for indexing of broadcast data or two-speaker conversations have been shown to yield accuracies of about 80-90%.
6.2. Further Research

One major challenge faced in this research was the unavailability of standard multiple-speaker conversations databases for cross-evaluation. A further research topic could be to develop data enhancement techniques that will improve the performance of the system on conversations recorded in highly noisy environments. Some other possible areas for further research include (1) the investigation of prosodic speaker discrimination features such as intonation, rhythm and speaking rate, to be used in combination with the cepstral features, (2) the improvement of model formation techniques by determining speaker change-points \textit{a priori}, (3) the use of individual phonemes for speaker model formation and comparison, and (4) an investigation of the use of unvoiced speech, combined judiciously with voiced speech to increase speaker information.
BIBLIOGRAPHY


