EFFECTIVENESS OF LOW-FREQUENCY PARAMETERS FOR AUTOMATIC SPEAKER VERIFICATION.

MARIO CARLOS. VIDALON

University of Windsor

Follow this and additional works at: https://scholar.uwindsor.ca/etd

Recommended Citation
https://scholar.uwindsor.ca/etd/1842

This online database contains the full-text of PhD dissertations and Masters' theses of University of Windsor students from 1954 forward. These documents are made available for personal study and research purposes only, in accordance with the Canadian Copyright Act and the Creative Commons license—CC BY-NC-ND (Attribution, Non-Commercial, No Derivative Works). Under this license, works must always be attributed to the copyright holder (original author), cannot be used for any commercial purposes, and may not be altered. Any other use would require the permission of the copyright holder. Students may inquire about withdrawing their dissertation and/or thesis from this database. For additional inquiries, please contact the repository administrator via email (scholarship@uwindsor.ca) or by telephone at 519-253-3000ext. 3208.
The quality of this microfiche is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us a poor photocopy.

Previously copyrighted materials (journal articles, published tests, etc.) are not filmed.

Reproduction in full or in part of this film is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30. Please read the authorization forms which accompany this thesis.

**THIS DISSERTATION HAS BEEN MICROFILMED EXACTLY AS RECEIVED**
EFFECTIVENESS OF LOW FREQUENCY PARAMETERS FOR AUTOMATIC SPEAKER VERIFICATION

by

Mario Carlos Vidalon

A Dissertation
Submitted to the Faculty of Graduate Studies through the Department of Electrical Engineering in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy at the University of Windsor

Windsor, Ontario, Canada 1976
ABSTRACT

An analysis of the spectral properties of male speech reveals that most of the spectral energy is distributed in the frequency range of 50-2000 Hz. In this dissertation an attempt is made to develop a computer algorithm for speaker verification that utilize parameters extracted directly from the digitized speech. The original speech signal was sampled at 4 kHz after filtering it with a 2 kHz low-pass filter. The investigations reveal that by the use of low-order linear predictor model a feasible set of features could be realized for application to speaker verification. Also parametric representation of speech derived from a short-time spectrum analysis and the use of a new concept of composite reference for speaker verification were studied.

A simple linear shifting warping procedure modifies the parameter contour of the unknown speaker so that the correlation with the reference contour is maximized. The verification decision was based on a set of selected distances of the test sample contour from the reference contour for the claimed speaker. If the overall distance was less than a fixed threshold,
then the speaker was accepted.

Among all the parameters investigated, the PARCOR (reflection) and the autocorrelation coefficients were found to be the most effective, providing a verification accuracy of 98% for speech 2 sec. in duration, which increased to more than 98% for a duration of 3 sec. In addition, verification results obtained using composite reference and measurement of energy at specific bands of frequency look promising.
ACKNOWLEDGMENTS

The completion of this research would not have been possible without the generous help of many people. I am particularly grateful to my supervisor, Dr. M. Shridhar, for his invaluable guidance, support, and encouragement. In addition, the author wishes to express his appreciation to Dr. G. A. Jullien for writing the sampling routine and to all the "speakers" who cheerfully participated in the experiment. Thanks are due to Mrs. Judy Assaf, who did an excellent job of typing this dissertation.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF ILLUSTRATIONS</td>
<td>ix</td>
</tr>
<tr>
<td>LIST OF APPENDICES</td>
<td>xiii</td>
</tr>
<tr>
<td>CHAPTER</td>
<td></td>
</tr>
<tr>
<td>I. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Analysis-Synthesis of Speech</td>
<td>1</td>
</tr>
<tr>
<td>1.2 State of the Art in Speaker Verification</td>
<td>2</td>
</tr>
<tr>
<td>1.3 Definition of the Problem</td>
<td>12</td>
</tr>
<tr>
<td>1.4 Questions to be Answered</td>
<td>14</td>
</tr>
<tr>
<td>1.5 Organization of Dissertation</td>
<td>16</td>
</tr>
<tr>
<td>II. SELECTION OF PARAMETERS</td>
<td>18</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>18</td>
</tr>
<tr>
<td>2.2 Desirable Properties of Speech Parameters</td>
<td>18</td>
</tr>
<tr>
<td>2.3 Categories of Speech Parameters</td>
<td>19</td>
</tr>
<tr>
<td>2.4 Description of Speech Parameters</td>
<td>22</td>
</tr>
<tr>
<td>III. ALGORITHM FOR SPEAKER VERIFICATION</td>
<td>29</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>29</td>
</tr>
<tr>
<td>3.2 Parameter Contour</td>
<td>30</td>
</tr>
<tr>
<td>3.3 Time Registration</td>
<td>41</td>
</tr>
<tr>
<td>3.4 Development of the Reference Contour</td>
<td>44</td>
</tr>
<tr>
<td>3.5 Computational Distances</td>
<td>48</td>
</tr>
<tr>
<td>3.6 Threshold Computation</td>
<td>61</td>
</tr>
</tbody>
</table>

vi
Table of Contents (Cont.)

IV. VERIFICATION USING COMPOSITE REFERENCE .......... 65
   4.1 Introduction ............................................. 65
   4.2 Construction of Composite Reference Contour ............ 66
   4.3 Algorithm for Speaker Verification Using Composite Reference ............................................. 67
   4.4 Alternative Procedure for Verification Using Composite Reference ............................................. 76

V. EXPERIMENTAL RESULTS .................................... 82
   5.1 Data Collection ........................................... 82
   5.2 Experimental Set-Up ...................................... 84
   5.3 Verification Results Using LPC Parameters .............. 86
   5.4 Verification Time ......................................... 93
   5.5 Verification Based on Composite Reference .............. 93
   5.6 Verification Accuracy as a Function of Duration of Speech ............................................. 96
   5.7 Verification Results Using Frequency-Domain Parameters ............................................. 98
   5.8 Discussion ................................................. 121

VI. CONCLUSIONS .............................................. 122

APPENDICES ..................................................... 127

BIBLIOGRAPHY .................................................. 156

VITA AUCTORIS .................................................. 160
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Set of Distances Using Speaker JN Reference</td>
<td>62</td>
</tr>
<tr>
<td>3.2</td>
<td>Set of Distances Using Speaker FE Reference</td>
<td>63</td>
</tr>
<tr>
<td>4.1</td>
<td>Set of Distances Using Composite Reference File for Sentence 4</td>
<td>80</td>
</tr>
<tr>
<td>4.2</td>
<td>Set of Distances Using Composite Reference File for Sentence 5</td>
<td>81</td>
</tr>
<tr>
<td>5.1</td>
<td>Verification Error Rates</td>
<td>90</td>
</tr>
<tr>
<td>5.2</td>
<td>Processing Time for Single Verification</td>
<td>92</td>
</tr>
<tr>
<td>5.3</td>
<td>Composite Reference Results</td>
<td>95</td>
</tr>
<tr>
<td>Illustration</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>--------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>3.1 Block Diagram for Speaker Verification</td>
<td>31</td>
<td></td>
</tr>
<tr>
<td>3.2 Set of Parameter Contours for Sentence 1</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>3.3 Set of Parameter Contours for Sentence 2</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>3.4 Set of Parameter Contours for Sentence 3</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>3.5 Set of Parameter Contours for Sentence 4</td>
<td>36</td>
<td></td>
</tr>
<tr>
<td>3.6 Set of Parameter Contours for Sentence 5</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>3.7 Discrete Linear Prediction Model of Speech Production</td>
<td>39</td>
<td></td>
</tr>
<tr>
<td>3.8 Contour Time Registration for Speaker JN</td>
<td>45</td>
<td></td>
</tr>
<tr>
<td>3.9 Contour Time Registration for Speaker FE</td>
<td>46</td>
<td></td>
</tr>
<tr>
<td>3.10 Design Set and Trial Reference for Reflection Coefficient $b_6$</td>
<td>49</td>
<td></td>
</tr>
<tr>
<td>3.11 Design Set and Trial Reference for Reflection Coefficient $b_3$</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>3.12 Design Set and Trial Reference for Linear Prediction Coefficient $a_3$</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>3.13 Design Set and Trial Reference for Autocorrelation Coefficient $\phi_7$</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>3.14 Registered Design Set and Final Reference for Reflection Coefficient $b_6$</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>3.15 Registered Design Set and Final Reference for Reflection Coefficient $b_3$</td>
<td>54</td>
<td></td>
</tr>
<tr>
<td>3.16 Registered Design Set and Final Reference for Linear Prediction Coefficient $a_3$</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>Illustration</td>
<td>Page</td>
<td></td>
</tr>
<tr>
<td>------------------------------------------------------------------------------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>3.17 Registered Design Set and Final Reference for Autocorrelation Coefficient $\phi_7$</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td>4.1 Design Set for Three Speakers and Their Trial Composite Reference Contour</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>4.2 Registered Design Set for Three Speakers and Their Final Composite Reference Contour</td>
<td>70</td>
<td></td>
</tr>
<tr>
<td>4.3 Design Set and Trial Subtracted Reference for Reflection Coefficient $b_3$</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>4.4 Registered Design Set and Final Subtracted Reference for Reflection Coefficient $b_3$</td>
<td>74</td>
<td></td>
</tr>
<tr>
<td>4.5 Composite Reference, Original and Subtracted Contours for Sentence 5</td>
<td>77</td>
<td></td>
</tr>
<tr>
<td>4.6 Composite Reference, Original and Subtracted Contours for Sentence 4</td>
<td>78</td>
<td></td>
</tr>
<tr>
<td>5.1 Block Diagram of Experimental Set-up</td>
<td>85</td>
<td></td>
</tr>
<tr>
<td>5.2 Verification Accuracy as a Function of Duration of speech</td>
<td>97</td>
<td></td>
</tr>
<tr>
<td>5.3 Design Set and Trial Reference Contour for Energy Parameter Computed up to 1 kHz in Bands of 16 Hz</td>
<td>99</td>
<td></td>
</tr>
<tr>
<td>5.4 Registered Design Set and Final Reference Contour for Energy Parameter Computed up to 1 kHz in Bands of 16 Hz</td>
<td>100</td>
<td></td>
</tr>
<tr>
<td>5.5 Parameter Contours for Bands of Energy</td>
<td>101</td>
<td></td>
</tr>
<tr>
<td>Illustration</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>-------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>5.6</td>
<td>Registered Design Set and Final Reference Contour for Spectral Energy for 750 to 1550 Hz</td>
<td>104</td>
</tr>
<tr>
<td>5.7</td>
<td>Spectral Energy Contours for 5 Bands</td>
<td>105</td>
</tr>
<tr>
<td>5.8</td>
<td>Spectral Energy Contours for 6 Bands</td>
<td>106</td>
</tr>
<tr>
<td>5.9</td>
<td>Spectral Energy Contours for 8 Bands</td>
<td>107</td>
</tr>
<tr>
<td>5.10</td>
<td>Design Set of the 9th Power Spectrum Amplitude Parameter</td>
<td>110</td>
</tr>
<tr>
<td>5.11</td>
<td>Design Set of the 9th Square Power Spectrum Amplitude Parameter</td>
<td>111</td>
</tr>
<tr>
<td>5.12</td>
<td>Registered Parameter Contours and Final Reference Contour for the 9th Power Spectrum Amplitude</td>
<td>112</td>
</tr>
<tr>
<td>5.13</td>
<td>Registered Parameter Contour and Final Reference Contour for the 9th Square Power Spectrum Amplitude</td>
<td>113</td>
</tr>
<tr>
<td>5.14</td>
<td>Power Spectrum Amplitude Contours up to 10 Frequency Points</td>
<td>114</td>
</tr>
<tr>
<td>5.15</td>
<td>Design Set of the 28th Power Spectrum Amplitude Parameter</td>
<td>117</td>
</tr>
<tr>
<td>5.16</td>
<td>Registered Parameter Contours and Final Reference Contours for the 28th Power Spectrum Amplitude</td>
<td>118</td>
</tr>
<tr>
<td>5.17</td>
<td>Design Set for the 28th Square Power Spectrum Amplitude</td>
<td>119</td>
</tr>
<tr>
<td>5.18</td>
<td>Registered Parameter Contours and Final Reference Contour for the 28th Square Power Spectrum Amplitude</td>
<td>120</td>
</tr>
</tbody>
</table>
List of Illustrations (Cont.)

<table>
<thead>
<tr>
<th>Illustration</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>Schematic Diagram of the Human Vocal Mechanism</td>
<td>129</td>
</tr>
<tr>
<td>A2</td>
<td>Block Diagram of a Simplified Model of Speech Production</td>
<td>132</td>
</tr>
<tr>
<td>A3</td>
<td>A Fortran Program for Solving the Auto-correlation Equations</td>
<td>141</td>
</tr>
</tbody>
</table>

xii
# List of Appendices

<table>
<thead>
<tr>
<th>Appendix</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Mechanism and Model of Speech Production</td>
<td>127</td>
</tr>
<tr>
<td>B</td>
<td>Linear Predictor Parameters Estimation</td>
<td>134</td>
</tr>
<tr>
<td>C</td>
<td>Brief Comparison of Speaker Verification and Speaker Identification</td>
<td>142</td>
</tr>
<tr>
<td>D</td>
<td>Brief Description of Experimental Set-Up</td>
<td>147</td>
</tr>
<tr>
<td>E</td>
<td>Instructions to the Speakers</td>
<td>153</td>
</tr>
<tr>
<td>F</td>
<td>Fortran Program Listings for Speaker Verification</td>
<td>156</td>
</tr>
</tbody>
</table>
LIST OF SYMBOLS

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_k$</td>
<td>kth predictor coefficient</td>
</tr>
<tr>
<td>$b_k$</td>
<td>kth PARCOR (reflection) coefficent</td>
</tr>
<tr>
<td>$\phi_k$</td>
<td>kth autocorrelation coefficient</td>
</tr>
<tr>
<td>$R(k)$</td>
<td>autocorrelation matrix function</td>
</tr>
<tr>
<td>$\psi(k)$</td>
<td>covariance matrix function</td>
</tr>
<tr>
<td>$F_k$</td>
<td>kth formant</td>
</tr>
<tr>
<td>UTT</td>
<td>utterance</td>
</tr>
<tr>
<td>LPC</td>
<td>linear predictor coefficient</td>
</tr>
<tr>
<td>P.S.A.</td>
<td>power spectrum amplitude</td>
</tr>
<tr>
<td>$E(t)$</td>
<td>energy function</td>
</tr>
<tr>
<td>ZCR</td>
<td>zero crossing rate</td>
</tr>
<tr>
<td>$w(t)$</td>
<td>window function</td>
</tr>
<tr>
<td>$s(t)$</td>
<td>speech signal function</td>
</tr>
<tr>
<td>FFT</td>
<td>Fast Fourier Transform</td>
</tr>
<tr>
<td>DFT</td>
<td>Discrete Fourier Transform</td>
</tr>
<tr>
<td>$H(z)$</td>
<td>all-pole filter</td>
</tr>
<tr>
<td>$S_n$</td>
<td>nth speech sample</td>
</tr>
<tr>
<td>$\hat{S}_n$</td>
<td>nth predicted speech sample</td>
</tr>
<tr>
<td>$u(t)$</td>
<td>excitation function</td>
</tr>
<tr>
<td>${x_i}$</td>
<td>the set of elements $x_i$</td>
</tr>
<tr>
<td>$\rho$</td>
<td>coefficient of correlation</td>
</tr>
<tr>
<td>$D$</td>
<td>overall distance</td>
</tr>
<tr>
<td>$D_T$</td>
<td>threshold distance</td>
</tr>
</tbody>
</table>

xiv
**List of Symbols (Cont.)**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E_d )</td>
<td>( i )th exponential distance</td>
</tr>
<tr>
<td>NSEG</td>
<td>number of segments</td>
</tr>
<tr>
<td>SD</td>
<td>standard deviation</td>
</tr>
<tr>
<td>( u )</td>
<td>mean value</td>
</tr>
<tr>
<td>C.R.</td>
<td>composite reference</td>
</tr>
<tr>
<td>COR</td>
<td>correlation</td>
</tr>
<tr>
<td>PARCOR</td>
<td>partial correlation</td>
</tr>
</tbody>
</table>
CHAPTER I
INTRODUCTION

1.1 Analysis-Synthesis of Speech

In the past few years, a number of research investigations have been carried out to analyse and reproduce human speech with the use of Digital Computers. Generally, most of these efforts have been in the direction of reproducing basic human speech processes such as recognition, verification, and understanding. In speech recognition, for example, the different linguistic elements of a given acoustic signal can be identified by human analysis using the knowledge of the contextual, grammatical and semantic constraints of the given language. On the automatic level, it has not been possible to match human ability as far as the precise application of linguistic constraints are concerned. As of this day, it is only possible to simulate approximately the initial operations performed on the acoustical signal by the human. The main reason for not achieving a full automatic speech recognition, as the human does it, is the lack of knowledge of the linguistic constraints. Research in speech synthesis has been providing most of the insight for the knowledge of linguistic

---

1 A model and mechanism of speech production is discussed in Appendix A.
constraints. Speech synthesis experiments have also provided knowledge in finding speech characteristics that are effective for automatic speaker recognition. Early work in speech synthesis was centered around speech processing devices such as the vocoder. The results obtained when using these devices were quite good as far as intelligibility was concerned; but the techniques were deficient with respect to the recognizability of speakers. This failing provided a strong motivation for the subsequent work aimed towards obtaining speaker-dependent information from the speech signal. Some of the speech characteristics that have been investigated are spectrographic data, formants, pitch, gain, linear prediction coefficients and related parameters. A survey of the fundamental studies for full automatic speaker recognition follows.

1.2 State of the Art in Speaker Verification

The literature in speaker verification can be broadly divided into two categories:

1) Speaker recognition systems with prescribed text; and

2) Speaker recognition systems that are text independent.

The majority of the automatic speaker verification systems are text dependent. However, there are some limited efforts in the line of verification with no prescribed text.
For example, in the text dependent case, if the speech material has a number of speech events that can be extracted using segmentation and recognition techniques, it becomes somewhat text independent.

The text independent case is close to what we as human listeners go through in order to recognize people by their voices. For example, we are able to recognize people over the telephone even though the spoken text is different from one occasion to another. Further work is needed to be able to understand a speaker verification system that is text independent in the human sense.

In the early stages of verification systems, visual analysis of spectrograms were used to discriminate among speakers\(^{(1,2,3)}\). The experimental error rates obtained using this system varied anywhere between 1% to 30%, depending upon the experimental constraints. Best results were obtained by sorting a set of spectrograms of a given utterance into a known number of speaker sets, each containing a known number of spectrograms, and poorest results were obtained for matching a single spectrogram with one of a set of reference of spectrograms, with the possibility that no match exists. Speaker recognition by audition, on the other hand, gives better results than recognition by visual analysis of spectrograms\(^{(4)}\). In a recent experiment\(^{(5)}\), 40 speakers were asked to utter an all-voiced
sentence, "we were away a year ago". The verification task was carried out by human listeners who were asked to verify if a test and reference utterance belonged to the same or different talkers. The human listeners achieved approximately 96% correct verification.

One of the main drawbacks of aural or voiceprint verification is that human ability to pick up the clues used for speaker verification is absent. In order to replace human expertise in the process of verification, efforts have been directed to finding an automatic decision algorithm. The same population of 40 speakers and the same speech utterance were used in an automatic verification system (6). The automatic method made use of three different speech parameters such as formants, pitch and intensity. The verification accuracy obtained by this method was 98%.

A comparison between human performance and machine performance in verification tasks is difficult to assess. There is no common base for comparison, since the speech data base is different and most of the constraints under which the verification tests are carried out are different. However, one can safely say that, under limited conditions, the performance of the automatic method is at least comparable, if not better, than auditory methods.

A survey of automatic methods for verification in the last decade follows. Since most of the useful information
for speaker verification comes from speech synthesis work, the most frequently used techniques for speaker recognition were based on filter bank systems. Filter banks provide an efficient overall analysis of speech signals.

The spectral representation of speech obtained from a bank of 15 bandpass filters, with frequency range of 300 to 4000 Hz, was used by Li in a verification system. A two-level adaptive linear threshold element system was used to perform speaker discriminations. At the first level, four different kinds of training processes take place. As a result, a set of weights are obtained from the frequency bands and time segments. The weights are speaker dependent. At the second level the decision process takes place. Three different test phrases were employed: "My name is (name)", "Alibi", and "I.D. Reference (name)". The two phrases where the speaker's name are used were chosen so that the talker would emphasize the pronunciation of his name rather than emphasizing on the time segment of the phrase. A large number of specimen utterances were collected over the telephone lines using standard telephone hand sets in typical office acoustical environments. Over 90% accuracy in the verification system is reported.

Cepstral measurements over short segments of speech were used in an automatic speaker verification system by James E. Luck. The cepstral measurements were obtained for each
of the first two vowels of the standard test phrase "My code is ...". The length of the word "my" and the pitch information were used as additional parameters. The above parameters are collected under the reference data and stored. They represent the normal variations of a given speaker's voice. The decision algorithm is based on multidimensional distance measures between the test sample and the nearest reference sample. If this comparison yields a 'distance' which is below a given threshold value, then the identity of the speaker is accepted. A population of four authorized (true) speakers and 30 impostors were examined in the verification procedure with error rates ranging from 6% to 13%. Collection of specimen utterances for the construction of reference data, over a period of time to take into account the variations in the speaker's voice from one recording session to another is suggested. Also, it has been suggested that impostors attempting to mimic a true speaker are not able to improve their ability to fool the system significantly. Professional mimics were not incorporated and the rehearsal time was minimal.

Another similar technique using the spectral representation of speech obtained from filter banks is one reported by Das and Mohn (9). In addition to the speech parameters used by Li, it incorporates formant information and uses a comprehensive and quite complicated segmentation procedure.
The segmentation technique is essentially a pattern recognition scheme in which various events in the speech utterances are identified. The alignment of the utterances is carried out using the identified speech events. After alignment, parameters are extracted from the segmented utterance and used either to construct the reference information for a given speaker or as a test to compare against the reference profile of the claimed identity. A total of approximately 100 utterances, spanning a period of five recording sessions, was obtained for each of 50 true speakers. They were divided into two equal groups of 50 utterances for the purposes of training and recognition. Sixty-eight impostors provided their specimens of 20 replications in a single session. The recordings were made in a sound booth and the phrase used was "check available terminals". An equal-error rate of about 1% was obtained. But unfortunately, this low rate was corrupted by a 10% of 'no decision' rate. The difficulty with this kind of system was due to a faulty segmentation scheme.

A novel non-linear time-warping was introduced by Doddington\(^6\), and this method smoothed out some of the problems that had been encountered in the previous methods of segmentation. The non-linear warp was achieved by a steepest-ascent algorithm. Pitch, intensity, and formant frequency values were extracted directly from the speech
waveform and used as features. A crucial part of the algorithm is the automatic time registration of the sample contour to the reference contour retrieved as part of the reference data of the claimed identity. Registration, based on second formant, modifies the parameter contour of the unknown speaker so that the correlation with the reference contour is maximized. The decision algorithm is based on a set of distances, computed with several heuristically chosen formulas, which measure the dissimilarity between the sample and reference contours. A population of 40 male speakers who uttered the phrase "we were away a year ago" achieved a verification of 98.5%. Identical twins included in the experiment were differentiated 100% of the time.

A subsequent experiment on the same population of 40 speakers was carried out by R. C. Lummis (10). Registration was based on the intensity contour. Also, a more sophisticated set of distances was used to measure the dissimilarity between the registered sample contour and the reference contour. An individual utterance was represented by the following parameters: voice pitch, low-frequency intensity and three lowest formant frequencies. Verification based solely upon voice pitch and intensity achieved an average error rate below 1%.

Formant data has been shown to be significant in reducing the mimic acceptance rate. In a study conducted by
Lummis and Rosenberg\textsuperscript{(11)}, 27\% of the utterances of four professional mimics were accepted. With the omission of formant data, the mimic acceptance rate increased to 41\%. Consequently, if a verification system is going to be mimic-resistant, additional sophistication is required.

Linear prediction parameters were added to the already existing pitch and intensity analysis of sentence-long utterances by Rosenberg and Sambur\textsuperscript{(12)}. In addition, a new method of overall distance computation and an enlarged sample of male speech utterances of two sentences, "we were away a year ago" and "I know when my lawyer is due", were used during the experimentation. A verification error rate of approximately 1\% with respect to casual impostors and 4\% with respect to well-trained mimics was achieved.

Another investigation of parametric representation of speech derived from the linear prediction model is that of B. S. Atal\textsuperscript{(13)}. His experiment included predictor coefficients and speech parameters derived from them, such as the impulse response function, the autocorrelation function, the area function and the cepstrum function. The speech data consisted of 60 utterances, made up of six repetitions of the same sentence spoken by ten speakers. All the speakers were female and they spoke the sentence, "may we all learn a yellow lion roar". Among all the parameters investigated, the cepstrum was found to be the most effective.
The cepstrum provided a verification accuracy of approximately 83% for 50 m sec. of speech, increasing to 98% for speech duration of one sec. In a separate study to determine the feasibility of text-independent speaker identification, an identification accuracy of 93% was achieved for speech two sec. in duration even though the texts of the test and reference samples were different.

Among recent experiments, a large-scale operating control system has been installed by Doddington at Texas Instruments\(^{(14,15)}\). Using a four-phase sequential decision strategy, 180 speakers, of which 13% were female, made an average of more than 400 entries a day. The average number of utterances required per transaction is 1.3 phrases.

The average verification time is 5.8 seconds. Because of the fast execution time, the Texas Instrument automatic speaker verification system allows the comfortable use of a multiphase sequential decision strategy. In the sequential decision strategy, the speaker is accepted after the first phrase if the decision function is less than some threshold. Otherwise, additional phrases are requested combining the decision functions at each stage. A verification accuracy of 0.3% was achieved for customer reject rate and 1% for impostor accept rate.

In another recent experiment, M. R. Sambur of Bell Laboratories\(^{(16)}\) has introduced a new verification system
which does not require any time normalization or segmentation. A set of orthogonal parameters are formed by a linear transformation of the linear prediction parameters (Sambur (17)). The set of orthogonal parameters are essentially independent of all linguistic information but are highly indicative of the identity of the speakers. The speech data consisted of six repetitions of the sentence, "I was stunned by the beauty of the view", by 21 male speakers on six separate occasions spaced over a three-week period. Among the various linear prediction parameters examined, the orthogonal parameters derived from the parcor or reflection coefficients yielded the best results. A verification accuracy of 99.2% was obtained when only the seven least significant orthogonal parameters were used in the distance computation. Finally, A. E. Rosenberg also of the Bell Laboratories (18), has evaluated an automatic speaker-verification system accessed by test customers from their own telephones over dialed-up lines. The test population consisted of over 100 male and female speakers who called up nominally once each working day over a period of five months. The system was based on an acoustic analysis of a fixed, sentence-long utterance, "we were away a year ago", resulting in a parameter contour for each feature analyzed. Features selected for analysis were pitch and intensity. The system was implemented on a NOVA 800
laboratory computer. Telephone line access to the computer was via a data set hook up. Identity claims were made by keying an identification number on a Touch-Tone dial. Instructions and responses to the customer were made by means of a programmed voice-response system. Reference data was computed off-line and updated with the new analysis data of accepted utterances. A verification accuracy of approximately 90% for new customers, and approximately 95% for adapted customers was obtained.

In conclusion, speaker recognition systems have reached a turning point. Despite the existence of large gaps in our knowledge with respect to speaker recognition techniques, we are able to conclude that automatic speaker recognition within a small speaker population is possible. Further, a number of experiments have demonstrated that a high accuracy for speaker discrimination can be achieved, at least under laboratory environments. A generalized system of man-machine communication by natural speech is not too far away.

1.3 **Definition of the Problem**

Speaker verification is a process by which

1) a sample utterance of an unknown speaker and a claimed identity is given;
2) A more detailed comparison of speaker verification and speaker identification is given in Appendix B.

A decision of acceptance or rejection is reached with respect to the similarity between the sample and reference utterance associated with the claimed identity.

The sample utterance is obtained by a request to speak a pre-assigned sentence. The speaker, being a cooperative one, does not make any attempt to alter his speaking behaviour from trial to trial. The identity claim can be initiated by a keyed-in identification number. The similarity between the sample and reference contours is obtained by comparing a set of measurements of the sample contour with the reference file. Only one comparison of the features is needed regardless of the number of persons in the population. Because of this number of comparisons, there exists a finite probability of an incorrect decision. However, in a more general problem, that is, of speaker identification, given n speakers in the population, one has to compare the sample contour with each of the reference files. Because of this number of comparisons, there exists a finite probability of an incorrect decision. This probability of an incorrect decision is generally independent of the population size.
The single comparison of unknown sample to reference contour and a simple accept or reject decision, makes speaker verification a practical and more tractable problem. In a practical implementation of a speaker verification system the important limitations are storage capacity and speed in accessing the reference file. However, with the sophistication of computers, the commercial use of these systems is in sight. Some of the possible areas of applications are the conducting of business transactions over the telephone, identity checks, etc.

1.4 Questions to be Answered

The purpose of this Thesis was briefly to investigate:

1) whether a sampling frequency of 10 kHz is essential, especially since most of the spectral energy is within 2 kHz;

2) which of the linear predictor coefficients is most sensitive to speaker variation as opposed to speech variation;

3) whether there is a linear transformation of the linear predictor parameters that can yield greater reliability;

4) the development of a software package for performing the parameter extraction and for computing the measure of similarity between the unknown speaker's sample
and the reference contour;

5) A time warping technique based strictly on correlation. The procedure should be such that the speech events of a given parameter contour are brought into the best possible alignment with the corresponding speech events of the reference contour.

The linear prediction characteristics provide a representation of the spectral envelope of the speech signal. Thus they represent the combined information about the formant frequency, their bandwidth, and the glottal waveform. Another advantage of using linear prediction characteristics or a linear transformation of them is that they are easily determined directly from the digitized speech signal and they are reasonably reliable for measuring speaker sensitive features of the speech waveform.

As previously mentioned, researchers have conducted verification experiments to determine the speaker verification effectiveness of predictor coefficients as opposed to that of formant and bandwidth data. Their conclusions show that the linear prediction coefficients were slightly more informative and that they were a more effective choice of verification parameters as far as computation time was concerned. In this dissertation these results are confirmed. Furthermore, we show that lower order analysis is
sufficient for a reasonable verification system. A sixth order analysis and a 4 kHz sampling rate was used. This system offers an optimum compromise between storage requirements, error performance and computation time. A simple warping procedure was used in order to bring into alignment the parameter contour of the unknown speaker with the reference contour of the identity claimed so that the correlation between them is maximum.

1.5 Organization of Dissertation

Chapter II deals with the selection of suitable parameters for speaker verification. Properties that take into account the fixed physical characteristics of a speaker such as his vocal apparatus, as well as the important aspects of his learned habits of speaking, are presented. Discussion of the advantages and disadvantages of time-varying and steady state acoustic parameters follows. A brief description of the most important speech parameters that have been used for speaker verification is also presented.

Chapter III describes the basic tasks of the speaker verification algorithm. The speaker verification algorithm does not use the speech waveform itself, but instead deals with the parametric representation of the signal. The extraction of these acoustic parameters that were used in our experiment are described. The important property of
automatic time registration of the contours of the sample utterance to the contours retrieved as part of the reference data is discussed. Construction of the reference contour and distances which quantify the dissimilarity between the sample and reference contour measurements are also presented.

Chapter IV describes the new concept of composite reference and its use in an automatic speaker verification scheme. An advantage of using a single parameter contour, namely the composite reference, which comprises the information of the identities of several speakers is that it represents an extra feature for verification purposes. This extra feature can be used in a verification procedure to guard against accepting voice samples that might otherwise be accepted as true speakers. The verification task consists, as in the conventional verification procedure, in authenticating an individual's claimed identity by the analysis of his speech. Construction of the composite reference contour and description of the algorithm for speaker verification using this new concept are also presented.

Chapter V presents all the results obtained from our experiments. Verification as a function of the duration of spoken material is also reported. In addition, a brief discussion and results obtained with parameters derived from the short-time spectrum analysis of speech are presented.
Chapter VI contains the conclusions of the automatic speaker verification scheme that was implemented on a laboratory computer, NOVA 840.
CHAPTER II

SELECTION OF PARAMETERS

2.1 Introduction

The problem of speaker verification may be divided into two parts: measurement and classification. In the first part, given a specified speech utterance, a number of measurements are made so as to be able to obtain a set of parameters which, hopefully, will characterize the speaker. In the second part, a classification scheme is implemented by the application of appropriate decision rules to verify if the measurements belong to the speech of the claimed speaker.

In the following sections we discuss the different aspects for a good measurement.

7.2 Desirable Properties of Speech Parameters

One of the most important tasks for a successful speaker verification system is that of the selection of an efficient set of acoustic parameters. They should be efficient in discriminating speakers and amenable to automatic implementation. The majority of the speech characteristics are measured directly from the waveform or after spectral transformation to the frequency domain.
Any selection of them for purposes of speaker verification should take into account the unique properties of a speaker's vocal apparatus, as well as the important aspects of his learned pattern of speaking. Ideally, the chosen speech parameters should have the following properties (19):

1) should be highly indicative of the identity of the speaker;
2) be easily measurable;
3) occur naturally and frequently in speech;
4) not change over time;
5) change little from one speaking environment to another;
6) not be susceptible to mimicry.

In a real implementation of a verification system, the simultaneous fulfillment of these criteria is beyond the actual state of the art. Present verification tests which are mainly conducted in laboratory environment acknowledge partial relaxation of some of these standards. In our experiment, the last two factors were not investigated, but were controlled.

2.3 Categories of Speech Parameters
A typical acoustic parameter generally can fall into either of the following two categories: steady state (time-
invariant) or time-varying.

Time-invariant parameters can be obtained by two methods:

a) by a suitable averaging process of the time-varying behaviour of the parameter, or

b) by measuring the fixed anatomical properties of a speaker's vocal apparatus.

The disadvantages of the first method is that time-averaging of certain voice characteristics can be easily mimicked. If the entire time-variant behaviour of the parameter is taken into account, however, mimicry becomes much more difficult. Also, many speaker-dependent characteristics are affected by learned speech habits. Time-invariant parameters cannot represent these useful speech characteristics, since the idiosyncrasies in the speaking habits of individuals vary from one utterance to another.

The main advantage of time-invariant parameters is that they contain almost no linguistic information. Consequently, they are suited for text-independent speaker verification analysis.

Time-varying parameters can be obtained over

a) long enough intervals to be context independent, or
b) specific contexts.
Measurements that are obtained continuously as a function of time result in a large set of data that have a high degree of redundancy. Such difficulties can be solved by using techniques such as the 'F ratio' or divergence tests (19, 20, 21). Measurements over short intervals at appropriately chosen locations in the speech utterance give much smaller data sets and avoid the problem of redundancy. Further, it includes advantageous information of phoneme-specific effects and most of the learned characteristics. However, the automatic location of certain speech events of interest is a difficult problem.

All results presented in our experiment have been accomplished by the use of parameters measured at regularly spaced intervals throughout a given entire utterance. In this procedure, the measurements are not optimally suited to every speech segment of the utterance. Some of them will result in useless information such as that which occurs during a pause or silence interval. To avoid this problem, the measurements during voiceless intervals are removed. Also, because of the timing differences occurring within different speech utterances of a single speaker, corresponding speech events do not occur at exactly the same times. Consequently, the comparison of parameter values, at points where the
utterances are out of alignment, will result in severe errors. To solve this problem, approximate time synchronization is introduced by aligning the beginning and the end of the two parameter contours. Further, the synchronization is improved by an additional operation of time registration. In this operation, the events of one of the parameter contours are brought into the best possible alignment with corresponding events of the second parameter contour so that the correlation between them is maximum.

2.4 Description of Speech Parameters

This section describes the acoustic parameters that are useful for a speaker verification system. Parametric characterization of speech signal can be derived by simple measurements on the waveform itself or on the frequency domain characterization.

A brief discussion of some acoustic parameters follows.

1) Energy: One of the simplest representations of a signal is its energy. For a time-varying energy calculation of a nonstationary signal such as speech, energy is defined as

$$ E(t) = \int_{t-T/2}^{t+T/2} s^2(\tau) d\tau $$

(2.1)
where $T$ is the measurement interval. The choice of $T$ is quite important. If it is less than a pitch period, the energy will fluctuate very rapidly and, if it is a multiple of a pitch period, then the energy will vary very little, and will not reflect the changing properties of the speech signal. A value between 10 to 30 m sec. is adequate in most cases.

A direct application of energy measurement is in the differentiation of voiced speech segments from unvoiced speech segments. The silent portions and pauses in an utterance are eliminated from the analysis interval by the use of energy measurements.

2) Zero crossing: Zero crossing is another very simple measurement in the time domain. It is defined as the number of zero crossings in a fixed-frame length. Its implementation is straightforward and it is often used as a gross estimate of the frequency content of a speech signal.

In a digital implementation of zero crossing measurements, it is important to observe that if one requires a fine resolution, then the sampling rate has to be high. Also, to avoid noise, dc offset, etc., it is better to use a bandpass filter rather than a low-pass filter prior to sampling. Some of the applications of this kind of measurement are:
a) zero crossing rate (ZCR) measurements along with energy are used for making a decision about whether a particular segment of speech is voiced or unvoiced. If the zero crossing rate is high, then the segment of speech is unvoiced. Otherwise, it is voiced.

b) Zero crossing measurements along with pitch information are useful for estimation of excitation parameters.

c) Zero crossing measurements are also useful as representation of speech signals for speech recognition.

3) Pitch: Pitch is the fundamental frequency of the vocal-cord vibrations. It can be measured directly in the time domain by determining the period of the speech waveform or in the frequency domain by measurements on the frequency spectrum of the signal.

The variations of the fundamental frequency are used, along with the formants, for synthesizing speech. Also, pitch information has been used for speech and speaker recognition.

4) Formant Frequencies and Bandwidths: Formants are more difficult parameters to measure. They are defined as the resonance frequencies of the vocal tract.
There are several methods to determine the formants from the speech waveform. One approach is to estimate formants by searching the smooth spectra for peaks and then to decide which peaks correspond to formants (22). The problem with methods based on peak picking is that often times two peaks will be close enough to each other so that only one distinct peak will be found in the spectrum. In such a case the spectrum is enhanced by the chirp Z-transform. This technique gives better formant resolution by evaluating the vocal tract transfer function on a contour which passes closer to the poles, thereby sharpening the resonances (38).

Formants are speaker dependent so that they are very useful for speaker verification.

5) Predictor Coefficients: Linear prediction analysis is perhaps one of the most useful methods of speech analysis. Its importance is due to the accuracy and speed in characterizing the important spectral properties of speech in the time domain. Basically, in this method each sample of speech can be approximated by a linear combination of the past p samples. The predictor coefficients are determined by minimizing the squared difference between the actual speech samples and the linearly predicted ones. Typically, 12 predictor coefficients are adequate for speech which is band-limited to
5 kHz. For speaker recognition purposes, it has been shown that a lower order analysis and calculation of the predictor coefficients once every 20 msec is sufficient. Thus, the effective storage requirements and the computation time can be reduced without affecting significantly the verification error rates.

6) Short-time Spectrum: Given a speech signal \( S(t) \), the short-time power spectrum is defined as

\[
G(f, t) = \left| \int_{-\infty}^{\infty} S(\lambda)w(t-\lambda)e^{-j2\pi f\lambda}d\lambda \right|^2
\]

(2.2)

where \( w(t) \) is the window function \(^{(27)}\). By the above definition, the function \( G(f, t) \) contains information about the frequency, time and energy of the speech signal. Consequently, it provides most of the acoustical characteristics of speech. The reason it is called short-time analysis is because, over a sufficiently short time interval, the speech signal can be considered to be stationary. Thus, the Fourier transformation of a short segment of speech should give a good spectral representation of the speech during that time interval. Generally, the short-time spectrum analysis is implemented either by the use of a bank of bandpass filters or by the use of the Fast Fourier Transform (FFT) algorithm. The FFT method is, in general, computationally superior to the bank of filters model. Whichever method is used, the short-time
spectrum has been found to be effective for automatic speaker recognition.

7) Spectral Correlation: It has been reported that the correlation between the short-time spectrum at different frequencies varies from one speaker to another\(^{(28)}\). In fact, statistical analysis of speech spectra reveals that the long term average correlation coefficients of speech spectra exhibit strong intertalker similarities and intratalker differences. Also, the estimates of these coefficients converge to a set of stable values after averaging over long utterances. A duration of 30 sec. of speech at least is required for the stability of these coefficients.

8) Nasal Coarticulation: During connected speech, the shape of the vocal tract at any given time depends not only on the phoneme being spoken at that time but also on the neighboring phonemes. This phenomenon is known as coarticulation and it has been reported that the nature of it in a given context is speaker dependent. The problem of using this kind of information for speaker verification is its difficulty to measure such difference from speech. Coarticulation during the production of nasal consonants has been found useful for speaker verification.
9) Duration of Speech Events: The duration of a word within an utterance differs from one speaker to another. Thus, the relative timing of different speech events are good measures to use to discriminate speakers. Among all the techniques to measure such differences, one of the most useful is time warping. The warping procedure basically aligns the equivalent times of any two contours by a suitable nonlinear deformation of the time axis.
CHAPTER III

ALGORITHM FOR SPEAKER VERIFICATION

3.1 Introduction

Speaker verification algorithms enable one to make a decision to either accept or reject the identity claim of a person on the basis of a prearranged verification phrase. In a legitimate situation, a cooperative individual is willing to indicate his claimed identity so that the machine can judge whether his claim is correct or not. The verification task is a practical one, since it requires only one comparison. Because of its simplicity, it has opened new and revolutionary services in many different fields. Some of the many possible applications of an automatic speaker verification are efficient banking and credit transactions by voice and access to restricted areas by means of "voice-locks". The latter will facilitate the access of authorized personnel through an instantaneous-voice analysis to restricted locations or identity checks for persons seeking classified information over the telephone or similar media.

The task of verification can be summarized as follows:

1) The person making an identity claim is asked to speak a specified utterance.
2) The computer makes measurements on the offered voice sample and compares the results to stored reference data that is available for the claimed identity.

3) Based on the comparison, it accepts or rejects the identity claim.

Figure 3.1 illustrates the basic structure of our verification scheme.

3.2 Parameter Contour

The proposed speaker verification system does not deal with the speech waveform itself, but rather with parameter contours. The parameter contours are functions of time and they describe the speech signal parametrically. The different parameters that are used for speaker verification have been described in the previous chapter. Parameters, such as formants, pitch and intensity have been used to discriminate among speakers. The determination of these parameters, especially the formants, requires extensive computation. It takes approximately 200-300 times the duration of the utterance to compute the formant data. Moreover, Doddington(6) has found that the distances based on formant measurements contributed relatively little to the system performance. Area function, short-time spectrum, spectral correlation,
Fig. 3.1. Block Diagram for Speaker Verification
and the impulse response function were either found to be not very reliable or were difficult to measure.

Recently, the use of predictor coefficients in verification has been proposed \((12,13)\). A success rate higher than 90% has been claimed. Generally, the predictor coefficients were extracted from speech sampled at 8-10 kHz and the order of the linear predictor analysis varied between 8 and 12, the latter being more commonly used.

In this experiment, the linear predictor coefficients, the PARCOR (partial correlation) coefficients, and the autocorrelation coefficients were used as speaker sensitive parameters. The parameters are extracted approximately every 10 m sec. from the speech waveform which is band limited to 2 kHz and sampled at a frequency of slightly above 4 kHz. The duration of each segment was made proportional to the duration of the utterance to provide an approximate alignment of the time scales of the different utterances. 10 m sec. of speech for each frame instead of 20 m sec. for a six-order analysis, was found to be reasonable. The resulting parameter contour was smoothed by a simple linear smoothing process, and the smoothed version constituted the final parameter contour.
In the linear prediction model, the filtering action of the vocal tract, the radiation, and the glottal flow is represented by a discrete linear filter with \( p \) poles. Thus, the output of the all-pole filter \( H(z) \) which is excited by a sequence of pulses separated by the pitch period for voiced sounds, or by white noise for unvoiced sounds, is given by

\[
S_n = - \sum_{k=1}^{p} a_k S_{n-k} + G U_n. \tag{3.1}
\]

In the above formula \( U_n \) is the input excitation, \( G \) is the gain of the filter, and the \( a_k \)'s are the linear prediction coefficients that characterize the filter. The linear prediction model is shown in Fig. 3.2 in the time and frequency domains.

There exist several methods to determine the predictor coefficients \( a_k \). The procedure that has been chosen is the well known least squares approach\(^3\). By this method, we assume that the input \( U_n \) is totally unknown. It can be modelled by an uncorrelated random sequence\(^{40}\). It turns out that a useful prediction technique of \( S_n \) can be obtained from a linearly weighted summation of past samples, say \( \hat{S}_n \), where

---

\(^3\)The method of least squares used here is described in detail in Appendix B.
Fig. 3.2: Discrete Linear Prediction Model of Speech Production
\[ S_n = - \sum_{k=1}^{p} a_k S_{n-k} \]  

By minimizing the mean squared error between \( S_n \) and its estimate \( \hat{S}_n \), we obtain the following set of linear normal equations:

\[ \sum_{k=1}^{p} a_k \sum_{n-k}^{n} S_{n-i} = -\sum_{n}^{n} S_{n} S_{n-i}, \quad 1 \leq i \leq p \]  

(3.3)

or

\[ \frac{1}{p} \sum_{k=1}^{p} \phi_{ik} a_k = -\phi_{i0}, \quad 1 \leq i \leq p \]  

(3.4)

where

\[ \phi_{ik} = \frac{1}{N} \sum_{n}^{n} S_{n-i} S_{n-k} \]  

(3.5)

represents the autocorrelation coefficients.

Solving the above set of equations, one obtains the predictor coefficients \( a_k \). Most of the techniques used to solve the normal equations use recursion relations \((29,30,31,32)\). The PARCOR (partial correlation) coefficients \( b_i \) are computed from the predictor coefficients by a backward recursion. The recursion is as follows:

Initially set: \( a_i^{(p)} = a_i, \quad 1 \leq i \leq p \)

then

1) \( b_i = a_i^{(i)} \)
\[ a_{(i)j}^{(i-1)} = \frac{a_j^i - a_i^i}{1 - b_i^2}, \quad 1 \leq j \leq i-1 \]  

(3.6)

where \( i = p, p-1, \ldots, 1 \)

Figures 3.3 through 3.7 illustrate the four sets of parameters for all the utterances used in our experiment. These parameters are time registered with the reference contour before they are used for verification.

3.3 Time Registration

In this operation, the speech events of the sample contour are brought into the best possible registration with the corresponding speech events of the reference contour. This alignment is accomplished by replott ing the sample contour versus a modified function of time. This step is crucial since, if successful, it solves the problems of the normal and expected variations in the speaking behaviour observed in the repetition of a sample utterance by the same speaker. Doddington developed a procedure by which a sample contour is time registered, by means of a method of steepest ascent, with a stored reference contour. A dynamic programming technique for time warping was used by Itakura.

In the present implementation, a simple linear time shifting was used. The description of this technique follows.
Parameter Contours
Speaker: JN
Utterance: "Cash this bond, please".

Fig. 3.3. Set of Parameter Contours - Sentence 1
Parameter Contours
Speaker: JN
Utterance: "Papa needs two singers"

Fig. 3.4. Set of Parameter Contours - Sentence 2
Parameter Contours
Speaker: JN
Utterance: "Pay the man first, please"

Fig. 3.5. Set of Parameter Contours - Sentence 3
Parameter Contours
Speaker: JN
Utterance: "We were away a year ago"

Fig. 3.6. Set of Parameter Contours - Sentence 4
Parameter Contours

Speaker: JN

Utterance: "My name is Miller, cash this bond, please"

Fig. 3.7. Set of Parameter Contours - Sentence 5
At the beginning of the warping procedure the parameter contours are divided into an equal number of segments. Corresponding segments are then aligned in time and intermediate points are linearly interpolated. The description of the registration procedure follows.

Let $S(t)$ be a function of time which is some parametric representation of the original speech waveform. If $S(t_1)$ represents $S(t)$ during replication one and similarly $r(t)$ represents a simple average of $N$ replications of $S(t)$, then it is desired to maximize, segment wise, the similarity between $S(t_1)$ and $r(t)$. The maximization of the similarity, say in segment $i$, is accomplished by:

1) a proper selection of the function $\lambda(t)$.

$\lambda$ has been chosen to be a shifting function; i.e.,

$$\lambda(t) = t + \Delta t, \quad t_i \leq t \leq t_{i+1}$$

where $\Delta t$ is an integer number. If $\Delta t = 0$, then $\lambda$ represents the identity function, $\lambda = t$. If $\Delta t > 0$, then $\lambda$ represents $\Delta t$ units of shift to the right. If $\Delta t < 0$, then $\lambda$ represents $\Delta t$ units of shift to the left.

2) assigning values to $\Delta t$. In this work, $\Delta t = \{0, \pm 1, \pm 2, \pm 3\}$, and the coefficient
of correlation, \( \rho_{sr,i} \), between the shifted function \( S(\lambda) \) and the function with which it is to be registered, \( r(t) \), is computed. The coefficient of correlation in segment \( i \) is defined as

\[
\rho_{sr,i} = \frac{S(\lambda) r(t)}{[S^2(\lambda) \cdot r^2(t)]^{1/2}}, \quad t_i \leq t \leq t_{i+1}
\]  (3.8)

The overbar means time average computed over the range, \([t_i, t_{i+1}]\), where \( t \) and \( \lambda \) are defined.

3) determining the relative value of \( \Delta t \) for which the coefficient of correlation is the largest.

The shifted \( S(\lambda) \) corresponding to the \( \Delta t \) obtained from step 3) becomes the final registered sample contour. Once the alignment is completed in a segmental manner the intermediate points between the segments are determined by a simple linear interpolation. A process of filling up with zeros, instead of linear interpolation, for points between segments has also been investigated. Results obtained by this method do not show any degradation. However, this method when compared with the method of linear interpolation is less complicated. Empirically determined constraints incorporated in the process prevent either function from being outside the domain in which both are well defined.
Figure 3.8 and Fig. 3.9 illustrate the relationship between a sample and reference contour for the two different speakers before and after time warping. In both cases the parameter contour is the PARCOR or reflection coefficient $b_6$.

3.4 Development of the Reference Contour

Among all the specimen contours available for a given speaker, a subset of them, called the training set, is used to characterize the speaker. In a verification system, the specimens of the training set would be collected when the speaker's identity was not in question. The specimens, five for this work, are analyzed and the data is stored in disk files. This data is used to construct the speaker's initial reference file. The reference file serves as the standard against which future test samples are compared.

The construction of the reference file is carried out off-line and updated with the analysis data of recorded specimens from time to time.

The reference file is constructed from the design set by the following algorithm.

a) Stretch or contract all sample contours to a standard length of 190 pts. The standardization of length is an inherent
Reference: JN
Test Contour: JN
Parameter: $b_6$
Utterance: "My name is Miller, cash this bond, please"

Fig. 3.8. Time Registration - Speaker JN
Reference: FE
Test Contour: JN
Parameter: $b_6$
Utterance: "My name is Miller, cash this bond, please"

Fig. 3.9. Time Registration - Speaker FE
property of this verification procedure since all utterances are analyzed to yield the same number of parameters with the result that the contours are equal in length.

b) Smooth all the sample contours.

c) Average all sample contours at each sample point, producing the first "trial reference".

d) Register each sample contour to the trial reference by means of a linear time shifting procedure.

e) Average the registered sample contours to form a new trial reference.

f) Determine whether or not the new trial reference is significantly different from the previous one. The criterion used to determine whether a significant difference exists is to test if at least 2% of the values differ by at least 2%.

g) If it is not significantly different, stop. Otherwise, repeat from step (d) onwards.

An alternative procedure to step (c) has also been investigated. In this procedure, instead of taking the
average of all sample contours to form the first trial prototype, one of the sample contours is used as the first trial prototype. Results obtained using the alternative procedure, although reducing computations, did not show any significant improvement.

The final reference file obtained at step (g) is stored, as reference information for a particular speaker, for subsequent use in a verification system. Reference files are constructed for all the parameters, namely, linear prediction, parcor and autocorrelation coefficients. Figures 3.10 through Fig. 3.13 illustrate sample contours of the design set and the first trial reference for all the parameters for the speaker JN. Figure 3.14 through Fig. 3.17 illustrate the final warped version for the same set of parameters and the same speaker. This final reference file is kept in disk storage for later use in the task of speaker verification.

3.5 Computational Distances

Following the automatic time registration of the sample contour to the reference contour, a set of measurements is applied to the sample contour and compared with the same set of measurements applied to the reference contour. Distances are calculated which quantify the closeness or dissimilarity between the sample and
Speaker: JN
Parameter: $b_6$
Utterance: "My name is Miller, cash this bond, please"

Fig. 3.10. Design Set and Trial Reference - Parameter $b_6$
Speaker: JN
Parameter: $b_3$
Utterance: "My name is Miller, cash this bond, please"

Fig. 3.11: Design Set and Trial Reference - Parameter $b_3$
Fig. 3.12. Design Set and Trial Reference - Parameter $a_3$

Speaker: JN
Parameter: $a_3$
Utterance: "My name is Miller, cash this bond, please"
Spec. 1
Spec. 2
Spec. 3
Spec. 4
Spec. 5
Trial Reference

Speaker: JN
Parameter: $\phi$
Utterance: "My name is Miller, cash this bond, please"

Fig. 3.13. Design Set and Trial Reference - Parameter $\phi$
Speaker: JN

Parameter: $b_6$

Utterance: "My name is Miller, cash this bond, please"

Fig. 3.14. Registered Design Set and Final Reference - Parameter $b_6$
Speaker: JN
Parameter: $b_3$
Utterance: "My name is Miller, cash this bond, please"

Fig. 3.15. Registered Design Set and Final Reference - Parameter $b_3$
Speaker: JN
Parameter: $a_3$
Utterance: "My name is Miller, cash this bond, please"

Fig. 3.16. Registered Design Set and Final Reference - Parameter $a_3$
Speaker: JN
Parameter: $\phi_7$
Utterance: "My name is Miller, cash this bond, please"

Fig. 3.17. Registered Design Set and Final Reference – Parameter $\phi_7$
reference contour measurements. These distances are such that a small value results if a sample contour is identical to the reference contour to which it is compared, and a large value results if the sample and reference contours differ. The greater the differences, the larger the distances.

Finally, the distances are combined into an overall distance which is compared with a given threshold distance value to determine whether the identity claim should be accepted or rejected.

There are two kinds of distances—those based on short segmental correlation and those computed over the whole contour.

For the segmental distance, the contours are divided into 19 equal length segments (10 pts per segment), and then the following formula is applied to each segment

\[ d_1 = \frac{1}{19} \sum_{i=1}^{19} \frac{(1-\rho_{yr,i})^2}{\sigma_{p,i}^2} \tag{3.9} \]

where \( \rho_{yr,i} \) is the coefficient of correlation between the unknown sample contour \( y \) and the reference contour \( r \) in segment \( i \).

The quantity in the denominator is the measure of variability which represents the average, over the set of registered sample contours, of the squared difference
indicated in the numerator, that is

$$\sigma_{p,i}^2 = \frac{1}{n_s-1} \sum_{s=1}^{n_s} (1-\rho_{SR,i})^2 \quad 1 \leq i \leq 19$$  \hspace{1cm} (3.10)

where $s$ represents a final warped sample contour from the design set and $n_s$ is the number of sample contours. In our study, $n_s = 5$.

The second distance used in our verification system is defined over the whole contour instead of segment-by-segment. The formula for the second distance is

$$d_2 = \frac{(1-\rho_{yx})^2}{\sigma_{\rho}^2}$$  \hspace{1cm} (3.11)

where $\rho_{yx}$ again is the correlation coefficient between the unknown sample and the reference contours computed over the whole length of the contours. The quantity in the denominator $\sigma_{\rho}^2$ is the average over the sample set of $(1-\rho_{SR})$.

The third distance also is defined over the whole contour. It is a simple percentage error defined as

$$d_3 = \frac{\sum (x_j - y_j)^2}{\sum x_j^2} \times 100$$  \hspace{1cm} (3.12)

The overall distance $D_i$ is computed for both the design and test sets as follows

$$D_i = d_{1i}^2 + d_{2i}^2$$  \hspace{1cm} (3.13)
where \( i \) represents the number of sample contours in the design and test sets. The percentage error measurement and the correlation values, obtained at the time of registration, between the sample and reference contours are used as 'extra' distances for verification cases that are doubtful. In addition to the different distances discussed above, the following new distances were investigated.

1) Exponential Distance: This measure was evaluated segment-by-segment, and also over the whole contour.

The segmental exponential distance is defined by

\[
Ed_1 = \frac{1}{NSEG} \sum_{i=1}^{NSEG} \exp \left( \frac{1}{\rho_{yr,i}} \right)
\]  

(3.14)

where \( \rho_{yr,i} \) is the coefficient of correlation between the unknown sample contour \( y \) and the reference contour \( r \) in segment \( i \). \( NSEG \) represents the number of segments. In our case, \( NSEG \) is equal to 19.

The second exponential distance is defined over the whole contour and is given by

\[
Ed_2 = \exp \left( \frac{1}{\rho_{yr}} \right)
\]  

(3.15)

where \( \rho_{yr} \) is the correlation coefficient between the unknown sample and reference contours computed over the whole length of the parameter contour.
Both exponential measures applied to each of the four parameters used in our experimentation yielded results that were not much different from $d_1$ and $d_2$.

2) Distance Measurements of Lummis: The segmental distances given by R. C. Lummis\(^{(10)}\) were implemented. These measures use three coefficients obtained from an orthogonal polynomial representation of the parameter contour samples comprising a particular time segment. Furthermore, in our experiment, these three coefficients were obtained from a Fourier expansion instead of any other orthogonal polynomial representation. Results obtained using these segmental distances did not improve our verification accuracy.

3) Number of Points per Segment: The number of points assigned to each segment was also investigated. Ten points per segment were used in our studies. Segmental measures computed with higher number of points per segment did not yield better verification results. Consequently, the number of points in each segment was fixed to ten points for subsequent experimentation.

In conclusion, the segmental distance $d_1$ with ten points per segment, the distance $d_2$, and distance $d_3$, both computed over the whole range of the parameter contour, were used to implement our automatic verification system.
3.6 **Threshold Computation**

A threshold distance is computed for each utterance and each speaker. The threshold is based on the distances obtained when the four test contours of the test set of each speaker is compared against his own reference. It is defined by

\[ D_T = \mu + 2 \times SD \] (3.16)

where \( \mu \) is the mean and SD the standard deviation of the individual overall distances.

The sample contours of the test set normally yield greater distances than the sample contours of the design set used to construct the reference file. Thus, the estimation of the threshold is based on distances obtained from the test set rather than the design set.

The quantity of \( 2 \times SD \) that is added to the mean is somewhat arbitrary and was arrived at by inspection of the overall distances obtained from comparing the sample and reference files. Tables 3.1 and 3.2 show a typical set of distances and a threshold distance for two speakers for the utterance of "my name is Miller, cash this bond, please". The distances were obtained comparing the sample contours of the speakers JN and FE against their own reference contours. Note that distances obtained for the
## TABLE 3.1

DISTANCE MEASUREMENTS

Reference: JN
Test Contour: JN
Parameter: $b_6$
Threshold: 43.545
Utterance: "My name is Miller, cash this bond, please"

<table>
<thead>
<tr>
<th>Utterance</th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
<th>$D$</th>
<th>Cor.</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTT1</td>
<td>.6102</td>
<td>.7439</td>
<td>13.26</td>
<td>.9257</td>
<td>.93385</td>
</tr>
<tr>
<td>UTT2</td>
<td>2.287</td>
<td>2.604</td>
<td>23.98</td>
<td>12.01</td>
<td>.87625</td>
</tr>
<tr>
<td>UTT3</td>
<td>.6592</td>
<td>1.605</td>
<td>25.24</td>
<td>3.009</td>
<td>.90284</td>
</tr>
<tr>
<td>UTT4</td>
<td>2.615</td>
<td>2.540</td>
<td>23.37</td>
<td>13.29</td>
<td>.8778</td>
</tr>
<tr>
<td>UTT5</td>
<td>.7461</td>
<td>.7194</td>
<td>15.50</td>
<td>1.074</td>
<td>.93405</td>
</tr>
<tr>
<td>UTT6</td>
<td>2.152</td>
<td>5.706</td>
<td>35.14</td>
<td>37.18</td>
<td>.81681</td>
</tr>
<tr>
<td>UTT7</td>
<td>.5918</td>
<td>.7654</td>
<td>13.72</td>
<td>.9361</td>
<td>.93290</td>
</tr>
<tr>
<td>UTT8</td>
<td>1.937</td>
<td>2.660</td>
<td>26.44</td>
<td>10.82</td>
<td>.87493</td>
</tr>
<tr>
<td>UTT9</td>
<td>.6990</td>
<td>1.166</td>
<td>20.10</td>
<td>1.848</td>
<td>.91719</td>
</tr>
</tbody>
</table>

Reference: JN
Test Contour: PE
Parameter: $b_6$

<table>
<thead>
<tr>
<th>Utterance</th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
<th>$D$</th>
<th>Cor.</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTT1</td>
<td>21.00</td>
<td>17.53</td>
<td>70.64</td>
<td>748.7</td>
<td>.67886</td>
</tr>
<tr>
<td>UTT2</td>
<td>35.50</td>
<td>21.19</td>
<td>68.48</td>
<td>1709</td>
<td>.64695</td>
</tr>
<tr>
<td>UTT3</td>
<td>10.95</td>
<td>13.41</td>
<td>61.41</td>
<td>299.5</td>
<td>.71918</td>
</tr>
<tr>
<td>UTT4</td>
<td>15.12</td>
<td>25.18</td>
<td>91.38</td>
<td>862.8</td>
<td>.61513</td>
</tr>
<tr>
<td>UTT5</td>
<td>13.24</td>
<td>15.53</td>
<td>71.25</td>
<td>416.5</td>
<td>.69976</td>
</tr>
<tr>
<td>UTT6</td>
<td>11.37</td>
<td>18.56</td>
<td>67.41</td>
<td>473.9</td>
<td>.66967</td>
</tr>
<tr>
<td>UTT7</td>
<td>6.216</td>
<td>8.731</td>
<td>54.26</td>
<td>114.9</td>
<td>.77339</td>
</tr>
<tr>
<td>UTT8</td>
<td>17.66</td>
<td>20.95</td>
<td>79.99</td>
<td>750.5</td>
<td>.64900</td>
</tr>
<tr>
<td>UTT9</td>
<td>8.264</td>
<td>15.93</td>
<td>69.48</td>
<td>322</td>
<td>.69394</td>
</tr>
</tbody>
</table>
### TABLE 3.2

**DISTANCE MEASUREMENTS**

Reference: FE  
Test Contour: FE  
Parameter: $a_3$  
Threshold: 34.899  
Utterance: "My name is Miller, cash this bond, please"

<table>
<thead>
<tr>
<th>Utterance</th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
<th>D</th>
<th>Cor.</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTT1</td>
<td>.7177</td>
<td>1.503</td>
<td>17.75</td>
<td>2.773</td>
<td>.92392</td>
</tr>
<tr>
<td>UTT2</td>
<td>.8598</td>
<td>1.296</td>
<td>17.02</td>
<td>2.418</td>
<td>.92935</td>
</tr>
<tr>
<td>UTT3</td>
<td>.3460</td>
<td>.5228</td>
<td>11.86</td>
<td>.3930</td>
<td>.95512</td>
</tr>
<tr>
<td>UTT4</td>
<td>3.796</td>
<td>3.755</td>
<td>31.90</td>
<td>28.51</td>
<td>.87973</td>
</tr>
<tr>
<td>UTT5</td>
<td>.8642</td>
<td>.8235</td>
<td>10.95</td>
<td>1.421</td>
<td>.94367</td>
</tr>
<tr>
<td>UTT6</td>
<td>.9386</td>
<td>1.459</td>
<td>15.39</td>
<td>3.010</td>
<td>.92503</td>
</tr>
<tr>
<td>UTT7</td>
<td>.7635</td>
<td>.7902</td>
<td>11.64</td>
<td>1.207</td>
<td>.94483</td>
</tr>
<tr>
<td>UTT8</td>
<td>.9978</td>
<td>1.638</td>
<td>19.95</td>
<td>3.679</td>
<td>.92112</td>
</tr>
<tr>
<td>UTT9</td>
<td>.6128</td>
<td>1.361</td>
<td>17.04</td>
<td>2.228</td>
<td>.92759</td>
</tr>
</tbody>
</table>

Reference: FE  
Test Contour: JN  
Parameter: $a_3$

<table>
<thead>
<tr>
<th>Utterance</th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
<th>D</th>
<th>Cor.</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTT1</td>
<td>12.76</td>
<td>6.270</td>
<td>28.80</td>
<td>202.1</td>
<td>.84458</td>
</tr>
<tr>
<td>UTT2</td>
<td>43.27</td>
<td>19.80</td>
<td>52.61</td>
<td>2264</td>
<td>.72381</td>
</tr>
<tr>
<td>UTT3</td>
<td>61.16</td>
<td>22.90</td>
<td>59.96</td>
<td>4265</td>
<td>.70296</td>
</tr>
<tr>
<td>UTT4</td>
<td>22.50</td>
<td>24.50</td>
<td>60.19</td>
<td>1106</td>
<td>.69280</td>
</tr>
<tr>
<td>UTT5</td>
<td>9.564</td>
<td>20.10</td>
<td>58.17</td>
<td>495.3</td>
<td>.72177</td>
</tr>
<tr>
<td>UTT6</td>
<td>8.144</td>
<td>16.36</td>
<td>50.29</td>
<td>333.9</td>
<td>.74897</td>
</tr>
<tr>
<td>UTT7</td>
<td>14.80</td>
<td>22.13</td>
<td>64.63</td>
<td>708.8</td>
<td>.70803</td>
</tr>
<tr>
<td>UTT8</td>
<td>7.452</td>
<td>8.796</td>
<td>34.58</td>
<td>132.9</td>
<td>.81592</td>
</tr>
<tr>
<td>UTT9</td>
<td>19.60</td>
<td>27.21</td>
<td>92.25</td>
<td>1125</td>
<td>.67621</td>
</tr>
</tbody>
</table>
test set are larger than the ones obtained for the design set.
CHAPTER IV

VERIFICATION USING COMPOSITE REFERENCE

4.1 Introduction

This is a new concept in automatic verification systems. In the existing speaker verification systems, the main task is the authentication of an individual's claimed identity by analysis of his spoken utterances. The acoustic analysis of the individual's fixed, sentence-long utterances result in a function of time or contour for each feature analyzed. Then the system compares the set of sample contours obtained from the unknown individual with the set of reference contours corresponding to the identity claimed by that individual. If the comparison results in an overall measure of dissimilarity which is smaller than a predetermined threshold, the identity claim is accepted. Otherwise, it is rejected.

In the proposed new verification system using composite reference contours, the combined information of the identity of several speakers is given. That is, it is assumed the existence of an overall reference contour which contains the information corresponding to all the identities of all the speakers in a given population of speakers. The verification task consists in authenti-
cating an individual's claimed identity, as in the above mentioned procedures, by the analysis of his speech. The individual's claimed identity need not be in the composite reference contour.

4.2 Construction of Composite Reference Contour

The construction of the composite reference file is done off-line, and it is updated from time to time using new set of analyzed data.

The composite reference file is constructed, using parameter contours of the design set of each speaker, by the following algorithm.

1) Stretch or contract all smoothed sample contours of all speakers for whom the composite reference contour is being constructed.

2) Average all sample contours at each sample point, producing the first "composite trial reference".

3) Register each sample contour of each speaker to the composite trial reference by means of the linear time shifting registration procedure.

4) Average the registered sample contours of each speaker to form a new composite trial reference.
5) Determine whether or not the new composite trial reference is significantly different from the previous one. The criterion of significantly different is that at least 2% of the values differ by at least 2%.

6) If it is not significantly different, stop. Otherwise, repeat from step (3) above.

The final composite reference file obtained at the end of step (6) is kept on disk as final reference information for all the speakers whose sample specimens enter in the construction of the composite reference. This information is used in subsequent verification tasks.

Figure 4.1 and Fig. 4.1A show original parameter contours representing the LPC coefficient $a_4$ for the speakers JN, FE and MS, and its resulting composite trial reference. The fixed utterance, "cash this bond, please", was sampled at 8 kHz, and an eighth-order linear prediction analysis was implemented. Figure 4.2 and Fig. 4.2A illustrate the final warped version of the sample contours and the final composite reference contour for the same speakers.

4.3 **Algorithm for Speaker Verification Using Composite Reference**

Once the composite reference contour is constructed, a subtracted reference contour for each of the speakers,
Speakers: JN, FE, MS
Parameter: a
Utterance: "Cash this bond, please"

Fig. 4.1. Design Set and Trial Composite Reference Contours
Speakers: JN, PE, MS
Parameter: $a_4$
Utterance: "Cash this bond, please"

Fig. 4.1A. (Cont.)
Speakers: JN, FE, MS
Parameter: $a_4$
Utterance: "Cash this bond, please"

Fig. 4.2. Registered Design Set and Final Composite Reference Contours
Speakers: JN, FE, MS
Parameter: $a_4$
Utterance: "Cash this bond, please"

Fig. 4.2A (Cont.)
whose specimens entered in the construction of the composite reference, is obtained in the following fashion. Suppose speaker Y is part of the composite reference. The subtracted reference contour of speaker Y, say DYR, is formed by the following steps:

1) Subtract each of the parameter contours of the design set of speaker Y from the composite reference contour.

2) Using the subtracted sample contours of speaker Y, construct his subtracted reference contour DYR, by the procedure described in Chapter III, Section 3.4.

Figure 4.3 shows four subtracted sample contours of the design set of speaker JN and his subtracted trial reference. The parameter contour represents the partial correlation coefficient $b_3$ and the utterance is "my name is Miller, cash this bond, please". Figure 4.4 illustrates the final warped subtracted parameter contours and the final subtracted reference contour for the same speaker and the same utterance.

If the speaker Y is not part of the composite reference contour, then his subtracted reference contour is obtained by the same procedure described above. The composite reference contour along with the subtracted
Subtracted Spec. 1
Subtracted Spec. 2
Subtracted Spec. 3
Subtracted Spec. 4
Trial Reference

Original version
Speaker: 'JN'
Parameter: $b_3$
Utterance: "My name is Miller, cash this bond, please"

Fig. 4.3. Design Set and Trial Subtracted Reference Contours
Warped Version
Speaker: JN
Parameter: b_3
Utterance: "My name is Miller, cash this bond, please"

Fig. 4.4. Registered Design Set and Final Subtracted Reference Contours
reference contour of selected speakers is constructed off-line, and used for verification purposes. In a practical implementation of a verification system, suppose speaker X enters his identity claim Y and speaks his prearranged verification sentence-long utterance. After a linear prediction analysis is performed on the offered voice sample, a parameter contour is constructed for the speaker X. Then, the task of verification is carried out by the following algorithm.

1) Subtract the parameter contour of speaker X, sample by sample, from the composite reference contour. The result is a new subtracted parameter contour, call it DX.

2) Register the subtracted parameter contour DX of speaker X against the subtracted reference contour of speaker Y, namely, DYR. Time registration is accomplished by the linear shifting registration procedure discussed in Chapter III, Section 3.3.

3) Compute distances between the parameter contours DX and DYR.

4) Compute threshold and decide whether claim is valid or not.
Figure 4.5 illustrates the composite reference contour for a population of four speakers: JN, FE, MS, and WCM. Also, the test parameter contour of speaker FE and its subtracted parameter contour are shown. The utterance is "My name is Miller, cash this bond, please" and the parameter is the partial correlation coefficient $b_3$. Figure 4.6 illustrates the composite reference contour for a population of four speakers: JN, FE, MS, and PHA. The test parameter contour is the same as that of the previous figure. The utterance is the all-voiced utterance "we were away a year ago".

4.4 Alternative Procedure for Verification Using Composite Reference

In an automatic speaker verification procedure using composite reference contour, the principal idea is that the composite reference file theoretically represents the combined information of the identity of all speakers.

Consider a population of $N$ speakers for whom a composite reference file has been constructed. If an unknown speaker $X$ claims the identity of speaker $Y$ whose identity information is merged in the composite reference file, then the problem is, how to extract the relevant information of the identity of speaker $Y$ from the composite reference file. In the suggested previous algorithm,
Composite Reference

Speakers: JN, FE, MS, WCM

Test Contour: FE

Utterance: "My name is Miller, cash this bond, please"

Fig. 4.5. Composite Reference, Original and Subtracted Contours - Sentence 5
Composite Reference

Speakers: JN, FE, MS, PHA

Test Contour: FE

Utterance: "We were away a year ago"

Fig. 4.6. Composite Reference, Original and Subtracted Contours - Sentence 4
the problem was solved by a direct subtraction of the sample contours of speaker \( Y \) from the composite reference. However, it is also realistic to be able to solve the problem by removing, from the composite reference file, all the relevant identity information of all the speakers except \( Y \). That is, if we denote the relevant identity information of speaker \( Y \) by \( E_Y \), then

\[
E_Y = CR - \sum_{i=1}^{N-1} X_i
\]

where \( CR \) is the composite reference file and \( X_i \) represents the relevant identity information of speaker \( i \).

As of this writing, preliminary investigations using this alternative procedure did not yield results that were significantly better than the previous method.

Table 4.1 and Table 4.2 show a typical set of distances and a threshold distance using composite reference for sentences 4 and 5, respectively. In the foregoing discussion, namely section 4.3, the operation of direct subtraction was used to construct the subtracted reference contour of the speaker whose identity was claimed. The reason being that any other operation such as addition had the effect of cancelling out most of the important speech events within the parameter contour.
### TABLE 4.1
**DISTANCE MEASUREMENTS USING COMPOSITE REFERENCE**

<table>
<thead>
<tr>
<th>Utterance</th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
<th>$D$</th>
<th>Cor.</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTT1</td>
<td>.8294</td>
<td>.9934</td>
<td>31.21</td>
<td>1.675</td>
<td>.87704</td>
</tr>
<tr>
<td>UTT2</td>
<td>2.383</td>
<td>1.384</td>
<td>30.06</td>
<td>7.595</td>
<td>.85485</td>
</tr>
<tr>
<td>UTT3</td>
<td>.1979</td>
<td>.2587</td>
<td>22.91</td>
<td>.1061</td>
<td>.93724</td>
</tr>
<tr>
<td>UTT4</td>
<td>4.838</td>
<td>2.125</td>
<td>41.42</td>
<td>27.92</td>
<td>.82014</td>
</tr>
<tr>
<td>UTT5</td>
<td>.9892</td>
<td>2.269</td>
<td>38.05</td>
<td>6.129</td>
<td>.81415</td>
</tr>
<tr>
<td>UTT6</td>
<td>2.470</td>
<td>1.013</td>
<td>34.87</td>
<td>7.127</td>
<td>.87581</td>
</tr>
<tr>
<td>UTT7</td>
<td>.6224</td>
<td>.4100</td>
<td>16.97</td>
<td>.5555</td>
<td>.92101</td>
</tr>
<tr>
<td>UTT8</td>
<td>1.220</td>
<td>1.124</td>
<td>31.76</td>
<td>.2753</td>
<td>.86920</td>
</tr>
<tr>
<td>UTT9</td>
<td>4.643</td>
<td>1.947</td>
<td>40.38</td>
<td>25.34</td>
<td>.82787</td>
</tr>
</tbody>
</table>

**Reference:** JN  
**Test Contour:** JN  
**Parameter:** $b_3$  
**Threshold:** 33.870  
**Utterance:** "We were away a year ago"  

<table>
<thead>
<tr>
<th>Utterance</th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
<th>$D$</th>
<th>Cor.</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTT1</td>
<td>400.1</td>
<td>52.67</td>
<td>184.1</td>
<td>162800</td>
<td>.10464</td>
</tr>
<tr>
<td>UTT2</td>
<td>1385</td>
<td>116.5</td>
<td>319.3</td>
<td>193200</td>
<td>-.33159</td>
</tr>
<tr>
<td>UTT3</td>
<td>817.4</td>
<td>64.64</td>
<td>171.7</td>
<td>672200</td>
<td>.00814</td>
</tr>
<tr>
<td>UTT4</td>
<td>256.6</td>
<td>32.96</td>
<td>116.6</td>
<td>66930</td>
<td>.29169</td>
</tr>
<tr>
<td>UTT5</td>
<td>1163</td>
<td>105</td>
<td>300.1</td>
<td>136300</td>
<td>-.26426</td>
</tr>
<tr>
<td>UTT6</td>
<td>806.4</td>
<td>56.34</td>
<td>171.2</td>
<td>653500</td>
<td>.07398</td>
</tr>
<tr>
<td>UTT7</td>
<td>1326</td>
<td>97.67</td>
<td>248.8</td>
<td>1767000</td>
<td>-.21924</td>
</tr>
<tr>
<td>UTT8</td>
<td>82.70</td>
<td>9.492</td>
<td>69.22</td>
<td>6929</td>
<td>.61990</td>
</tr>
<tr>
<td>UTT9</td>
<td>1144</td>
<td>80.69</td>
<td>206.1</td>
<td>1315000</td>
<td>-.10823</td>
</tr>
</tbody>
</table>
TABLE 4.2
DISTANCE MEASUREMENTS USING COMPOSITE REFERENCE

| Reference: JN | Test Contour: JN | Parameter: $b_3$ | Threshold: 8.91 | Utterance: "My name is Miller, cash this bond, please"
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Utterance</strong></td>
<td>$d_1$</td>
<td>$d_2$</td>
<td>$d_3$</td>
<td>$D$</td>
</tr>
<tr>
<td>UT1</td>
<td>.8739</td>
<td>1.135</td>
<td>40.59</td>
<td>2.052</td>
</tr>
<tr>
<td>UT2</td>
<td>1.576</td>
<td>1.816</td>
<td>47.10</td>
<td>5.783</td>
</tr>
<tr>
<td>UT3</td>
<td>.5966</td>
<td>1.115</td>
<td>41.86</td>
<td>1.600</td>
</tr>
<tr>
<td>UT4</td>
<td>1.316</td>
<td>1.903</td>
<td>47.38</td>
<td>5.356</td>
</tr>
<tr>
<td>UT5</td>
<td>.2538</td>
<td>.5099</td>
<td>23.39</td>
<td>.3244</td>
</tr>
<tr>
<td>UT6</td>
<td>1.525</td>
<td>2.290</td>
<td>65.64</td>
<td>7.572</td>
</tr>
<tr>
<td>UT7</td>
<td>.8508</td>
<td>1.240</td>
<td>38.66</td>
<td>2.261</td>
</tr>
<tr>
<td>UT8</td>
<td>1.725</td>
<td>2.145</td>
<td>42.10</td>
<td>7.577</td>
</tr>
<tr>
<td>UT9</td>
<td>1.511</td>
<td>2.199</td>
<td>64.95</td>
<td>7.118</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference: JN</th>
<th>Test Contour: FE</th>
<th>Parameter: $b_3$</th>
<th>Utterance:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Utterance</strong></td>
<td>$d_1$</td>
<td>$d_2$</td>
<td>$d_3$</td>
</tr>
<tr>
<td>UT1</td>
<td>18.90</td>
<td>21.92</td>
<td>230.9</td>
</tr>
<tr>
<td>UT2</td>
<td>21.87</td>
<td>30.10</td>
<td>275.8</td>
</tr>
<tr>
<td>UT3</td>
<td>53.22</td>
<td>39.65</td>
<td>231.9</td>
</tr>
<tr>
<td>UT4</td>
<td>16.83</td>
<td>25.61</td>
<td>244.0</td>
</tr>
<tr>
<td>UT5</td>
<td>23.65</td>
<td>34.96</td>
<td>253.7</td>
</tr>
<tr>
<td>UT6</td>
<td>39.13</td>
<td>46.93</td>
<td>319.7</td>
</tr>
<tr>
<td>UT7</td>
<td>50.55</td>
<td>54.95</td>
<td>293.1</td>
</tr>
<tr>
<td>UT8</td>
<td>29.76</td>
<td>34.53</td>
<td>291.4</td>
</tr>
<tr>
<td>UT9</td>
<td>41.18</td>
<td>47.35</td>
<td>257.0</td>
</tr>
</tbody>
</table>
CHAPTER V
EXPERIMENTAL RESULTS

5.1 Data Collection

The utterances chosen for the experiment were:
1) Cash this bond, please.
2) Papa needs two singers.
3) Pay the man first, please.
4) We were away a year ago.
5) My name is Miller, cash this bond, please.

The speech data consisted of 54 utterances, consisting of nine repetitions of the same sentence, spoken by six speakers. All the speakers were male, ranging in age from 23 to 38. They were all native Canadian speakers of English with no overt speech defects. None of the speakers attempted to imitate anyone. They spoke in their own natural voice and they were not trained in any manner. All the recordings were made in a studio environment.

The recorded speech signal was low-pass filtered at 2 kHz and sampled at a rate of slightly above 4 kHz.

---

4 Instructions given to all the speakers can be found in Appendix E.
by means of a 14-bit analog-to-digital converter, and stored in disk for further processing. Also the procedure of down sampling from 8 kHz to 4 kHz by extracting every other sample of the original sequence gave good results. In our experiments both procedures were used.

The stored samples were analyzed utterance by utterance. Each utterance, for example, was divided into 190 equally long segments. The duration of each segment was made proportional to the duration of the utterance to provide an approximate alignment of the time scales of the different utterances. Thus, the resulting contours were equal in length. The analysis interval consisted of approximately 40 speech samples corresponding to an interval of 10 m sec. However, the actual analysis interval varied with the duration of each utterance.

The silent portions and pauses were removed from the analysis interval. These portions of the signal were determined automatically by computing the energy of the utterance segment by segment. The resulting segmental energy, within 10 m sec. duration, was compared against a fixed threshold. If the segment had energy less than the threshold, it was classified as either a silent portion or a pause and was eliminated from the analysis interval.
Six predictor and parcor coefficients and seven autocorrelation coefficients were computed for each of the 190 uniformly spaced time frames in an utterance. They were determined using the procedure outlined in Chapter III, Section 3.2. The analysis was carried out for all the sentences and for all of the 54 utterances of the six speakers. The resulting nine sample contours for each speaker were partitioned into design and test sets. The sample contours in the design set were used to construct the reference (or prototype) contour; the sample contours in the test set were used to test the effectiveness of the automatic speaker-verification procedure. Each of the nine sample contours obtained from the remaining speakers was used in turn as the test set to provide a total of 54 judgments for verification.

5.2 Experimental Set-up

A block diagram of the experimental set-up is shown in Fig. 5.1.

The basic units are the signal generator and the signal processor.

There are six pieces of hardware that form the main elements of the signal processor. The Data General NOVA 840 computer with a core memory storage capacity
Fig. 5.1. Block Diagram of Experimental Set-Up
of 32 K words is used. Two Diablo disks with a combined storage capability of 2.494 million words are also an essential element of the processor. In it reside all the automatic speaker-verification programs.

A Tustin Electronics Corporation analog-to-digital converter with 14-bit resolution and up to a 100 kHz sampling rate is used to digitize the speech signal. An INFOTON VISTAR CRT terminal is used to key-in the instructions for processing in the computer. In addition, a Textronix-4013 Graphics terminal and a Textronix-4610 Hard Copy Unit are used to display and copy waveforms and functions generated during the processing.

The main components of the signal generator consist of tapes recorded in a studio using high-quality equipment. They are played back on a Crown tape recorder, Model SX800 with a remote control option. The filtering requirements are achieved using a Krohn-Hite variable filter, Model 3750.°

5.3 Verification Results Using LPC Parameters

A. E. Rosenberg(12) reported that an eighth-order linear prediction model was sufficient for speaker

---

5 A detailed description of all the elements is given in Appendix D.
verification and that there exists redundancy between the linear prediction coefficients. The redundancy in the LPC parameters is not surprising in view of the fact that the speech signal can be synthesized with only three formant parameters, namely $F_1$, $F_2$ and $F_3$. A description of this redundancy is given by Sambur. His results show that there is an extremely high correlation between coefficients and that widely spaced predictor coefficients are reasonably uncorrelated with one another. An eigenvalue analysis of the covariance matrix of the predictor coefficients was used by Sambur to further demonstrate the redundant nature of the coefficients by showing that the 12-predictor coefficients can be adequately represented in a space less than 12 dimensions.

The analysis carried out by Rosenberg, based on the high correlation between adjacent predictor coefficients, suggests that all the contours are not needed to obtain good speaker verification results, and that the best strategy for choosing the coefficient contours is to select contours that are widely spaced in numbered ordering such as $a_2$, $a_4$, $a_8$. The experimental results confirm these observations. The error rate for all 8 contours was not appreciably better than the error rates for the best set of three contours. For the final selection of features, the number of predictor coeffi-
cient contours was further reduced from 3 to 2, $a_4$ and $a_8$ yielding a considerable saving in storage requirements with no significant degradation in performance. However, in the tests conducted during our experiments, the eighth-order or any higher order model did not yield satisfactory results when applied to speech sampled at 4 kHz. Tests were conducted to determine the optimum order of the linear model, and it was found that a sixth-order model yielded the best results. Also, with the sixth-order model it was found that only two of the six parameters, viz $a_3$ and $a_6$, yielded reliable results. Hence, these two parameters were selected for speaker verification.

By a similar study, it was also determined that of the six PARCOR (reflection) coefficients, only $b_3$ and $b_6$ were needed. Generally, the autocorrelation coefficients were not very satisfactory, but the coefficient $\phi_7$ yielded satisfactory results. Thus, the parameter set chosen for this study included $a_3$, $b_6$, $b_3$ and $\phi_7$ (33).

Seven sentences were evaluated for these set of parameters. Results are shown for each sentence separately and for two sentences combined. The first five sentences were:
Sentence 1: Cash this bond, please.
Sentence 2: Papa needs two singers.
Sentence 3: Pay the man first, please.
Sentence 4: We were away a year ago.
Sentence 5: My name is Miller, cash this bond, please.
Also, the following two sentences were artificially created by attaching two of the previous sentences.
They were:
Sentence 6: Cash this bond, please--Papa needs two singers.
Sentence 7: Cash this bond, please--Pay the man first, please.
The first three sentences are about 1.5 seconds in duration and the remaining two are longer. Sentence 4 is about two seconds in duration and sentence 5 is three seconds long. All of the sentences except sentence 4 are mixtures of voiced and nasal sounds. Sentence 4 is all voiced. These sentences, except sentences 5, 6 and 7, were used by other researchers in previous implementations of a speaker verification system.
The results presented here are in terms of the two types of error that can occur: rejecting a legitimate speaker claim and accepting an impostor claim.
The speaker-rejection rate was calculated simply from a tabulation of the number of rejections experienced by all the speakers in our fixed population during the evaluation process. Excluded from these tabulations were trials in which the speaker deliberately altered his utterance.

Table 5.1 summarizes the results of the evaluations. The results are presented for two parcor coefficients, one linear predictor coefficient and one autocorrelation coefficient. Also, results are shown for a set of selected parameters, namely $b_6$ and $\phi_7$.

The error rate obtained with all the measurements for the first three sentences varied from 6 to 7%. The very same measurements applied to $b_6$ and $\phi_7$ yielded an error rate between 3 to 3.6%. The error rates decreased as the length of the utterance increased. The all-voiced sentence 4 yielded an error rate of 4% using all the parameters and yielded 2% using a selected number of parameters. Further, for a sentence of three seconds in duration, namely sentence 5, the overall error rate dropped to 1% using all the parameters and to .5% using $b_6$ and $\phi_7$. In addition, low error rates were also obtained when two short sentences were attached together artificially.
### Table 5.1

**Verification Error Rates (%)**

<table>
<thead>
<tr>
<th>Sentence</th>
<th>All Parameters</th>
<th>Selected Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sent. 1</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Sent. 2</td>
<td>6.5</td>
<td>3.2</td>
</tr>
<tr>
<td>Sent. 3</td>
<td>7</td>
<td>3.6</td>
</tr>
<tr>
<td>Sent. 4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Sent. 5</td>
<td>1</td>
<td>.5</td>
</tr>
<tr>
<td>Sent. 6</td>
<td>3</td>
<td>1.5</td>
</tr>
<tr>
<td>Sent. 7</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Sent. 4 + Sent. 5</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
and treated as one long utterance. Sentence 6 and sentence 7 yielded error rates between 2 and 3% for all four parameters and error rates between 1 and 1.5% for the selected parameters. These results are reasonable considering the fact that the tests were implemented in a laboratory computer.

Improvements in performance are even greater by the inclusion of a second sentence for verification. The addition of a second sentence in the verification process provides a considerable amount of information that is independent of the first sentence. Thus, by combining the information obtained for sentence 4 and sentence 5, the error rates drop considerably. This is equivalent to an additional selection process in which the sentence providing the least error rate is assigned to each speaker. The error rate obtained by this procedure is approximately half the error rate for the sentences considered separately (34).

5.4 Verification Time

A breakdown of the computation time for two seconds of speech is shown in Table 5.2. The entire process from silence removal to actual verification takes 35 to 45 seconds. Because of the nature of the linear prediction analysis, all the parameters viz parcor,
**TABLE 5.2**

PROCESSING TIME FOR SINGLE VERIFICATION

<table>
<thead>
<tr>
<th>OPERATION</th>
<th>TIME (SEC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silence Removal</td>
<td>5 (approximate)</td>
</tr>
<tr>
<td>Parameter Extraction</td>
<td>10</td>
</tr>
<tr>
<td>Smoothing</td>
<td>3</td>
</tr>
<tr>
<td>Time Registration</td>
<td>5</td>
</tr>
<tr>
<td>Comparison and Decision</td>
<td>5</td>
</tr>
<tr>
<td>Miscellaneous</td>
<td>10 (approximate)</td>
</tr>
<tr>
<td>TOTAL</td>
<td>38</td>
</tr>
</tbody>
</table>
linear predictor and autocorrelation coefficients can be obtained by a single set of operations on the speech samples. It takes approximately 10 seconds. The procedure of registration, comparison and decision takes approximately 10 seconds.

Parameters based on the short-time spectrum analysis are effective measures for speaker verification. However, a good amount of processing time is consumed in the computation of these parameters.

5.5 Verification Based on Composite Reference

Verification accuracy using the new concept of composite reference for each of the three sentences are presented. The sentences used in the experimentation were sentence 1, sentence 4 and sentence 5. The features for which composite reference were constructed are the linear predictor coefficient \( a_4 \) and the PARCOR coefficient \( b_3 \).

The speech data consisted of 45 utterances, consisting of nine repetitions of the same sentence spoken by five speakers. The composite reference file for sentence 1 was constructed from seven specimens of each speaker of a population of four speakers. The remaining two specimens of each speaker were used as test samples to determine the accuracy of the verification.
system: The composite reference file for sentences 4 and 5 were constructed from four specimens of the design set of each speaker of a population of four speakers. The remaining five specimens of each speaker and the specimen of the fifth speaker were used as test samples.

Table 5.3 summarizes the results of the evaluation. The error rates obtained with the linear predictor coefficient $a_4$ for sentence 1 is 4.5% for all the speakers whose specimens entered in the construction of the composite reference file. Speakers whose specimens did not enter in the construction of the composite reference file were not used for verification for sentence 1. The error rates obtained with the parcor coefficient $b_3$ for sentence 4 is 4% for four speakers whose specimens entered in the construction of the composite reference file. For the speaker that did not enter in the construction of the reference contour, all his sample contours were verified correctly. For sentence 5, the error rate is in the same order of magnitude as that of sentence 4, namely 4%. However, this error rate of 4% occurs when sample contours of the fifth speaker are tested against the composite reference contour.
<table>
<thead>
<tr>
<th>SENTENCE</th>
<th>PARAMETER</th>
<th>SPEAKERS INSIDE C.R.</th>
<th>SPEAKERS OUTSIDE C.R.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SENT. 1</td>
<td>a₄</td>
<td>4.5</td>
<td>NONE</td>
</tr>
<tr>
<td>SENT. 4</td>
<td>b₃</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>SENT. 5</td>
<td>b₃</td>
<td>0</td>
<td>4</td>
</tr>
</tbody>
</table>
The overall results of these evaluations show that verification with the use of the new concept of composite reference is viable. Further investigations involving large population of speakers and new features for verification are needed.

5.6 Verification Accuracy as a Function of Duration of Speech

The results presented so far have been based on sentences of length 1.5, two and three seconds in duration. In addition to these results, if a long sentence of three seconds in duration is restricted to 1.5 seconds of speech material, then the verification accuracy for the restricted sentence decreases. The utterance used for this study was sentence 5 of three seconds of duration. The restricted sentence was "Cash this bond, please".

Verification computed as a function of the duration of the spoken material is illustrated in Fig. 5.2. The results show a verification accuracy of about 93.5% for a duration of 1.5 seconds of speech. Accuracy of greater than 98% was obtained for durations of three seconds of speech and higher.
Fig. 5.2. Verification Accuracy as a Function of Duration of Speech
5.7 Verification Results Using Frequency-Domain Parameters

In addition to the linear predictor coefficients and an equivalent set of parameters derived from them, results were obtained using features obtained from a frequency-domain analysis of speech. A brief summary of our experimentation follows.

Among the frequency-domain speech processing techniques, the use of the Discrete Fourier Transform for a short-time spectrum analysis was investigated. The Fast Fourier Transform (FFT) was used as an efficient algorithm for evaluation of the Discrete Fourier Transform (DFT).

Energy computation up to 1 kHz in bands of 16 Hz was investigated. Basically, the procedure consisted of the following steps:

1) Compute an 8192 point FFT.

2) Obtain energy information up to 1 kHz for each band. The bandwidth was approximately 16 Hz.

3) Form a parameter contour using the computed value of energy at each band.

Figure 5.3 and Fig. 5.4 illustrate a trial and reference contour, respectively, for sentence 2 of the speaker JN. Figure 5.5 illustrates the energy
Speaker: JN
Parameter: Energy up to 1 kHz in Bands of 16 Hz
Utterance: "Papa needs two singers"

Fig. 5.3. Design Set and Trial Reference Contours
Speaker: JN

Parameter: Energy up to 1 kHz in bands of 16 Hz

Utterance: "Papa needs two singers"

Fig. 5.4. Registered Design Set and Final Reference Contours
Speaker: JN
Parameter: Energy Contours
Utterance: "Papa needs two singers"

Fig. 5.5. Parameter Contours for Bands of Energy
contents in three separate regions spaced at 1 kHz for the same utterance and the same speaker. Notice that the contour corresponding to the first region, namely, the 0 to 1 kHz region, has clear excursions and the peaks are well spaced. Verification results, using energy information up to 1 kHz, show error rates ranging between 5 to 8%.

Spectral energy computed over specific band of frequencies was also studied. The basic steps were:

1) Specify number of bands for which analysis will be carried out. In our case, schemes of 5, 6, 8 and 12 bands were studied.

2) Fix number of segments for the analysis of the pre-assigned verification sentence.

3) A value of \( N = 1024 \) for the FFT has been used so that if \( M \) is the length of each segment, then \((1024-M)\) zero samples must be appended to the windowed segment of speech samples.

4) Compute 1024 points FFT.

5) Compute energy within the specified bands.
6) Form parameter contour using the computed value of energy at each segment.

Figure 5.6 illustrates a final reference contour with three final warped sample contours for the speaker MS and sentence 1. The parameter shown in the figure is the spectral energy computed between 750 Hz and 1550 Hz in a 5-band scheme. The break-frequencies for the five bands were 50, 150, 350, 750, 1550 and 3150 Hz.

Figure 5.7, Fig. 5.8, and Fig. 5.9 illustrate parameter contours in a 5, 6, and 8 bands scheme for speakers MS, FE, and JN, respectively. The utterance in each case was sentence 2.

In a 5-band scheme using two sentence-long utterances, "cash this bond, please" and "Papa needs two singers", energy computed between 750 Hz and 1550 Hz was found to be a good measure. Verification accuracy of about 93 to 95% was achieved.

In a 6 and 8 bands scheme, energy computed between 800 Hz and 1600 Hz, and 600 Hz and 1000 Hz, respectively, were found to be good measures. They yielded verification error rates between 5 and 9%.

Finally, power spectrum computed via a Fast Fourier Transform was investigated. The main procedure is as follows:
Speaker: MS
Parameter: 750 to 1550 Hz Spectral Energy (5 bands)
Utterance: "Cash this bond, please"

Fig. 5.6. Registered Design Set and Final Reference Contours
Speaker: MS
Parameter: Spectral Energy (5 bands)
Utterance: "papa needs two singers"

Fig. 5.7. Spectral Energy Contours for 5 Bands
Speaker: FE
Parameter: Spectral Energy (6 bands)
Utterance: "Papa needs two singers"

Fig. 5.8. Spectral Energy Contours for 6 Bands
Speaker: JN
Parameter: Spectral Energy (8 bands)
Utterance: "Papa needs two singers"

Fig. 5.9. Spectral Energy Contours for 8 Bands
Speaker: JN
Parameter: Spectral Energy (8 bands)
Utterance: "Papa needs two singers"

Fig. 5.9. (Cont.)
1) Fix number of segments for analysis.
2) Equalize energy segment-by-segment.
3) Compute 128 points FFT of the windowed speech frame.
4) Compute power spectrum amplitudes for all frequency points up to the Nyquist frequency or a limited range of frequencies less than the Nyquist frequency. In our case, the power spectrum amplitudes were computed in the range of 0 to 10, to 20, to 40 frequency points.
5) Form the parameter contour from the power spectrum amplitude values computed framewise.

Figure 5.10 and Fig. 5.11 illustrate the original parameter contours of the speakers JN and FE for sentence 2. The parameter contours are the 9th and 9th squared power spectrum amplitudes, respectively. Figure 5.12 and Fig. 5.13 illustrate three final warped parameter contours and the final reference for the 9th and 9th squared power spectrum amplitudes.

Figure 5.14a and Fig. 5.14b illustrate power spectrum amplitudes up to 10 frequency points of the speaker FE for sentence 2.
Speaker: JN
Parameter: 9th Power Spectrum Amplitude
Utterance: "Papa needs two singers"

Fig. 5.10. Design Set - Parameter: 9th Power Spectrum Amplitude
Speaker: FE
Parameter: \((9^{\text{th}})^2\) Power Spectrum Amplitude
Utterance: "Papa needs two singers"

Fig. 5.11. Design Set - Parameter: \((9^{\text{th}})^2\) Power Spectrum Amplitude
Fig. 5.12. Registered Parameters and Final Reference Contours - 9th Power Spectrum Amplitude

Speaker: JN
Parameter: 9th Power Spectrum Amplitude
Utterance: "Papa needs two singers"
Speaker: FE
Parameter: $(9th)^2$ Power Spectrum Amplitude
Utterance: "Papa needs two singers"

Fig. 5.13. Registered Parameters and Final Reference Contours - $(9th)^2$ Power Spectrum Amplitude
Speaker: PE
Parameter: Power Spectrum Amplitudes
Utterance: "papa needs two singers"

Fig. 5.14a. Power Spectrum Amplitudes
Speaker: FE
Parameter: Power Spectrum Amplitudes
Utterance: "Papa needs two singers"

Fig. 5.14b. (Cont.)
After an exhaustive search among the power spectrum amplitudes up to 10 frequency points, the 9th power spectrum amplitude was found to be a good measure. Proceeding beyond 10 frequency points, the 28th power spectrum amplitude was also found to be a good measure. Figure 5.15 illustrates a design set for the 28th power spectrum amplitude for the speaker JN and the same utterance. Figure 5.16 shows registered parameters and final reference contours for the same speaker and the same utterance. Further, the square of the 9th and 28th power spectrum amplitudes showed better results. Figure 5.17 illustrates a design set for the 28th squared power spectrum amplitude for the speaker JN and the same sentence. Figure 5.18 shows registered parameters and final reference contours for the same speaker and the same sentence. Using the above parameters, verification error rates of about 6 to 8% were obtained. Also, it was found that there is a major limitation with this kind of automatic verification procedure. When the utterance, "papa needs two singers" was changed to the utterance, "cash this bond, please", the 9th and the 28th power spectrum amplitudes no longer were good measures for verification. That is, with the new utterance, "cash this bond, please", a new exhaustive search had to be done to be able to find good measures for
Speaker: JN
Parameter: 28th Power Spectrum Amplitude
Utterance: "Papa needs two singers"

Fig. 5.15 Design Set - Parameter: 28th Power Spectrum Amplitude
Speaker: JN
Parameter: 28th Power Spectrum Amplitude
Utterance: "Papa needs two singers"

Fig. 5.16. Registered Parameters and Final Reference Contours - 28th Power Spectrum Amplitude
Speaker: JN
Parameter: (28th)² Power Spectrum Amplitude
Utterance: "Papa needs two singers"

Fig. 5.17. Design Set - Parameter: (28th)² Power Spectrum Amplitude
Speaker: JN
Parameter: \((28\text{th})^2\) Power Spectrum Amplitude
Utterance: "Papa needs two singers"

Fig. 5.18. Registered Parameters and Final Reference Contours - \((28\text{th})^2\) Power Spectrum Amplitude
5.8 Discussion

The practical acceptability of any verification system depends on its expected error rate and the purpose for which it is intended to be applied. The overall error rates obtained in this experiment are reasonable and can be used to provide reliable speaker verification by machines. The error rate obtained for a selected number of features, namely the parcor coefficient \( b_6 \) and the autocorrelation coefficient \( \phi_7 \), gave quite good results considering its implementation on a laboratory computer. In addition, the error rates achieved with the features derived from the frequency-domain analysis are acceptable. There are many noncritical screening applications in which the present error rate of 6% to 8% achieved with this kind of parameter is acceptable \(^1\). Further research on measuring energy at specific bands is needed which could improve the error rates in speaker verification systems. Finally, the use of the new concept of composite reference information looks promising.
CHAPTER VI

CONCLUSIONS

The effectiveness of several different parametric representations of speech derived from the linear prediction model was determined for the automatic verification of speakers by their voices. The different speaker-dependent parameters investigated were the linear prediction coefficients, the parcor (partial correlation) coefficients, and the autocorrelation coefficients. In addition to the prediction parameters, a number of measures derived from the short-time spectrum analysis of speech was investigated. Such parameters included the computed energy values in frequency bands up to 1 kHz, spectral energy computed over specific band of frequencies, and power spectrum amplitudes computed via the Fast Fourier Transform. Also, automatic speaker verification using the new concept of composite reference was studied.

(i) Linear prediction parameters: An improved system for automatic speaker verification using linear prediction parameters has been implemented on a laboratory computer. The improvement over previous implementations includes the following:

122
a) Reliable results could be achieved by using sampling frequency as low as 4 kHz.

b) The order of the linear prediction model, and consequently, the computational effort, could be significantly reduced. A typical order of such a model is 6.

c) Only two of the six parameters were required.

d) The parcor or reflection coefficients and the autocorrelation coefficients were found to increase reliability.

e) The duration of the utterance should be longer than one sec. for acceptable verification accuracy.

(ii) Frequency-domain parameters: The various parameters derived from the short-time spectrum analysis of speech via the Fast Fourier Transform differed somewhat in the realized accuracy of verification of speakers, but not by a wide margin. Error rates of about 6% were obtained in a preliminary investigation. Further improvement can be obtained by measuring energy over specific bands of frequencies. In addition to the new frequency-domain parameters, a larger number of utterances and a larger number of utterance replications have been used to evaluate the performance of the
system. Also, a selected number of distances were used to determine the degree to which a given unknown sample contour differs from a reference contour.

(iii) Verification by composite reference: A new concept of verification by using composite reference information has been introduced. This new concept can be used to great advantage in combining several speaker identity information in a single file. Preliminary results using this new concept look promising. Further investigation is needed for implementation in a practical verification system. In addition to composite reference, a new registration procedure based strictly on correlation was developed. By this new technique, a sample parameter contour is time registered with a stored reference contour, representative of the identity claimed, by a linear shifting and warping of the sample contour time axis.

The results of this dissertation show that the predictor coefficients, or an equivalent set of parameters derived from them, obtained from speech sampled at a rate as low as 4 kHz, provide a very effective representation of speech for automatic speaker verification. The linear prediction characteristics provide a representation of the spectral envelope of the speech signal. That is, they represent combined information
about the formants, their bandwidths and the glottal waveforms.

Also, parametric representations derived from a short-time spectrum analysis of speech and the new concept of composite reference gave promising results. The verification accuracy achieved by the linear prediction parameters was found to be relatively higher than that achieved by the parameter derived from the short-time spectrum analysis. Of course, in any practical implementation, all of these speech characteristics could be used to provide reliable speaker verification by machines.

There are several facets of the speaker recognition problem which deserve further attention. Most of the 'systems' considered so far in speaker recognition (verification and identification) are exploratory studies. Research of speaker recognition systems that would operate under "real world" conditions is needed. Considerable work needs to be done with various types of noise and degradation. For example, the issue of optimal determination of the linear prediction parameters in the presence of noise should be given further attention. In addition, there are several interesting theoretical and practical issues associa-
ted with the quest for new speaker-dependent features for automatic speaker recognition. Parameters that are relatively constant across an analyzed utterance, and can thus be considered independent of the linguistic information and be highly indicative of the speaker are needed. Thus, it is the characterization of the speaker, and not the linguistic content, that will lead possibly to achieving text-independent speaker recognition and will certainly eliminate time-normalization of speech.
APPENDIX A

MECHANISM AND MODEL OF SPEECH PRODUCTION
APPENDIX A

MECHANISM AND MODEL OF SPEECH PRODUCTION

Figure A1 illustrates the important parts that are involved in the production of speech\(^2\). The principal organs involved in speech production are the lungs, trachea, throat, mouth and nose. Speech sounds are produced as a result of acoustical excitation of the vocal tract which consists of the cavities in the pharynx and the mouth. The vocal tract may be thought of as an acoustic tube of variable cross-sectional area which is terminated at the upper end by the lips and the lower end by the vocal cords. The shape of the vocal tract is changed continually during speech production by the movement of different articulators such as the tongue, the jaw and the lips.

For most speech sounds, the vocal cords are made to vibrate at a particular frequency which, for a normal adult, is about 60 to 400 Hz. Bursts of air are allowed to enter the throat cavity. The air then enters the vocal tract and the sound is radiated at the lips. In the case of nasal sounds, the front part of the vocal tract is closed and the vocal tract is coupled through the velar opening to the nasal cavities,
Fig. A1. Schematic Diagram of the Human Vocal Mechanism
thereby producing sound radiation from the nostrils. During the production of nonnasal sounds, the velar opening is closed and no sound is radiated from the nostrils.

Sounds are classified according to their mode of excitation, and there are three modes of exciting the vocal tract.

Voiced sounds are produced by forcing air to flow through the vocal cord orifice causing the cords to vibrate. The opening and closing of the glottis produces a series of nearly periodic pulses of air which excite the vocal tract. The fundamental frequency of the vocal cord vibrations is determined by the mass and tension of the vocal cords, and by the subglottal air pressure. Both the tension of the vocal cords and the subglottal air pressure change during speech production. Examples of voice sounds are vowels, semi-vowels, voiced stops and nasals.

Unvoiced sounds are produced by forming a constriction at some point in the vocal tract, anywhere from the glottis to the lips. Air is forced through these constrictions, causing a noise-like turbulent flow of air to be produced which is heard as hissing noise. The resulting excitation of the vocal tract has a relatively uniform spectrum and a very low intensity.
Examples of unvoiced sounds are various fricatives, such as f, s, sh, etc.

Finally, the plosive sounds are produced by making a complete closure and abruptly releasing it. The abrupt release of pressure provides an impulsive excitation of the vocal tract. Examples of plosive sounds are stop consonants like b, d, g, p, t, and k.

In continuous speech, combinations of the different types of excitation of the vocal tract mentioned above are also used.

Figure A2 illustrates a simple model of speech production. In this model, the vocal tract is represented by a linear time-varying filter which is excited by either random noise or a quasi-periodic pulse train, or both. The output of the vocal tract is then expressed as a convolution between the excitation function and the impulse response of the time-varying linear filter representing the vocal tract. Thus, the speech signal $S(t)$ is given by

$$S(t) = \int_{-\infty}^{t} u(\tau)v(t, \tau)d\tau \quad (A.1)$$

where $v(t, \tau)$ is the response of the vocal tract at time $t$ to a delta function input applied at time $\tau$, and $u(t)$ is the excitation function.
Fig. A2. Block Diagram of a Simplified Model of Speech Production
The process of speech production can be considered to be quasi-stationary, because the shape of the vocal tract changes slowly compared to the variations in the excitation waveform. The quasi-stationarity of the model is a strong motivation for the study of the speech characteristics in the frequency domain. For voiced speech, the impulse response of the vocal tract is characterized by its eigen-frequencies. The spectrum of the vocal-tract response consists of a number of eigen-frequencies whose locations depend upon the vocal-tract shape. In speech perception, only the first three eigen-frequencies or formants are important. For unvoiced speech, the excitation spectrum is uniformly distributed over a wide frequency range. The speech spectrum for unvoiced sounds reflects entirely the vocal-tract response.
APPENDIX B

LINEAR PREDICTOR PARAMETERS ESTIMATION
APPENDIX B

LINEAR PREDICTOR PARAMETERS ESTIMATION

In the linear prediction model, the combined filter transfer function of the vocal tract, the radiation, and the glottal-wave shape are represented by a discrete all-pole linear filter with \( p \) poles as shown in Fig. 3.2.

The output of the all-pole model is the speech signal \( S_n \) which at the \( n \)th sampling instant is estimated as

\[
S_n = - \sum_{k=1}^{p} a_k S_{n-k} + G U_n
\]  

(B.1)

where \( a_k \) are the predictor coefficients

\( U_n \) is the excitation function, and

\( G \) is the gain factor

The predictor coefficients are obtained here using the least square approach. By this method, it is assumed that \( U_n \) is totally unknown, and \( S_n \) can be predicted only approximately from a linearly weighted summation of past \( p \) samples, i.e.

\[
\hat{S}_n = - \sum_{k=1}^{p} a_k S_{n-k}
\]  

(B.2)
The error between the actual value $S_n$ and the predicted value $\hat{S}_n$ is given by

$$e_n = S_n - \hat{S}_n = S_n + \sum_{k=1}^{p} a_k S_{n-k}$$  \hspace{1cm} (B.3)

Next, the predictor coefficients $a_k$ are obtained as a result of the minimization of the mean or total squared error with respect to each of the coefficients.

If the total square error is denoted by $E$, where

$$E = \sum_{n} e_n^2 = \sum_{n} [S_n + \sum_{k=1}^{p} a_k S_{n-k}]^2$$  \hspace{1cm} (B.4)

then $E$ is minimized by setting

$$\frac{\partial E}{\partial a_i} = 2 \sum_{n} \left[ S_n + \sum_{k=1}^{p} a_k S_{n-k} \right] S_{n-i} = 0, \; 1 \leq i \leq p$$ \hspace{1cm} (B.5)

Rearranging terms we get the normal equations:

$$\sum_{k=1}^{p} a_k \sum_{n-k} S_{n-i} = -\sum_{n} S_n S_{n-i}, \; 1 \leq i \leq p$$ \hspace{1cm} (B.6)

The specification of the range of summation over $n$ in the above equation leads to two distinct methods for the estimation of the predictor parameters.

1) Autocorrelation method: In this method, the error in (B.4) is minimized over the infinite interval $-\infty < n < \infty$. Equation (B.6) is then reduced to
\[ \sum_{k=1}^{p} a_k R(i-k) = -R(i), \quad 1 \leq i \leq p \quad (B.7) \]

where

\[ R(i) = \sum_{n=-\infty}^{\infty} s_n s_{n+i} \quad (B.8) \]

represents the autocorrelation matrix function of the speech samples \( s_n \). This matrix is symmetric and all the elements along each of its diagonal are equal.

(2) **Covariance method**: In contrast with the autocorrelation method, here the total squared error \( E \) in (B.4) is minimized over a finite interval, say, \( 0 \leq n \leq N-1 \). \( N \) represents the number of samples in each interval. In our case, \( N \) is approximately equal to 40 samples of speech.

Equation (B.6) is then reduced to

\[ \sum_{k=1}^{p} a_k \psi_{ki} = -\psi_{oi}, \quad 1 \leq i \leq p \quad (B.9) \]

where

\[ \psi_{ik} = \sum_{n=0}^{N-1} s_{n-i} s_{n-k} \quad (B.10) \]

represents the covariance matrix function of the speech samples \( s_n \) in the given interval. The covariance matrix (B.10) is also symmetric, but the terms along each diagonal are not equal. This fact can be seen by
expanding (B.10)

\[ \psi_{i+1,k+1} = S_{i-1}S_{k-1} + S_{i}S_{k} + \ldots + S_{N-2-i}S_{N-2-k} \]  
\hspace{1cm} (B.11)

From the above expression, \( \psi_{i+1,k+1} \) can be written in terms of \( \psi_{ik} \) as follows:

\[ \psi_{i+1,k+1} = \psi_{ik} + S_{i-1}S_{k-1} - S_{N-1-i}S_{N-1-k} \]  
\hspace{1cm} (B.12)

In this equation, values of the speech signal \( S_n \) for \(-p \leq n \leq N-1\) must be known. In fact, a total of \( p + N \) samples are needed.

The covariance method reduces to the autocorrelation method as the interval over which \( n \) varies approaches infinity. In each of the two formulations, namely, the autocorrelation and covariance methods, the predictor coefficients \( a_k, 1 \leq k \leq p \), can be computed by solving a set of \( p \) linear equations with \( p \) unknowns. These equations are (B.7) for the autocorrelation method, and (B.9) for the covariance method. There exist several procedures for obtaining solutions to these equations. Most of these solutions are based on the useful fact that the matrix of coefficients in both methods is a covariance matrix. The nice property of covariance matrices is that they are symmetric and in general positive semidefinite, although in practice
they are usually positive definite.

In particular, the autocorrelation normal equations (B.7) can be expanded in matrix form as

\[
\begin{bmatrix}
R_0 & R_1 & R_2 & \cdots & R_{p-1} \\
R_1 & R_0 & R_1 & \cdots & R_{p-2} \\
R_2 & R_1 & R_0 & \cdots & R_{p-3} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
R_{p-1} & R_{p-2} & R_{p-3} & \cdots & R_0
\end{bmatrix}
\begin{bmatrix}
a_1 \\
a_2 \\
a_3 \\
\vdots \\
a_p
\end{bmatrix}
= -
\begin{bmatrix}
R_1 \\
R_2 \\
R_3 \\
\vdots \\
R_p
\end{bmatrix}
\]  

(B.13)

where the p×p autocorrelation matrix is symmetric and the elements along any diagonal are identical. Levinson (35), Robinson (29), and later Durbin (36) derived an elegant recursive procedure for solving this type of equation.

Once the predictor coefficients \( a_k \) are computed, then the PARCOR (partial correlation) coefficients are obtained from them by a backward recursion. The recursion is as follows:

Initially set: \( a_i^{(p)} = a_i, 1 \leq i \leq p \). Then

1) \( b_i = a_i^{(i)} \)

2) \( a_i^{(i-1)} = a_j^{(i-1)} \frac{a_i^{(i)}}{1 - b_i^2}, 1 \leq j \leq i-1 \)
where \( i = p, p-1, \ldots, 1 \)

From all the above results, a Fortran computer program, such as shown in Fig. A3 can be written to solve for the predictor coefficients \( a_k \), the autocorrelation coefficient \( \phi_k \), and the parcor coefficients \( b_k \).
SUBROUTINE AKN(X,A,CK,R,NP,L)

THE ARRAY X REPRESENTS THE INPUT AND CONTAINS ALL
THE SPEECH SAMPLES IN A PARTICULAR SEGMENT
L - REPRESENTS THE LENGTH OF THE SEGMENT
NP - REPRESENTS THE ORDER OF THE LPC ANALYSIS
THE ARGUMENTS A, CK, R ARE THE OUTPUTS WHERE
A - REPRESENTS THE LINEAR PREDICTION COEFFICIENT
CK - REPRESENTS THE PARCOR (REFLECTION) COEFFICIENT
R - REPRESENTS THE AUTOCORRELATION COEFFICIENT

DIMENSION X(L),A(NP),CK(NP),R(NP)

IP=NP+1
DO 1 J=1,IP
SUM=0.
KK=L-J+1
DO 2 K=1,KK
SUM=SUM+X(K)*X(K+J-1)
2 CONTINUE
R(J)=SUM
1 CONTINUE
A(1)=-R(2)/R(1)
CK(1)=A(1)
AL=R(1)*A(1)*R(2)
BE=R(3)*A(1)*R(2)
MM1=NP-1
DO 10 I=1,MM1
IP1=I+1.
CC=-BE/AL
CK(IP1)=CC
I2=IP1/2
DO 20 J=1,I2
1J=IP1-J
TA=A(J)+CC*A(1J)
A(1J)=A(1J)+CC*A(J)
20 A(J)=TA
A(IP1)=CC
AL=AL+CC*BE
BE=R(I+3)
DO 10 J=1,IP1
NJ=I-J+3
10 BE=BE+A(J)*R(NJ)
RETURN
END

Fig. A3. Fortran Program for Solving the Autocorrelation Equations
APPENDIX C

BRIEF COMPARISON OF SPEAKER VERIFICATION
AND SPEAKER IDENTIFICATION
APPENDIX C

BRIEF COMPARISON OF SPEAKER VERIFICATION AND
SPEAKER IDENTIFICATION

The discussion of speaker verification and speaker identification will be limited to a population of $N$ speakers. The probability of error for speaker verification will be shown to be independent of $N$. In contrast, it will be shown that as $N$ becomes large, speaker identification becomes much more difficult, and the probability of error approaches $1^\text{(41)}$.

In the verification process, a given customer claims an identity and volunteers to speak a pre-arranged verification phrase. A set of measurements is applied to the offered voice sample, and the results are compared with the stored reference data that exists on file for the specific claimed identity. It must then be decided whether the claim is valid or not. To be more precise, given the voice sample $X$ and the claimed identity of speaker $j$, the claim will be accepted if $X$ is likely enough to have originated from speaker $j$. The criterion of likely enough means:

$$p_j(X) > c_j \left< p_k(X) \right>_k$$

143
where $C_j$ is a constant which is chosen to effect the desired balance between errors of false acceptance and false rejection, and

$$< p_k(X) >_k = \sum_{k=1}^{N} p_k(X) p_r(k)$$

where $N$ represents the population of possible users of the system, and $p_r(k)$ is the probability that the $k$th user accesses the system. As the population size $N$ becomes large, $< p_k(X) >_k$ will generally approach some limiting density, so that the tendency is for the system performance to be independent of $N$.

The overall error rate in a verification task depends upon the apriori probabilities that a random customer in the population is an honest person or an impostor. Generally, these apriori probabilities are assumed to be equal.

The machine can do nothing about a random customer who, having learned his closeness to a given legitimate customer, can fool the system at will. However, the decision can be made so that it can be classified as "doubtful" in cases of uncertainty. The extent of the doubtful cases can be tailored to the gravity of the requested transaction. Such situations can be corrected by asking the customer, for instance, to
speak another single utterance or set of utterances. By contrast, the task of speaker identification involves no claimed identity from the customer and the main problem is that of identifying an unknown speaker in a population of N-1 speakers. The machine must examine its library of stored references and, in absence of a more sophisticated strategy, make N-1 comparisons and indicate the closest match to the unknown speaker. In other words, given only the voice sample X, it must single out the speaker from whom X is most likely to have originated. That is, for all k choose the jth speaker for which

\[ P_j(X) > P_k(X). \]

The decision process can be thought of as N-1 comparisons of \( P_j(X) \) with other densities of X. If \( P_j(X) \) fails to be larger than \( P_k(X) \) for any of the N-1 other speakers, then the identification is incorrect. Thus, the probability of error for speaker j, \( P_{E_j} \), increases as N increases:

\[ P_{E_j} = 1 - P_x \text{ (no error)} \]

\[ = 1 - P_x \text{ (no error at every comparison)} \]
\[ N \sum_{k=1}^{N} \frac{1}{k-1} \text{ (no error when comparing } \sum_{k=1}^{N} \frac{1}{k} \text{ and } N) \]

Also the effort involved in identification is directly proportional to \( N \).
APPENDIX D

BRIEF DESCRIPTION OF EXPERIMENTAL SET-UP
APPENDIX D

BRIEF DESCRIPTION OF EXPERIMENTAL SET-UP

The basic block-units of the experimental set-up are the signal generator and the signal processor. See Fig. 5.1. The main elements of the signal generator are the Crown tape recorder and the Krohn-Hite variable filter.

Crown Tape Recorder, Model SX800: The recorded speech samples were played back on a Crown model SX800 four-track professional tape recorder. The operating speed was 7.5 inches per second which is the maximum speed for this specific machine. The recordings were made on Scotch 202 magnetic tape which was the recommended tape for this particular tape recorder.

Krohn-Hite Variable Filter Model 3750(R): The speech signal coming out of the tape recorder was pre-filtered prior to its digitation. The filtering of the recorded speech samples was achieved by using a Krohn-Hite variable filter, Model 3750(R). This model is a variable electronic filter covering the frequency range from 0.02 Hz to 20 kHz, and operable in any one of four functional modes: high-pass, low-pass, band-pass or band-reject. The pass-band gain is unity (0 dB)
or 10 (20 dB) with attenuation slopes of 6, 12, 18 or 24 dB per octave.

The cut-off frequencies of both the high-pass and the low-pass sections are independently adjustable from 0.02 Hz to 20 kHz. Operation in the band-pass or band-reject modes permits passage or rejection of any band of frequencies between these limits. In the low-pass and band-reject modes, the filter is direct coupled, passing all frequencies from DC to the selected cut-off frequency of the low-pass section. In the high-pass and band-reject modes, the upper 3 db point is at approximately 1 megahertz for 0 db gain, and approximately 200 kHz for 20 db gain.

As far as response characteristics, there is a choice of Butterworth (maximum flat response) for frequency domain operation and Low Q (damped response) for transient-free time domain operation. The input can vary between ±15 volts for a gain of 0 dB and ±1.5 volts for a gain of 20 dB. The input impedance is 10 megohms in parallel with a 200 pf capacitor and the output impedance is approximately 50 ohms.

The basic units of the signal processor consist of an analog-to-digital converter, and a laboratory computer with its peripheral devices.
Tustin X-1500 A/D Convertor: After the analog speech signal was low-pass filtered, it was then converted to digits using the Tustin X-1500 A/D Convertor. The A/D X-1500 Data System consists of a digital-to-analog conversion portion and an analog-to-digital conversion portion, housed in a common cabinet.

The Analog-to-Digital portion consists of four simultaneous sample and hold amplifier channels and twelve multiplexor channels for a total input capability of sixteen channels (4 buffered, 12 unbuffered). The analog input level is ±10 volts full scale. The simultaneous sample and hold amplifiers are connected to the first four channels of the multiplexor. Following the multiplexor is a sample and hold amplifier and a total of 14 binary bits (including sign) analog-to-digital converter. The input impedance is greater than 20 megohms shunted by less than 200 pf capacitance. The conversion time is 15 μsec with an accuracy of 0.005% of full scale, ±0.5 LSB. The actual coding is done using 2's complement.

Datagen Nova 840 Laboratory Computer and Peripheral Devices: The Nova 840 is a general purpose laboratory computer which operates using 16 bit word length. It is organized around four accumulators, two of which can be used as index registers. The working core memory
size is 32 K with a combined disk storage capability of approximately 2.5 million words. The peripheral equipment included a Vistar Infoton CRT terminal Textronix Graphics terminal, and the Textronix Hard Copy Unit. The infoton CRT terminal is an alphanumeric display designed for a two-way data transmission with the computer. It can be utilized as a substitute for a teletype in either an unbuffered (character-by-character) conversational mode or in a buffered (block) mode transmitting either a line or a page at a time.

The Textronix 4013 Graphics terminal consists of two parts: the Display Unit and the Pedestal Unit. The Display Unit houses the storage tube, keyboard, and their associated circuits. The Pedestal Unit contains the terminal control circuits, the low voltage power supply, and the interfaces. The terminal was used in our experimentation as means to display data from the computer such as parameter contours, and speech waveforms. The terminal's response to data is not restricted to writing on the screen. It can also perform a variety of functions in response to "control character" commands; such as controlling the display format, switching modes, and controlling outputs from the terminal.
The Textronix 4610 Hard Copy Unit was used to get reproductions of displays which appear on the Graphics Terminal screen.
APPENDIX E

INSTRUCTION TO THE SPEAKERS
APPENDIX E
INSTRUCTION TO THE SPEAKERS

All the speakers were male, Canadian and members of the Engineering Faculty of the University of Windsor. They were all briefed informally on the nature of the verification experiment, and in addition, the following written description was given to each of them during every recording session:

This is a speaker verification experiment. A verification task consists of distinguishing a known cooperative talker from the set of all other talkers. In order to do this, the speaker in question identifies himself and gives a natural voice utterance. Then the computer must decide the question, "Is this speaker so-and-so?"

Suppose you go to the bank one morning to find the money withdrawal slip-check system replaced by a voice-check system. The system compares your utterance of a selected key phrase with previous recording samples. Similarity of your on-the-spot utterance with the previous recorded
samples governs your acceptance or rejection to withdraw money.

In this experiment, it is important for you to maintain consistency between recordings of the key phrase. I suggest that you try merely to speak naturally, using your own natural stress and speed.

First give your complete name followed by a pause. Then speak the following key phrase: "My name is Miller, cash this bond, please". In order to reduce the large variations in talking behaviour, you may like to repeat the key phrase a few times so that you can get familiar with the utterance.

If you feel that you are "out of voice" with a head cold or whatever, please let me know.

If any speaker indicated that his voice was not normal for some reason, his utterance for that session was discarded.
APPENDIX F

FORTRAN PROGRAM LISTINGS FOR SPEAKER VERIFICATION
DIMENSION NAME(5), DC(500), IX(200), E(1500)

C
C REMOVAL OF UNVOICED SPEECH SEGMENT
C THE ARRAY IX CONTAINS ORIGINAL SAMPLE CONTOUR
C THE ARRAY DC CONTAINS THE DC VALUES SEGMENTWISE
C THE ARRAY E CONTAINS THE ENERGY VALUES SEGMENTWISE
C
ACCEPT '# OF POINTS PER SEGMENT = ', N
ACCEPT 'CUTOFF FACTOR, ATT = ', ATT
6 ACCEPT 'LENGTH OF FILE = ', NT
NSEG=NT/N
NL=NT-NSEG*N
TYPE "ENTER OUTPUT FILENAME"
READ(11,1) (NAME(I), I=1,5)
1 FORMAT(5A2)
OPEN 1,NAME
TYPE "ENTER INPUT FILENAME"
READ(11,1) (NAME(I), I=1,5)
OPEN 0,NAME
IC=0
AVE=0.
DO 2 K=1,NSEG
DC(K)=0.
READ(0) (IX(J), J=1,N)
DO 5 J=1,N
DC(K)=DC(K)+IX(J)
5 CONTINUE
DC(K)=DC(K)/FLOAT(N)
2 CONTINUE
CLOSE 0
OPEN 0,NAME
DO 3 K=1,NSEG
EK(K)=0.
READ(0) (IX(J), J=1,N)
DO 10 J=1,N
IX(J)=IX(J)-DC(K)+0.5*SIGN(1., FLOAT(IX(J)))*A=IX(J)
EK(K)=EK(K)+A**2
10 CONTINUE
AVE=AVE+EK(K)
CONTINUE
AVE=AVE/FLOAT(NSEG)
CLOSE 0
OPEN 0,NAME
DO 11 K=1,NSEG
READ(0) (IX(J),J=1,N)
IF(E(K).LT.AVE/ATT)GO TO 11
WRITE(1) (IX(J),J=1,N)
IC=IC+1
11 CONTINUE
IF(NL.EQ.0)GO TO 8
READ(0) (IX(J),J=1,NL)
WRITE(1) (IX(J),J=1,NL)
8 NSIZE=IC*N+NL
TYPE NSIZE
CLOSE 1
CLOSE 0
ACCEPT 'WISH TO CONTINUE ? YES-1,NO-0 ','IWISH
IF(IWISH.EQ.1)GO TO 6
END
COMMON NH, NSH, NSEG, NA, NS, NPS, NK
DIMENSION Y(3400), NAME(5), Y1(200), YA(200), YAA(200), YO(3400)

CONSTRUCTION OF THE REFERENCE CONTOUR
THE ARRAY Y CONTAINS THE DESIGN SET AND AVERAGE
THE ARRAYS YO AND Y1 ARE WORKING ARRAYS FOR
TEMPORARY STORAGE
THE ARRAYS YA AND YAA HOLD THE COMPUTED AVERAGES
OBTAINED DURING THE PROCESS

ACCEPT 'LENGTH OF INPUT FILE = ', LK
ACCEPT 'LENGTH OF EACH FILE = ', NK
ACCEPT ' # OF POINTS PER SEGMENT = ', NPS
ACCEPT ' # OF SPECIMENS = ', NSS
ACCEPT 'MAXIMUM SHIFT = ', NST
TYPE "ENTER INPUT FILENAME"
READ(11,1) (NAME(I), I=1,5)

1 FORMAT(5A2)
OPEN 0, NAME
READ(0) (Y(I), I=1, LK)
CLOSE 0

TYPE "ENTER FILENAME FOR OUTPUT"
READ(11,1) (NAME(I), I=1,5)

NSH=2*NST+1
NSEG=NK/NPS
NA=NSEG*NPS
NS=(NK-NA)/2
NL=NSS*NK
DO 13 I=1, LK

13 Y0(I)=Y(I)

DO 12 I=1, NK

12 YA(I)=Y(NL+I)

DO 11 L=1, NSS

11 NP=(L-1)*NK

DO 2 I=1, NK

2 Y1(I)=Y0(I+NP)

CALL SRSHIFT(Y1, YA)

DO 30 I=1, NK

30 Y(I+NP)=Y1(I)

CONTINUE

DO 4 I=1, NK

4 SUM=0.

DO 20 J=1, NSS

20 SUM=SUM+Y(I+(J-1)*NK)

4 YAA(I)=SUM/FLOAT(NSS)
SUM=0.
SUM1=0.
DO 5 J=1,NK
Z=YA(J)-YAA(J).
SUM=SUM+Z*Z
SUM1=SUM1+YA(J)*YA(J)
AE=SUM/SUM1
DO 17 J=1,NK
17 YA(J)=YAA(J)
IF(AE.LT.0.002)GO TO 9
GO TO 3
9 DO 8 J=1,NK
8 Y(NL+J)=YA(J)
LK=NS
NS1=NSS+1
DO 16 I=1,NS1
L=(I-1)*NA
DO 15 J=1,NA
15 Y(L+J)=Y(L+J+LK)
LK=LK+2*NS
16 CONTINUE
L=NA*NS1+2
Y(L-1)=FLOAT(NSH)
Y(L)=FLOAT(NS)
WRITE(10,7) SUM,SUM1,AE
7 FORMAT(3(1X,G19.4))
OPEN 0,NAME
WRITE(0) (Y(I),I=1,L)
CLOSE 0
END
SUBROUTINE SRSMSHIFT(Y, X)

This subroutine time shifts a given contour
the array Y contains the input and output
sample contours
the array X contains the reference contour.

COMMON NSW, NSEG, NA, NS, NPS, N
DIMENSION X(200), Y(200), Z(200)
DO 5 K=1, NSEG
BIG=0
ISH=0
L=(K-1)*NPS
DO 2 KS=1, NSH
SUM=0
SUMX=0
SUMY=0
DO 1 J=1, NPS
LJ=L+J+NS
IS=LJ*(KS-4)
IF(IS.LT.1 OR IS.GT.N)GO TO 1
SUM=SUM+X(LJ)*Y(IS)
SUMX=SUMX+X(LJ)**2
SUMY=SUMY+Y(IS)**2
CONTINUE
1 COR=SUM/SQRT(SUMX*SUMY)
IF(BIG.GE.COR)GO TO 2
BIG=COR
ISH=KS-4
CONTINUE
2 CONTINUE
DO 4 J=1, NPS
LJ=L+J+NS
LJI=LJ+ISH
IF(LJI.LT.1 OR LJI.GT.N)GO TO 3
Z(LJ)=Y(LJI)
GO TO 4
3 Z(LJ)=0
4 CONTINUE
5 CONTINUE
DO 8 J=1, N
8 Y(J+NS)=Z(J+NS)
RETURN
END
DIMENSION NAME(5),X(3400),Y(200),Z(200)

WARPING ALGORITHM USING LINEAR SHIFTING BASED ON
CORRELATION
THE ARRAY X CONTAINS THE REFERENCE INFORMATION
THE ARRAY Y CONTAINS THE SAMPLE CONTOUR TO BE WARPED
THE ARRAY Z CONTAINS THE WARPED SAMPLE CONTOUR

ACCEPT 'LENGTH OF EACH FILE =',N
ACCEPT 'LENGTH OF REFERENCE FILE =',LK
ACCEPT '# OF POINTS PER SEGMENT =',NPS
ACCEPT '# OF SPECIMENS =',NSS
ACCEPT 'BEGINNING OF FILE =',IB
TYPE "ENTER REFERENCE FILENAME"
READ(11,9) (NAME(I),I=1,5)
FORMAT(5A2)
OPEN 0,NAME
READ(0) (X(I),I=1,LK)
CLOSE 0

TYPE "ENTER UNKNOWN FILENAME"
READ(11,9) (NAME(I),I=1,5)
OPEN 0,NAME
READ(0) (Y(I),I=1,N)
CLOSE 0

NSH=X(LK-1)+0.5
NS=X(LK)+0.5
NSEG=N/NPS
NA=NSEG*NPS
NK=NSS*NA
DO 6 I=1,NA

Y(I)=Y(IB+I)
DO 5 K=1,NSEG
BIG=.FALSE.
ISH=0
L=(K-1)*NPS
DO 2 KS=1,NSH
SUM=0.
SUMX=0.
SUMY=0.
DO 1 J=1,NPS
LJ=L+J
IS=LJ+(KS-4)*NS
IF(IS.LT.1.OR.IS.GT.NJ)G0 TO 1
SUM=SUM+X(LJ+NK)*Y(IS)
SUMX=SUMX+X(LJ+NK)**2
SUMY=SUMY+Y(IS)**2
1 CONTINUE
PP=SUMX*SUMY
COR=SUM/SQRT(PP)
IF(BIG.GE.COR)GO TO 2
BIG=COR
ISH=KS-4
2 CONTINUE
DO 4 J=1,NPS
LJ=L+J
LJI=LJ+ISH+NS
IF(LJI.LT.1.OR.LJI.GT.N)GO TO 3
Z(LJ)=Y(LJI)
GO TO 4
3 Z(LJ)=0
4 CONTINUE
5 CONTINUE
SUM=0.
SUMX=0.
SUMZ=0.
DO 7 J=1,NA,
SUM=SUM+X(NK+J)*Z(J)
SUMX=SUMX+X(NK+J)*X(NK+J)
SUMZ=SUMZ+Z(J)*Z(J)
N1=NA+1
PP=SUMX*SUMZ
Z(N1)=SUM/SQRT(PP)
TYPE Z(N1)
TYPE "ENTER FILENAME FOR OUTPUT"
READ(11,1) (NAME(I),I=1,5)
OPEN 0,FILE=NAME
WRITE(0) (Z(I),I=1,N1)
CLOSE 0
ACCEPT 'WISH TO CONTINUE ? YES-1,NO-0 'IWISH
IF(IWISH.EQ.1)GO TO 8
END
DIMENSION S(340),X(200),ROH(200),SIGMR(300),
*G(20),NAME(5),D4(200)

DISTANCE COMPUTATION USING R.C. LUMMIS'S ALGORITHM
THE ARRAY S CONTAINS THE REFERENCE INFORMATION
THE ARRAY Q CONTAINS THE CORRELATION VALUES
THE ARRAY X CONTAINS THE SAMPLE CONTOUR
THE ARRAYS ROH AND SIGMR CONTAIN THE SEGMENTAL
CORRELATION AND VARIANCE VALUES RESPECTIVELY
THE QUANTITIES D4 AND D5 REPRESENT THE SEGMENTAL
AND THE OVERALL DISTANCE RESPECTIVELY
THE QUANTITY D7 REPRESENTS THE PERCENTAGE ERROR

ACCEPT 'LENGTH OF EACH FILE = ' ,N
ACCEPT 'LENGTH OF REFERENCE FILE = ' ,LN
ACCEPT '# OF SEGMENTS = ' ,L1
ACCEPT '# OF POINTS PER SEGMENT = ' ,L2
ACCEPT '# OF SPECIMENS = ' ,NS
TYPE "ENTER REFERENCE FILENAME"
READ(11,1) (NAME(I),I=1,5)
FORMAT(5A2)
OPEN 0,NAME
READ(0) (S(I),I=1,SN)
CLOSE 0

TYPE "CORRELATION VALUES FILENAME"
READ(11,1) (NAME(I),I=1,SN)
OPEN 0,NAME
READ(0) (Q(I),I=1,NS)
CLOSE 0

TYPE "ENTER UNKNOWN UTTERANCE"
READ(11,1) (NAME(I),I=1,SN)
OPEN 0,NAME
N1=N+1
READ(0) (X(I),I=1,N1)
CLOSE 0
NC=NS*N
NCI=NC+N
DO 7 J=1,N
S(NC1+J)=X(J)
7 NS2=NS+2
BIG=0
DO 10 K=1,L1
SIGMR(K)=0
KL=(K-1)*L2
DO 2 I=1,NS2
SUM1=0
SUM2=0
SUM3=0
L=(I-1)*N+KL
2 SIGMR(K)=SIGMR(K)+L
DO 5 J=1, L2
SUM1=SUM1+S(L+J)*S(NC+KL+J)
SUM2=SUM2+S(L+J)*S(L+J)
5 SUM3=SUM3+S(NC+KL+J)*S(NC+KL+J)
XXX=SUM2*SUM3
ROH(I)=SUM1/SORT(XXX)
2 CONTINUE
DO 4 J=1, NS
SIGMR(K)=SIGMR(K)+(1.-ROH(JT)**2
4 CONTINUE
SIGMR(K)=SIGMR(K)/(NS-1.)
XXX=(1.-ROH(NS+2)**2/SIGMR(K)
D4(K)=XXX
IF(BIG.GT.XXX)GO TO 10
BIG=XXX
LS=K
10 CONTINUE
SUM3=0.
DO 8 J=1, L1
IF(J.EQ.LS)GO TO 8
SUM3=SUM3+D4(J)
8 CONTINUE
D44=SUM3/FLOAT(L1-1)
SIGMP=0.
DO 6 J=1, NS
BIG=SIGMP+(1.-Q(J))**2
6 SIGMP=BIG
BIG=SIGMP/FLOAT(NS)
D5=(1.-X(N1))**2/BIG
D6=D44*D44+D5*D5
SUM1=0.
SUM2=0.
DO 15 J=1, N
SUM1=SUM1+S(NC+J)**2
PP=(S(NC+J)-S(NC1+J))**2
15 SUM2=SUM2+PP
D7=SUM2/SUM1*100.
WRITE(10, 14) D44, D5, D6, D7
FORMAT(4(1X, G12.4))
ACCEP 'WISH TO CONTINUE? YES-1, NO-0', IWISH
IF(IWISH.EQ.1)GO TO 11
END
DIMENSION S(2200),X(200),ROH(200),NAME(5),DA(40)

EXPONENTIAL DISTANCE COMPUTATION
THE ARRAY S CONTAINS THE REFERENCE INFORMATION
THE ARRAY X CONTAINS THE SAMPLE CONTOUR
THE QUANTITIES D3 AND DS REPRESENT THE SEGMENTAL
AND OVERALL DISTANCES (EXPONENTIAL) RESPECTIVELY

ACCEPT 'LENGTH OF EACH FILE =', N
ACCEPT 'LENGTH OF REFERENCE FILE =', LN
ACCEPT '# OF SEGMENTS =', L1
ACCEPT '# OF POINTS PER SEGMENT =', L2
ACCEPT '# OF SPECIMENS =', NS
TYPE "ENTER REFERENCE FILENAME"
READ(11,1) (NAME(I),I=1,5)

1 FORMAT(5A2)
OPEN 0,NAME
READ(0) (S(I),I=1,LN)
CLOSE 0

11 TYPE "ENTER UNKNOWN UTTERANCE"
READ(11,1) (NAME(I),I=1,5)
OPEN 0,NAME
N1=N+1
READ(0) (X(I),I=1,N1)
CLOSE 0
NC=NS*N
NC1=NC+N
DO 7 J=1,N
7 S(NC1+J)=X(J)
NS2=NS+2
BIG=0.
DO 10 K=1,L1
KL=(K-1)*L2
I=NS2
SUM1=0.
SUM2=0.
SUM3=0.
L=(I-1)*N+KL
DO 5 J=1,L2
SUM1=SUM1+S(L+J)*S(NC+KL+J)
SUM2=SUM2+S(L+J)*S(L+J)
SUM3=SUM3+S(NC+KL+J)*S(NC+KL+J)
XXX=SUM2*SUM3

5 CONTINUE
ROH(I)=SUM1/SORT(XXX)
XXX=1./(ROH(NS2)*ROH(NS2))
IF(XXX.LT.0. OR XXX.GT.150.)XXX=150.
E=EXP(XXX)
D4(K)=E
IF(BIG.GT.E)GO TO 10
BIG=E
LS=K
10 CONTINUE
SUM3=0.
DO 15 J=1,LS
IF(J.EQ.LS)GO TO 15
SUM3=SUM3+D4(J)
15 CONTINUE
XXX=1./(X(N1)*X(N1))
D5=EXP(XXX)
D3=SUM3/FLOAT(L1-1)
D6=D3*D3+D5*D5
WRITE(10,14) D3,D5,D6
14 FORMAT(3G20.8)
ACCEPT 'WISH TO CONTINUE ? YES=1,NO=0 ',IWISH
IF(IWISH.EQ.1)GO TO 11
END
DIMENSION X(200),Y(3400),NAME(5),Q(30)

COMPUTATION OF CORRELATION BETWEEN TWO CONTOURS
THE ARRAY Y CONTAINS REFERENCE INFORMATION
THE ARRAY X CONTAINS SAMPLE CONTOUR
THE ARRAY Q CONTAINS THE OUTPUT

8 ACCEPT 'LENGTH OF EACH FILE = ',L
ACCEPT '# OF SPECIMENS = ',NS
NK=NS*L
N=NK+L
TYPE "ENTER REFERENCE FILENAME"
READ(11,1) (NAME(I),I=1,5)

1 FORMAT(5A2)
OPEN 0,NAME
READ(0) (Y(I),I=1,N)
CLOSE 0
SUM=0.
DO 2 J=1,L
  X1=Y(NK+J)*Y(NK+J)
  SUM=SUM+X1
2 DO 7 I=1,NS
  K=(I-1)*L
  DO 6 J=1,L
  X(J)=Y(K+J)
  SUM1=0.
  SUM2=0.
  DO 3 J=1,L
  SUM1=SUM1+X(J)*Y(NK+J)
  SUM2=SUM2+X(J)**2
  PP=SUM2-SUM

3 DNR=SORT(PP)
Q(I)=SUM1/DNR
CONTINUE

TYPE "ENTER OUTPUT FILENAME"
READ(11,1) (NAME(I),I=1,5)
OPEN 0,NAME
WRITE(0) (Q(I),I=1,NS)
CLOSE 0
ACCEPT 'WISH TO CONTINUE? YES-1, NO-0 ',IWISH
IF(IWISH.EQ.1) GO TO 8
END
SUBROUTINE SPLIT(IY,N)

C PLOT ROUTINE FOR GRAPHICS TERMINAL
C THE ARRAY IY CONTAINS DATA FOR DISPLAY
C # OF POINTS FOR DISPLAY

DIMENSION IY(N)
INTEGER GS,CR
DATA GS,CR/29,13/
OPEN 4, "STTO1"
WRITE(4) GS
BIG=ABS(IY(1))
DO 4 J=2,N
IF(BIG.LT.ABS(IY(J)))BIG=ABS(IY(J))
CONTINUE
4 M=N
IF(M.GT.4096)M=4096
L=M/4
DO 55 IK=1,4
K=(IK-1)*L
DO 5 J=1,L
IXPOS=(J-1.)*1024.*L/0.5
Z=4.-IK
IYPOS=IY(J+K)/BIG*100.*200.*Z+100.5
ITEMP=IXPOS/32
LOX=IXPOS-ITEMP*32+64
IHIXPOS=IXPOS/32+32
ITEMP=IYPOS/32
LOY=IYPOS-ITEMP*32+96
IHIXYPOS=IYPOS/32+32
DO 2 I=1,2
WRITE(4) IHYPOS,LOY,IHIXPOS,LOX
2 CONTINUE
5 CONTINUE
55 WRITE(4) GS
CONTINUE
WRITE(4) CR
CLOSE 4
RETURN
END
DIMENSION IBUF1(14336), NAME(5), IBUF2(2048)

ANALOG TO DIGITAL CONVERSION
THE ARRAY IBUF1 HOLDS THE SAMPLED DIGITS
THE ARRAY IBUF2 IS FOR TEMPORARY STORAGE

ACCEPT "SAMPLING RATE (IN MICROSECONDS) ? ", ISAMP
1 ACCEPT 
"# OF INPUT SAMPLES ? ", NBUF
NBUF=NBUF
ACCEPT "FIRST CHANNEL ? ", ICHAN1
ACCEPT "LAST CHANNEL ? " ,ICHAN2
ICHAN=ICHAN1*4000K+ICHAN2*1024+ICHAN1*64
TYPE "TO START SAMPLING, STRIKE ANY KEY"
CALL ADTEST(ISAMP, IBUF1, NBUF, 121000K, ICHAN)
IF(-NBUF.GT.2048) NBUF=-2048
CALL SPLOT(IBUF1, -NBUF)
ACCEPT "WISH TO PLOT REMAINING DATA ? YES-1, NO-0 ", IWISH1
IF(IWISH1.EQ.0) GO TO 3
10 ACCEPT "BEGIN OF FILE = ", IB
DO 2 J=1,2048
2 IBUF2(J)=IBUF1(IB+J)
CALL SPLOT(IBUF2, -NBUF)
ACCEPT "WISH TO CONTINUE PLOTTING ? YES-1, NO-0 ", IWISH2
IF(IWISH2.EQ.0) GO TO 3
GO TO 10
3 ACCEPT "WANT TO STORE ? YES-1, NO-0 ", IWANT
IF(IWANT.EQ.0) GO TO 9
11 ACCEPT "BEGINNING OF FILE = ", IB
ACCEPT "LENGTH OF NEW FILE = ", LF
TYPE "ENTER OUTPUT FILENAME"
READ(11,8)(NAME(I), I=1,5)
8 FORMAT(5A2)
OPEN 1, NAME
WRITE(1)(IBUF1(IB+J), J=1,LF)
CLOSE 1
ACCEPT "FINISH STORING ? YES-1, NO-0 ", IWISH
IF(IWISH.EQ.0) GO TO 11
9 ACCEPT "WISH TO CONTINUE SAMPLING ? YES-1, NO-0 ", IWISH
IF(IWISH.EQ.1) GO TO 1
END
SUBROUTINE FOR ANALOG TO DIGITAL CONVERSION
ISAMP IS THE SAMPLING RATE
THE ARRAY IBUF1 CONTAINS THE SAMPLED DATA

*TITL ADTEST
*ENT ADTEST
*EXTD *SAVE *RETN
*ZREL

ADTEST: *ADTEST
*NREL

ADTEST: *ADTEST
JMP @SAVE
LDA 0,01,2
STA 0,ISAMP
LDA 0,2,2
STA 0,IBUF1
LDA 0,03,2
STA 0,NBUF
LDA 0,04,2
DOAC 0,21
LDA 0,05,2
DOA/0,21
SYSTEM
GCHAR
JMP +1
LDA 0,IBUF1
DOB 0,21
LDA 0,NBUF
DOC 0,21
LDA 0,ISAMP
DOAP 0,21
SKPDRN 21
JMP -1
JMP @RETN

ISAMP: 0
ITRIG: 0
ICHAN: 0
NBUF: 0
IBUF1: 0
*END
SUBROUTINE FFT(X,N,M,IS)
C SINGLE PRECISION VERSION OF BERGLAND FFT ALGORITHM
C X CONTAINS THE REAL VECTOR WHOSE FOURIER TRANSFORM
C IS REQUIRED
C THE NUMBER OF POINTS TO BE TRANSFORMED = 2**M
C
DIMENSION X(N)
IC=1
NT=N/2+1
IF(IS.EQ.-1)GO TO 2
X(1)=(X(1)+X(NT))/2.
X(NT)=X(1)-X(NT)
GO TO 9
2 NV2=N/2
DO 50 J=1,N
X(J)=X(J)/FLOAT(NV2)
50 CONTINUE
1 NV2=N/2
NM1=N-1
J=1
DO 17 I=1,NM1
IF(I.GE.J)GO TO 15
A=X(J)
X(J)=X(I)
X(I)=A
15 K=NV2
16 IF(K.GE.J)GO TO 17
J=J-K
K=K/2
GO TO 16
17 J=J+K
IF(IC.EQ.2)GO TO 7
9 MM1=M-1
PI=3.14159
NOVER2=N/2
NI=1
DO 20 IF=1,MM1
NI=2*NI
N2=NI
IF(IS.EQ.1)N2=N/NI
L=N2/2
NOP=NOVER2/N2
THETA=PI/N2
TB=SIN(THETA)
TA=SQRT(1.-TB*TB)
IC=1
IF(IS.EQ.-1)GO TO 65
IF(IS.EQ.1)IC=2
5 J2=L/2-1
IF(IS.EQ.-1)IC=3
IF(J2.LE.0)GOTO65
N1=N
160 DO 55 J=1,J2
N3=N1-J2-J
N5=J+J
N4=N3+N5
A=X(N3)
X(N3)=X(N4)
X(N4)=A
55 CONTINUE
N1=N1-N2
IF(N1.GE.N2)GO TO 160
65 IF(IC.EQ.3)GO TO 200
K0=0
DO 40 K=1,N,OP
C=1.*0
S=0.
K1=K0+1
K2=K1+L
K3=K2+L
K4=K3+L
X(K1)=X(K1)+X(K2)
X(K2)=X(K1)-2.*0*X(K2)
XI=-1.*0
IF(IS.EQ.1)XI=1.*0
T=XI*X(K4)
X(K3)=X(K3)-T
T=XI*X(K3)
X(K4)=2.*0*X(K4)+T
J1=L-1
IF(J1.EQ.0)GO TO 25
DO 30 J=1,J1
K1=K0+1+J
K2=K1+L
K3=K2+L
K4=K3+L
STEPM=TA*S+TB*C
CTEMP=TA*C-TB*S
S=STEPM
C=CTEMP
IF(IS.EQ.1)GO TO 4
AG=C*X(K3)
AC=S*X(K4)
G=AG+AC
AG=C*X(K4)
AC=S*X(K3)
H=AG-AC
\[ X(K4) = N - X(K2) \]
\[ X(K3) = X(K2) + H \]
\[ X(K2) = X(K1) - G \]
\[ X(K1) = X(K1) + G \]
GO TO 30

4 \[ G = X(K1) - X(K2) \]
\[ H = X(K3) + X(K4) \]
\[ X(K1) = X(K1) + X(K2) \]
\[ X(K2) = X(K3) - X(K4) \]
\[ AG = C \times G \]
\[ AC = S \times H \]
\[ X(K3) = AG - AC \]
\[ AG = S \times G \]
\[ AC = C \times H \]
\[ X(K4) = AG + AC \]
30 CONTINUE
25 K0 = K4
40 CONTINUE
IF(IS.EQ.-1 .AND. IC.EQ.1) GO TO 5
200 CONTINUE
20 CONTINUE
6 IF(IC.EQ.2) GO TO 1
\[ X(1) = X(1) + X(NT) \]
A = 2.0 \times X(NT)
\[ X(NT) = X(1) - A \]
7 RETURN
END
BIBLIOGRAPHY


Bibliography (Cont.)


Bibliography (Cont.)


VITA AUCTORIS

1946  Born in Ayacucho, Peru on July 18

1962  Graduated from High School in Ayacucho, Peru

1966  Graduated from National University of Education, Chosica, Peru, with a B.S. degree in Mathematics and Physics

1969  Graduated from Saint Louis University, Saint Louis, Missouri, with a B.S. degree in Mathematics

1971  Graduated from Saint Louis University, Saint Louis, Missouri, with a M.S. (R) degree in Mathematics

1976  Candidate for Ph.D. degree, Electrical Engineering Department, University of Windsor, Windsor, Ontario, Canada