On speech processing

Ying Yang

University of Nevada, Las Vegas

Follow this and additional works at: https://digitalscholarship.unlv.edu/rtds

Repository Citation
https://digitalscholarship.unlv.edu/rtds/1177

This Thesis is brought to you for free and open access by Digital Scholarship@UNLV. It has been accepted for inclusion in UNLV Retrospective Theses & Dissertations by an authorized administrator of Digital Scholarship@UNLV. For more information, please contact digitalscholarship@unlv.edu.
INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

Bell & Howell Information and Learning
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
800-521-0600

UMI®

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
ON SPEECH PROCESSING

by

Ying Yang

Bachelor of Science
University of Electronic Science & Technology of China, P.R.China
1993

A thesis submitted in partial fulfillment
of the requirements for the

Master of Science Degree
Department of Computer Science
Howard R. Hughes College of Engineering

Graduate College
University of Nevada, Las Vegas
August 2000
Thesis Approval
The Graduate College
University of Nevada, Las Vegas

July 7, 2000

The Thesis prepared by

YANG, YING

Entitled

ON SPEECH PROCESSING

is approved in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE IN COMPUTER SCIENCE

Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative
ABSTRACT

ON SPEECH PROCESSING

By

Ying Yang

Dr. Evangelos A. Yfantis. Examination Committee Chair
Professor of Computer Sciences
University of Nevada, Las Vegas

Speaking is one of the most important ways of communication between humans. Speech processing refers to the technology of speech signal transformations for more efficient storage and transmission, for enhanced intelligibility and ease of assimilation. According to the different purpose as above, technologies on speech signal processing falls into three main categories. They are speech coding, speech recognition and speech synthesis.

This thesis shows the experience of our approaching in this area. It focused on analyzing the different characteristics of speech by using mathematical and statistical methods. A test based on lag1 product of signals was taken for separating the voiced and unvoiced parts of the speech signals. Other tests for speech processing include pre-processing, collinear predicting, peak detected predicting as well as spectrum analysis.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>vi</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>vii</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>vii</td>
</tr>
<tr>
<td>CHAPTER 1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Overview</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Objective and Organization of the Thesis</td>
<td>3</td>
</tr>
<tr>
<td>CHAPTER 2 BACKGROUND</td>
<td>5</td>
</tr>
<tr>
<td>2.1 Speech Sound</td>
<td>5</td>
</tr>
<tr>
<td>2.2 Speech Coding</td>
<td>6</td>
</tr>
<tr>
<td>2.3 Speech Recognition</td>
<td>8</td>
</tr>
<tr>
<td>2.4 Speech Synthesis</td>
<td>10</td>
</tr>
<tr>
<td>CHAPTER 3 VOICE SIGNAL</td>
<td>11</td>
</tr>
<tr>
<td>3.1 Abstract</td>
<td>11</td>
</tr>
<tr>
<td>3.2 Introduction</td>
<td>11</td>
</tr>
<tr>
<td>3.3 Probability Distribution Function of the Normalized Lag1 Product</td>
<td>13</td>
</tr>
<tr>
<td>3.4 Interarrival Times of Zero-Crossings and Statistical Properties of Zero-Crossings</td>
<td>15</td>
</tr>
<tr>
<td>3.5 Probability Distribution of the Maximum and Minimum Amplitudes</td>
<td>15</td>
</tr>
<tr>
<td>3.6 Summary</td>
<td>17</td>
</tr>
<tr>
<td>CHAPTER 4 PRE-PROCESSING</td>
<td>23</td>
</tr>
<tr>
<td>4.1 Pre-emphasis</td>
<td>23</td>
</tr>
<tr>
<td>4.2 Endpoint Detection</td>
<td>25</td>
</tr>
<tr>
<td>4.3 Window Functions</td>
<td>29</td>
</tr>
<tr>
<td>CHAPTER 5 CO-LINEAR PREDICTING</td>
<td>31</td>
</tr>
<tr>
<td>5.1 Method</td>
<td>31</td>
</tr>
<tr>
<td>5.2 Data Analysis – Lossless Coding</td>
<td>33</td>
</tr>
<tr>
<td>5.3 Linear Parameters for Speech Recognition</td>
<td>40</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Fig.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Diagram of isolated-word recognition system</td>
<td>9</td>
</tr>
<tr>
<td>2.2</td>
<td>(a) Linear time alignment (b) Non-linear time alignment</td>
<td>10</td>
</tr>
<tr>
<td>3.1</td>
<td>Postfiltered noise block</td>
<td>18</td>
</tr>
<tr>
<td>3.2</td>
<td>Zero crossings detected</td>
<td>18</td>
</tr>
<tr>
<td>3.3</td>
<td>Histogram of interarrival time of zero crossings of noise block</td>
<td>19</td>
</tr>
<tr>
<td>3.4</td>
<td>Noise signal at LAG1</td>
<td>19</td>
</tr>
<tr>
<td>3.5</td>
<td>Histogram of noise signal at LAG1</td>
<td>20</td>
</tr>
<tr>
<td>3.6</td>
<td>Postfiltered signal block</td>
<td>20</td>
</tr>
<tr>
<td>3.7</td>
<td>Zero crossings detected</td>
<td>21</td>
</tr>
<tr>
<td>3.8</td>
<td>Histogram of interarrival time of zero crossings of signal block</td>
<td>21</td>
</tr>
<tr>
<td>3.9</td>
<td>Signal signal at LAG1</td>
<td>22</td>
</tr>
<tr>
<td>3.10</td>
<td>Histogram of signal signal at LAG1</td>
<td>22</td>
</tr>
<tr>
<td>4.1</td>
<td>The signals before pre-emphasis and after pre-emphasis</td>
<td>24</td>
</tr>
<tr>
<td>4.2</td>
<td>Energy function for a speech signal segment</td>
<td>26</td>
</tr>
<tr>
<td>4.3</td>
<td>Zero-crossing rate for a speech signal segment</td>
<td>27</td>
</tr>
<tr>
<td>4.4</td>
<td>Data range for a speech signal segment</td>
<td>28</td>
</tr>
<tr>
<td>4.5</td>
<td>Window functions</td>
<td>30</td>
</tr>
<tr>
<td>4.6</td>
<td>Relation of the nearest windows</td>
<td>30</td>
</tr>
<tr>
<td>5.1</td>
<td>Encoding and decoding process</td>
<td>33</td>
</tr>
<tr>
<td>5.2</td>
<td>Waveforms and histograms for co-linear method</td>
<td>35</td>
</tr>
<tr>
<td>5.3</td>
<td>Histogram($P=1.2.3.4$)</td>
<td>36</td>
</tr>
<tr>
<td>5.4</td>
<td>Histogram($P=5.6.7.8$)</td>
<td>37</td>
</tr>
<tr>
<td>5.5</td>
<td>Coefficients for different words</td>
<td>41</td>
</tr>
<tr>
<td>5.6</td>
<td>Coefficients for a word spoken by the same person at different time</td>
<td>42</td>
</tr>
<tr>
<td>6.1</td>
<td>Waveforms for peak detected method</td>
<td>46</td>
</tr>
<tr>
<td>6.2</td>
<td>Histogram</td>
<td>47</td>
</tr>
<tr>
<td>7.1</td>
<td>The word <em>bait</em> spoken by the same person at different time</td>
<td>50</td>
</tr>
<tr>
<td>7.2</td>
<td>The word <em>boat</em> spoken by the same person at different time</td>
<td>51</td>
</tr>
<tr>
<td>7.3</td>
<td>Word <em>bait</em>’s frequency spectrum</td>
<td>52</td>
</tr>
<tr>
<td>7.4</td>
<td>Word <em>boat</em>’s frequency spectrum</td>
<td>53</td>
</tr>
</tbody>
</table>
**LIST OF TABLES**

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Bandwidths and sampling rates for different audio quality</td>
<td>6</td>
</tr>
<tr>
<td>5.1</td>
<td>Entropy &amp; compression ratio</td>
<td>38</td>
</tr>
<tr>
<td>5.2</td>
<td>Coefficients</td>
<td>39</td>
</tr>
<tr>
<td>6.1</td>
<td>Peak detecting data segment</td>
<td>45</td>
</tr>
<tr>
<td>6.2</td>
<td>Compression without quantization (8kHz. 8bits. mono)</td>
<td>45</td>
</tr>
<tr>
<td>6.3</td>
<td>Compression after quantization (8kHz. 8bits. mono)</td>
<td>45</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

I would like to express my special thanks to Dr. Yfantis for his valuable advice, continuing encouragement and warmhearted guidance during the course of my graduate program in UNLV. I would like to thank all the members in our research group, Mr. A. Popovich, Mr. T. Lazarakis and Mr. A. Angelopoulos, for their contribution to this research work. I would also thank professors on my committee, Dr. Bein, Dr. Berghel and Dr. Wang, for their time and critical suggestions on this work. Thanks also go to Dr. Ogawa and Dr. Larmore for their advice.

It is also a great pleasure to thank all the professors for teaching me and all the staff of Department of Computer Science of UNLV for their enormous assistance. And thanks should go to the consultants of the UNLV writing center, they gave me a lot of help for correcting errors in this thesis.

Finally, Thanks must go to my family and friends for their love and support over years.
CHAPTER 1

INTRODUCTION

1.1 Overview

Speaking is one of the most important ways of communication between humans. Yet, there exist many obstacles, from the simple misunderstanding with a single dialect to the "untranslatable" slang between different languages and cultures. For the hard-of-hearing, the deaf, the speechless, and the deaf-mute, spoken language communication is impaired and impossible. But even for the sound listener, the screaming of an overhead plane, the noisy restaurant, the roar from a nearby highway, lawn mower or leaf blower can make normal speech perception difficult if not hopeless.

The large distances covered by modern trade and travel demand new means of rapid and reliable communication. The traveler on the road, in the air or on a train calls for new methods of mobile communication that offer maximum convenience and privacy without cluttering up scarce "air" space.

Many of the problems of spoken language communication, both ancient and those engendered by modern mores and technology, are amenable to amelioration by emerging strategies of speech processing: the transformations of speech signals for more efficient storage and transmission, for enhanced intelligibility and ease of assimilation [3]. These new departures in spoken language processing should be based on a thorough understanding of how humans speak, how they hear the sounds that impinge on their ears and how they absorb the intended message.
On the technical side, it is the emergence of fast algorithms, such as the Fast Fourier Transform [5], and the tiny transistor and its latter-day descendants, the integrated circuit and computer chip, that have made sophisticated signal processing a present reality. Modern computers and digital signal processors can do good things to speech that would have taken many bays of analog equipment not too long ago. The last few years have witnessed the transfer of speech technology from the research laboratory to the market place. It can be predicted that more new products will be available in the no longer future, such as digital hearing aids that not only amplify speech but suppress unwanted sounds: voice synthesizers that actually sound human: better book-reading aids for the blind: error-free automatic speech transcription (the voice-typewriter or “electronic” secretary): and reliable speaker verification for access to confidential data and limited resources. Speech processing is a topic that requires a wide range of knowledge, such as signal processing, electronics, computer science, linguistics, physiology, acoustic physics, mathematics, communication and information theory.

According to the purpose of different applications, technology on speech signal processing falls into three main areas. They are:

• Speech Coding
• Speech Recognition
• Speech Synthesis

Different areas use different techniques and algorithms. But at the same time, since they all depend on the acoustic mechanism and characteristics of speech signal, a lot of analysis or technique methods can be used in more than one area.

Another way of categorizing the speech systems is as either a store-playback system or as a real time speech system [8]. The constraints placed on each of them are
or as a real time speech system [8]. The constraints placed on each of them are considerably different and they are used in different applications. Applications of store-playback systems include voice-mail, answering machine, or applications where audio data is downloaded over the Internet to be played back after the data retrieval is complete. Real time systems include real time telephony and on-line broadcast. The objectives in a real time telephony based application typically are to devise low delay algorithms and schemes that are robust to transmission noise and usually it is needed to deal the datum frame by frame. In a store-playback type system, usually the delay is not a factor and there is no transmission error. However, the cost of hardware is a concern since each unit owned by the consumer needs to implement the decoding algorithm.

1.2 Objective and Organization of the Thesis

In this thesis, a lot of tests were taken in order to find the algorithms which are better than those we know for speech coding, recognition as well as speech synthesis. The importance, purpose and scope of these technologies are already discussed in the above section. Chapter 2 talks about the background concepts and methods which will be used in later chapters. The remaining chapters will go into more detail. Chapter 3 talks about one of the statistical properties of the speech signal. Chapter 4 (Pre-processing) discusses the methods usually applied before doing further processing, which include pre-emphasis, endpoints detection and window functions. Chapter 5 (Co-linear predicting) and chapter 6 (peak detected predicting) discuss the techniques based on the time domain analysis. In chapter 5, co-linear analysis is one of the most popular and efficient methods in speech processing. Several compression and recognition tests were based on it. Peak points of speech waveform include very
important information of a signal, such as its envelop (shape) and the length of the vocal tract. This is talked about in chapter 6. Chapter 7 (Spectrum Analysis) discusses the techniques based on the frequency domain analysis.

C. C++ and Mathlab are used in this thesis to calculate and analysis data. When it is needed to partial the signals frame by frame, a 10^-30ms speech signal segment is generally considered. because investigating shows the property of speech waveforms is more stable during this time interval. Most of the speech samples used in this thesis have the format of 8kHz-8bits-mono.
CHAPTER 2

BACKGROUND

This chapter gives a brief explanation of the properties of speech sound as well as three main kinds of speech processing techniques, which are speech coding, speech recognition and speech synthesis. Generally, all the three categories of speech processing techniques have two phases. The first phase is the same for all of these three processings, that is to find the parameters of a sound, which can be used to represent this specific speech signal (sound). Corresponding to speech code, speech recognition and speech synthesis, the second phase is to encode the speech signal, to compare the recognized pattern with those stored patterns and to produce the speech signal, respectively. The output parameters of the first phase are the input of the second phase.

2.1 Speech Sound

At the linguistic level, the unit of the speech sound is called phoneme. For different languages, the phonemes are different. For example, it has a set of around 40 phonemes in English, which can be divided into vowels, semi-vowels, consonants, etc. The same phoneme may sound differently, depending on those sounds surrounding it. The waveforms for vowel phonemes have significant characteristics than those of other phonemes. They play an important role in the speech signal processing. As far as the length of the vocal tract is concerned, which is the place that the sound
is generated, the frequency range of the speech is about 100-3000Hz. It is the major frequency range to be considered in this thesis.

2.2 Speech Coding

According to Nyquist's theorem, for digitizing the signal, the sampling rate is at least twice the bandwidth of the voice. This is necessary in order for the signal to be properly reconstructed from the digital samples. At here, sampling rate refers to the average number of samples selected per second. For example, the bandwidth of the telephony signal is about 4kHz, so the sampling rate of it is at least 8kHz (4kHz*2). The following Table 2.1 defines typical bandwidths and sampling rates for different audio quantities.

Table 2.1 Bandwidths and sampling rates for different audio quantity

<table>
<thead>
<tr>
<th>Audio Format</th>
<th>Sampling Rate</th>
<th>Frequency Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telephony</td>
<td>8 kHz</td>
<td>(200 - 3400 Hz)</td>
</tr>
<tr>
<td>Teleconferencing</td>
<td>16 kHz</td>
<td>(50 - 7000 Hz)</td>
</tr>
<tr>
<td>Compact Disk(CD)</td>
<td>44.1 kHz</td>
<td>(20 - 20000 Hz)</td>
</tr>
<tr>
<td>Digital Audio Tape(DAT)</td>
<td>48 kHz</td>
<td>(20 - 20000 Hz)</td>
</tr>
</tbody>
</table>

Bit rate is another terminology used very often. It is generally measured by bits per second or bits per sample. The number of bits per second is given by the production of sampling rate (measured in Hertz) and the average number of bits per sample of a certain coding method. For example, based on the sampling rates in the upper table, 8 bits per sample in the coding of telephone speech imply an overall bit rate of 64K bps (bits per second), and 16 bits per sample imply 128K bps. 16 bits per sample in the coding of each of two audio channels (left and right) in the CD-stereo format imply an overall rate of 1410K bps.

For the purpose of economic transmission or storage, such as holding large quan-
tities of speech on a voice mail system and transmitting the speech signal over a limited bandwidth mobile radio channel, it is important to reduce the size of digital speech representation, to have the required quantity and to maintain specified levels of processing delay and implementation complexity.

In this thesis, the file format used to record the sound is 8kHz-8bits-mono.

There are many methods for reducing the size of a file, such as run length coding and Huffman coding.

Run length encoding is one of the simplest data compression techniques, taking advantage of repetitive data. These repeating characters are called runs. Runs are represented by a count and the original data byte. For example, a source string of AAAABBBCCCCCCCCCCCDEEEEE could be represented with 4A5B8C1D4E.

Huffman codes are created by analyzing the data set and assigning shortest bit streams to the datum occurring most frequently. This coding method attempts to create codes that minimize the average number of bits per character.

Bit rate, signal quality, processing delay and the complexity of implementation are four fundamental parameters for measuring code-algorithms. To estimate the bit rate, the data can be analyzed by using mean, variance, entropy, which are defined as equation (2.1), (2.2), (2.3), and a histogram which is a figure used to show the occurring times of each values. Compression ratio is used to compare two different coding methods, which is defined as equation (2.4).

\[
\text{Mean} = \sum_{x=-\infty}^{\infty} xp(x) \quad (2.1)
\]

\[
\text{Variance} = \sum_{x=-\infty}^{\infty} (x - E(x))^2 P(x) \quad (2.2)
\]
\[ Entropy = - \sum_{x=-\infty}^{\infty} P(x) \log_2 P(x) \]  

\[ Compression_Ratio = \frac{Original\_file\_size}{compressed\_file\_size} \]

There are varieties of speech coding techniques being used currently. They belong to three different categories:

- Time domain coders, such as the most popular linear predictive coding (LPC)
- Frequency domain coders, such as sub-band coding and adaptive transform coding
- Model-based speech coders, such as sinusoidal transform coding, multi-band excitation coding and waveform interpolation coding.

### 2.3 Speech Recognition

The goal of research on speech recognition is to find valid methods which will accurately recognize normal human speech from any speaker. Those methods could be used in varieties of applications which include speech input to computers, office automation, security system, etc. But unfortunately to completely solve, the problem of automatic speech recognition has proved much more difficult. A number of speech recognition systems just perform reasonably well in some limited applications. The main obstacles in automatic speech recognition include speech by a different person, co-articulation in fast speech, dialect, noise, vocabulary, etc.

In order to implement practical automatic speech systems, the number of problems are minimized by allowing a limited vocabulary, reducing the number of speakers and requiring the user to speak each word from the vocabulary as an isolated entity. Depending on the minimization, recognition systems can be classified as speaker-
dependent or speaker-independent and isolated-word or connected-word. The system can be trained, that means it can continuously adapt the stored template to increase the recognition performance.

The diagram of recognition system is illustrated in figure 2.1. In the recognition system, the acoustic pattern or template of each word in the vocabulary is stored as a time sequence of features (frames), which are derived by using one or more of the parametric speech analysis techniques, such as autocorrelation function. FFT, cepstral analysis, etc. Recognition is performed by comparing the acoustic pattern of the word with the stored patterns and choosing the word which matches best with the recognized word.

![Diagram of isolated-word recognition system](image)

Figure 2.1 Diagram of isolated-word recognition system

The function of pattern matching in the diagram is to determine the similarity between the input word pattern and the stored word patterns. This involves not only computing the distance of amplitude for each sample, but also time-alignment of the input and reference patterns. A word spoken on different occasions, even by the same speaker, will exhibit variation in its time-scale. In order to properly match the internal parts of the patterns, the time-alignment can be linear as shown in figure 2.2(a) or can be non-linear which is the common case, as shown in figure 2.2(b). Dynamic Time Warping (DTW) and Hidden Markov Models are two of these kinds of pattern matching algorithms.
2.4 Speech Synthesis

Speech synthesis is the process of producing an acoustic signal by controlling a model of speech production with a set of parameters. If the model and parameters are sufficiently accurate, the production of intelligible synthetic speech is possible. There are two basic approaches in modeling the production process: one is called articulatory speech synthesis, and the other is called terminal analogue speech synthesis.
CHAPTER 3

VOICE SIGNAL

3.1 Abstract

Voice signals could be divided into the quiet periods and the period that voice is transmitted. The voice signals consist of the voiced segments, unvoiced segments, and transitional segments. A test is introduced here for separating the quiet periods from the voice signal. The same test also separates the voiced segments from the unvoiced segments, and from the transitional segments. This test should be administered in conjunction with tests introduced in the past for the same purpose. In this research the test is administered by itself and it has a high probability of correct classification and relatively low probability of misclassification. The probability of misclassification should be reduced if in borderline cases additional tests are used.

3.2 Introduction

Low frequency additive noise could create the appearance of a long period wave in the top of which the voice is added. So a voice signal which is quantized in a scale with mean zero, is being offseted by the noise to have mean on a low frequency sine-like wave which could be above zero for a relatively long period and then below zero for a relatively long period also. Thus appropriate high pass or band pass filters are necessary in order to filter out such environmental noise. Noise is a complex phenomenon with many sources. Thus the quiet periods of the filtered signal, although they have
mean zero, most of them do not have zero amplitude. Both the voiced and unvoiced part of the filtered voice signal are centered about zero. One of the attributes used for separating the quiet period from the voice signal is the frequency of zero crossings. Thus a cutoff point has been used, whereby if the number of zero crossings is greater than the cutoff point then the segment is characterized as noise otherwise as voice. In this research we investigate the probability distribution of the zero crossings during the quiet period, the probability distribution function of the zero crossings during the voiced period, the probability distribution function of the zero crossings during the unvoiced part of the signal, and finally the probability distribution function of the zero crossings during the transitional segment of the voice signal, where the transition is from voiced to unvoiced, and from unvoiced to voiced. We also consider the product of two consecutive numbers for segments of length $N$, where $N$ is greater than or equal to 200. The segments considered are segments of quiet period, segments of voiced part of the voice signal, segments of unvoiced part of the voice signal and segments of transitional part of the voice signal. We show that the probability distribution of the voiced signal is completely different than those of the quiet period and unvoiced signal, for both the zero crossings and the lag-1 product. However the unvoiced and quiet period probability distribution functions for number of zero crossings and quiet periods have a large overlap. Also the transitional period probability distribution functions of number of zero crossings, and lag1 products have a large overlap with the corresponding probability distribution functions of the voiced and unvoiced segments of the signal. The maximum likelihood estimate is used in this paper to decide which of segment the current segment is, based on the computed statistics. As we mentioned above the signal space is divided into strata. These strata are: the quiet segments, the voiced segments, the unvoiced segments, and the transition segments.
between voiced and unvoiced or unvoiced and voiced. A great deal of interest and
research has taken place in this area over the years. We cite some of it here, [9], [10],
[11], [12], [13].

3.3 Probability Distribution Function of the Normalized Lag1 Product
Let \(x_0, x_1, \ldots, x_{N-1}\) be \(N\) consecutive samples belonging to one of the four strata
considered in this paper. The expected value \(\mu = E(X_i) = 0\), \(i = 0, 1, 2, 3, \ldots, N - 1\).
hence an estimate of the variance is \(S^2 = \frac{\sum_{i=0}^{N-1} x_i^2}{N}\). Now consider the normalized lag1
products
\[
\begin{align*}
r_0 &= \frac{x_0 x_1}{S^2}, \\
r_1 &= \frac{x_1 x_2}{S^2}, \ldots, \\
r_{N-2} &= \frac{x_{N-2} x_{N-1}}{S^2}
\end{align*}
\]
If both \(x_i\) and \(x_{i-1}\) are both positive or both negative the product is positive, where
if one positive and the other is negative the product is negative. If \(R\) is a random
variable denoting the lag1 product. In totally uncorrelated data we have that:
\[
P(R > 0) = P(R < 0) = 0.5
\]
Where in correlated data where the conditional probability of \(x_i\) being positive given
\(x_{i-1}\) is positive is \(p_1\) and the conditional probability of \(x_i\) being negative given \(x_{i-1}\) is
negative is \(q_1\). Also if the probability for \(x_{i-1}\) to be positive is \(p\) and the probability
for \(x_{i-1}\) to be negative is \(q\), then the probability:
\[
P(R > 0) = pp_1 + qq_1
\]
and
\[
P(R < 0) = 1 - pp_1 + qq_1
\]
For \(p = q = 0.5\), \(p_1 = q_1 = 0.83\) then
\[
P(R > 0) = 0.83, P(R < 0) = 0.17
\]

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
The noise data are closer to being independent than the voice data. Therefore the histogram of the lag1 products for the noise data should be close to having as many negative numbers as many positive. Furthermore the normalized lag1 products of the noise data should have more numbers with large magnitude compared to the lag1 normalized products of the voiced data. This is because the values before the zero crossings of the voiced data are relatively small compared to the pick values of the voiced data, where the values right before the zero crossings of the voiced data are relatively large compared to the pick values of the noise data. In totally uncorrelated data the expected value of the normalized lag1 product is zero, because:

\[ E(R) = \int_{-\infty}^{\infty} rf(r)dr \]

since \( f(r) = f(-r) \) then

\[ E(R) = \int_{-\infty}^{0} rf(r)dr + \int_{0}^{\infty} rf(r)dr \]

or

\[ E(R) = 0 \]

In correlated data the probability distribution function of the normalized lag1 product is biased towards the positive values and therefore the expected value is positive. Also the variance of the normalized lag1 product is larger for totally uncorrelated data than for correlated data. The probability density function of the lag1 product for totally uncorrelated data is symmetric about zero, where the probability density function of correlated data has relatively few lag1 products being negative, and the majority of the lag1 products are positive. Figure 3.5 shows the histogram of the noise data for N=256 Figure 3.10 shows the histogram of the voiced data for N=256

It is clear that there is some overlap between the various distribution functions.
3.4 Interarrival Times of Zero-Crossings and Statistical Properties of Zero-Crossings

The interarrival time of zero-crossings is smaller for the noise and larger for the voiced part of the signal, for the transition periods, and the unvoiced part of the signal the interarrival times of the zero crossings are greater than or equal to the interarrival times of the noise and less than or equal to the interarrival times of voiced signal.

3.5 Probability Distribution of the Maximum and Minimum Amplitudes

Another indicator is the probability distribution of the maximum and minimum amplitudes. If

\[ X_0, X_1, X_2, \ldots, X_{N-1} \]

are random variables with the same probability distribution and

\[ U = \max\{X_0, X_1, X_2, \ldots, X_{N-1}\} \]

is a random variable equal to the maximum of these N random variables then the probability distribution function of U is given by:

\[ G(u) = P(U \leq u) = P(X_0 \leq u, X_1 \leq u, X_2 \leq u, \ldots, X_{N-1} \leq u) \]  \hspace{1cm} (3.1)

Due to independence of \( X_0, X_1, X_2, \ldots, X_{N-1} \) the above equation can be written as

\[ G(u) = P(X_0 \leq u)P(X_1 \leq u)\ldots P(X_{N-1} \leq u) \]  \hspace{1cm} (3.2)

or

\[ G(u) = [F(u)]^N \]  \hspace{1cm} (3.3)

where \( F(u) \) denotes the probability distribution function of the random variable \( X_i, i = 0, 1, 2, \ldots, N - 1 \). From the above we obtain that the probability density function
of the maximum is

$$g(u) = N f(u)[F(u)]^{N-1} \quad (3.4)$$

where $f(u)$ is the density function of the the random variable $X_i, \ i = 0, 1, 2, ..., N - 1$. Similarly the probability distribution function of the minimum denoted by

$$V = \min[X_0, X_1, X_2, ..., X_{N-1}]$$

is:

$$1 - H(v) = P(V > v) = P(X_0 > v, X_1 > v, ..., X_{N-1} > v) \quad (3.5)$$

Due to independence of $X_0, X_1, X_2, ..., X_{N-1}$ the above equation can be written as

$$1 - H(v) = P(X_0 > v)P(X_1 > v)...P(X_{N-1} > v) \quad (3.6)$$

or

$$H(v) = 1 - [1 - F(v)]^N \quad (3.7)$$

where $F(v)$ denotes the probability distribution function of the random variable $X_i, \ i = 0, 1, 2, ..., N - 1$ at the value $v$. and $f(v)$ denotes the corresponding density function. Thus the probability density function of the minimum is

$$h(v) = N f(v)[1 - F(v)]^{N-1} \quad (3.8)$$

The noise more so than the voice signal behaves like a random phenomenon. Furthermore the amplitude of the noise has different distribution function than that one of the signal. Therefore the above theory applies more to the noise than the signal. The more the departure of the signal is from the above theory the more likely
is that the signal is not noise. Where the closer the signal is in agreement with the above theory the more likely is to be noise.

After capturing the background noise at sampling rate of 8KHz. the block of 256 samples is extracted. This block is then filtered. Postfiltered noise signal is shown in figure 3.1.

3.6 Summary

A method for separating the quiet period from the voice signal was examined in this paper. The voice signals were divided into voiced segments, unvoiced segments, and transitional segments. Our test separates the quiet period from the voiced part of the voice signal. Preliminary results give us no misclassification error.
Figure 3.1 Postfiltered noise block

Figure 3.2 Zero crossings detected
Figure 3.3 Histogram of interarrival time of zero crossings of noise block

Figure 3.4 Noise signal at LAG1
Figure 3.5 Histogram of noise signal at LAG1

Figure 3.6 Postfiltered signal block

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 3.7 Zero crossings detected

Figure 3.8 Histogram of interarrival time of zero crossings of signal block
Figure 3.9 Signal signal at LAG1

Figure 3.10 Histogram of signal signal at LAG1
CHAPTER 4

PRE-PROCESSING

4.1 Pre-emphasis

In the spectrum of voiced speech, there is an overall -6dB/octave trend, as frequency increases. It is a combination of a -12dB/octave trend due to the voiced excitation source and +6dB/octave trend due to radiation from the mouth, see [6]. Therefore, it is desirable to compensate for the -6dB/octave roll-off by pre-processing the speech signal to give a +6dB/octave lift in the appropriate range so that the measured spectrum has a similar dynamic range across the entire frequency band. This is referred to as pre-emphasis.

Pre-emphasis can be implemented as a first-order, high-pass analogue filter or as a digital high-pass filter which processes the digitized speech signal. The equation is:

\[ y[n] = s[n] - \alpha \cdot s[n - 1] \]  

(4.1)

In the equation, \( y[n] \) is the output samples of the pre-emphasis filter, \( s[n] \) is the original input speech samples, and \( \alpha \) is a constant value (usually using \( \alpha = 15/16 = 0.937 \)). See figure 4.1 for comparing the difference of the signals before pre-emphasis and after pre-emphasis. It is clear to see that the pre-emphasis can average the noise backgrounds and the signal spectrum.

Although it is not needed to apply the pre-emphasis to the unvoiced speech. For simplicity, pre-emphasis is normally applied to the unvoiced part as well. In this thesis, by default, the pre-emphasis is the first step before doing further processing.
Figure 4.1 The signals before pre-emphasis and after pre-emphasis
4.2 Endpoint Detection

The second step in speech processing is to detect the beginning and ending of a speech utterance. This is often called endpoints detection. Endpoints are not easy to find accurately, especially in a noisy environment.

The existing algorithms used to detect the endpoints mostly are based on measurement of the signal short-time energy (See figure 4.2) and the zero-crossing rate (See figure 4.3). The short-time energy function used in figure 4.2 is:

\[ e[n] = \sum_{m=-128}^{128} |s[n]|W[n + m] \]  

(4.2)

In this equation, \( W[n] \) is a rectangle window function which means it is 1 inside the range. \( n=-128,-127,\ldots,127,128 \), and 0 outside the range. There are many other window functions and they will be talked about later. Zero-crossing rate is a measure of the number of times in a given time interval (frame) that the amplitude of a speech signal passes through the zero value.

Endpoints occur at the points where the signal crosses a threshold which can be decided by doing the same processing to the background noise samples prior to the speech signal.

It is also easy to notice that the data range is very different for unvoiced and voiced speech. Since voiced speech has larger amplitude than the unvoiced speech, the value range is larger than the unvoiced speech. Figure 4.4 was calculated by subtracting the minimum value from the maximum value in each frame.

It should be mentioned here that these three methods are linear real-time algorithms.
Figure 4.2 Energy function for a speech signal segment
Figure 4.3 Zero-crossing rate for a speech signal segment
Figure 4.4 Data range for a speech signal segment
4.3 Window Functions

Last chapter discussed the need to work with frames or short-terms of the signal because the speech signal is not stable along the time space. The way to strip the signals frame by frame is to pass them through a window function \( w(n) \). The most commonly used window functions are a rectangular window defined as equation (4.3) and the Hamming window defined as equation (4.4)

\[
\begin{align*}
    w(n) &= \begin{cases} 
        1, & n = 0, 1, \ldots, N - 1 \\
        0, & \text{otherwise}
    \end{cases} \\
    \text{(4.3)}
\end{align*}
\]

\[
\begin{align*}
    w(n) &= \begin{cases} 
        0.54 - 0.46 \times \cos(2\pi n/N - 1), & n = 0, 1, \ldots, N - 1 \\
        0, & \text{otherwise}
    \end{cases} \\
    \text{(4.4)}
\end{align*}
\]

In these equations, \( N \) is the size of the window which is another factor needed to be considered. Generally saying, a long window gives good frequency-domain resolution but poor time-domain resolution, and a short window gives a good time-domain resolution but poor frequency-domain resolution. In speech signals, a window of duration 20ms is generally selected. Since the speech signal is dynamically changed along the time, the 10-20ms is the duration in which the signal is stable.

The rectangular window function is used in time domain analysis, such as in the following short-term energy function and co-linear parameter function in chapter 4. The Hamming window is mostly used in the frequency domain analysis in order to filter the undesired frequencies generated by cutting off the signals at the beginning and ending of the window.

There are several ways to select the successive windows. The nearest two windows can not be overlapped or can be overlapped with each other within a certain time duration, as showing in figure 4.6
Figure 4.5 Window functions

(a) Rectangular window

(b) Hamming window

Figure 4.6 Relation of the nearest windows

No overlap

Overlap 2:1
CHAPTER 5

CO-LINEAR PREDICTING

5.1 Method

This section discusses speech coding & recognition by investigating the correlation characteristics of speech signals. The speech file format of 8kHz, 8 bits, mono is used.

Since the speech signal has a very strong correlation with its adjacent samples, the difference between the neighbor samples will be small. Suppose the predicted sample of the speech $s[n]$ is $s[n]'$ which is given by the equation:

$$s[n]' = \alpha_1 \cdot s[n - 1] + \alpha_2 \cdot s[n - 2] + \ldots + \alpha_p \cdot s[n - p] = \sum_{k=1}^{p} \alpha_k \cdot s[n - k] \quad (5.1)$$

where $p > 0$ and $\alpha_1, \alpha_2, \ldots, \alpha_p$ are the predictor coefficients. The prediction error, or difference, is given by

$$e[n] = s[n] - s[n]' \quad (5.2)$$

By minimizing the mean squared error, $E$, between the actual speech samples and the predicted ones, the predictor coefficients can be determined by solving a set of linear equations. The following are these equations:

$$E = \sum_{n} e^2[n] = \sum_{n} (s[n] - s[n]')^2 = \sum_{n} (s[n] - \sum_{k=1}^{p} \alpha_k \cdot s[n - k])^2 \quad (5.3)$$
\[
\frac{\partial E}{\partial \alpha_j} = -2 \sum_n s[n-j] \cdot (s[n] - \sum_{k=1}^p \alpha_k s[n-k]) = 0
\] (5.4)

So

\[
\sum_{k=1}^p \alpha_k \sum_n (s[n-j] \cdot s[n-k]) = \sum_n s[n] \cdot s[n-j], \quad j = 1, 2, ..., p
\] (5.5)

The above equation represents a set of \( p \) linear equations with the \( p \) unknown \( \alpha_k \).

Therefore, it should be possible to find a solution by matrix inversion. There exists an algorithm called Durbin's method to solve this equation. Let \( R(k) \) be defined as the autocorrelation value for a shift of \( k \) samples, that is

\[
R(k) = \sum_{n=-\infty}^{\infty} s[n] \cdot s[n+k]
\] (5.6)

This systems of equations can be written as

\[
\begin{bmatrix}
R(0) & R(1) & ... & R(p-1) \\
R(1) & R(0) & ... & R(p-2) \\
... & ... & ... & ... \\
R(p-1) & R(p-2) & ... & R(0)
\end{bmatrix}
\begin{bmatrix}
\alpha_0 \\
\alpha_1 \\
\alpha_2 \\
\vdots \\
\alpha_p
\end{bmatrix}
= \begin{bmatrix}
R(1) \\
R(2) \\
\vdots \\
R(p)
\end{bmatrix}
\] (5.7)

In Durbin's method, start with the autocorrelation coefficients \( R(i), i=0, ..., p \) and compute recursively the coefficients \( \alpha_i \), from the following equations:

\[
E(0) = R(0)
\] (5.8)

\[
K_i = \frac{R(i) + \alpha_1^{(i-1)} \cdot R(i-1) + \ldots + \alpha_{i-1}^{(i-1)} \cdot R(1)}{E(i-1)} = \frac{R(i) + \sum_{j=1}^{i-1} \alpha_1^{(i-1)} \cdot R(i-j)}{E(i-1)}.
\] (5.9)

where \( i = 1, 2, ..., p \)

\[
\alpha_i^{(i)} = K_i
\] (5.10)
\[ \alpha_j^{(i)} = \alpha_j^{(i-1)} - K_j \cdot \alpha_{i-j}^{(i-1)} \quad \text{where} \quad 1 \leq j \leq i - 1 \quad (5.11) \]

\[ E_i = (1 - K_i^2) \cdot E(i - 1) \quad (5.12) \]

The desired pth-order values, \( \alpha_j^{(p)} \), are the predictor coefficients:

\[ \alpha_j = \alpha_j^{(p)} \quad \text{where} \quad j = 1, 2, ..., p \quad (5.13) \]

### 5.2 Data Analysis – Lossless Coding

According to equation (5.2), it shows that the original speech samples can be reconstructed through the error and its predicted samples, as shown in Figure 5.1.

![Figure 5.1 Encoding and decoding process](image)

For a speech file containing word bait, Figure 5.2 shows its original waveform, the predicted waveform, the difference waveform and the histograms of the original and difference waveform when p=1.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
It is clear that there are more values nearing zero in the difference waveform than that of the original waveform. Figure 5.3 and 5.4 are the histograms of word bait's difference data when \( p = 1.2 \ldots 8 \).

A histogram is a bar chart which shows the distribution of different values. The more values are the same, the higher compression ratio which could be achieved later by using some coding methods. Entropy can also be calculated to to estimate the number of bits for per samples that are needed for encoding the data.

Other words' predicted coefficients (\( p = 1.2 \ldots 12 \)), entropies and their estimated compression ratios are showing in table 5.1 and table 5.2.

It can be noticed that the histogram becomes more narrow along the increasing of \( p \)'s value. The entropy also shows the same property that along the increasing of \( p \)'s value, the difference's entropy becomes smaller. It means less bits are needed to encoding the values of the signals. The average number of bits for coding each values can be estimated by casting the entropy value from float to integer. By using Huffman Coding, it can be achieved \( 2^4 \) bits reducing for per sample than the original number of bits for per sample and the compression ratio is about \( 1^3 \). Since this method for compression is lossless, the quantity of the voice will remain the same.

In the previous method, the coefficients were calculated for the whole voice signal. It can be used in a store-playback system. In a real time system, the speech signals should be stripped frame by frame by passing the signals through a rectangle window function.
Figure 5.2 Waveforms and histograms for co-linear method
Figure 5.3 Histogram(P=1,2,3,4)
Figure 5.4 Histogram (P=5,6,7,8)
Table 5.1 Entropy & compression ratio

<table>
<thead>
<tr>
<th>p</th>
<th>Difference_Entropy</th>
<th>C_Ratio</th>
<th>bat</th>
<th>Difference_Entropy</th>
<th>C_Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>bait</td>
<td>5.6405</td>
<td>1.33</td>
<td>bat</td>
<td>4.9197</td>
</tr>
<tr>
<td>2</td>
<td>bait</td>
<td>5.6309</td>
<td>1.33</td>
<td>bat</td>
<td>4.8507</td>
</tr>
<tr>
<td>3</td>
<td>bait</td>
<td>5.6276</td>
<td>1.33</td>
<td>bat</td>
<td>4.8325</td>
</tr>
<tr>
<td>4</td>
<td>bat</td>
<td>5.0969</td>
<td>1.33</td>
<td>bat</td>
<td>4.7284</td>
</tr>
<tr>
<td>5</td>
<td>bat</td>
<td>5.0725</td>
<td>1.33</td>
<td>bat</td>
<td>4.5841</td>
</tr>
<tr>
<td>6</td>
<td>bat</td>
<td>5.0677</td>
<td>1.33</td>
<td>bat</td>
<td>4.7776</td>
</tr>
<tr>
<td>7</td>
<td>bat</td>
<td>4.8422</td>
<td>1.60</td>
<td>bat</td>
<td>4.2837</td>
</tr>
<tr>
<td>8</td>
<td>bat</td>
<td>4.7552</td>
<td>1.60</td>
<td>bat</td>
<td>4.2674</td>
</tr>
<tr>
<td>9</td>
<td>bat</td>
<td>4.7588</td>
<td>1.60</td>
<td>bat</td>
<td>4.2547</td>
</tr>
<tr>
<td>10</td>
<td>bat</td>
<td>4.7147</td>
<td>1.60</td>
<td>bat</td>
<td>4.0338</td>
</tr>
<tr>
<td>11</td>
<td>bat</td>
<td>4.6908</td>
<td>1.60</td>
<td>bat</td>
<td>4.0313</td>
</tr>
<tr>
<td>12</td>
<td>bat</td>
<td>4.6137</td>
<td>1.60</td>
<td>bat</td>
<td>4.0314</td>
</tr>
<tr>
<td>1</td>
<td>beet</td>
<td>3.5988</td>
<td>2.00</td>
<td>boat</td>
<td>3.8244</td>
</tr>
<tr>
<td>2</td>
<td>beet</td>
<td>3.571</td>
<td>2.00</td>
<td>boat</td>
<td>3.314</td>
</tr>
<tr>
<td>3</td>
<td>beet</td>
<td>3.4874</td>
<td>2.00</td>
<td>boat</td>
<td>3.3163</td>
</tr>
<tr>
<td>4</td>
<td>beet</td>
<td>3.2341</td>
<td>2.00</td>
<td>boat</td>
<td>3.3425</td>
</tr>
<tr>
<td>5</td>
<td>beet</td>
<td>3.0001</td>
<td>2.00</td>
<td>boat</td>
<td>3.1771</td>
</tr>
<tr>
<td>6</td>
<td>beet</td>
<td>2.9138</td>
<td>2.67</td>
<td>boat</td>
<td>3.1615</td>
</tr>
<tr>
<td>7</td>
<td>beet</td>
<td>2.7761</td>
<td>2.67</td>
<td>boat</td>
<td>3.0499</td>
</tr>
<tr>
<td>8</td>
<td>beet</td>
<td>2.7013</td>
<td>2.67</td>
<td>boat</td>
<td>3.0648</td>
</tr>
<tr>
<td>9</td>
<td>beet</td>
<td>2.6722</td>
<td>2.67</td>
<td>boat</td>
<td>2.9356</td>
</tr>
<tr>
<td>10</td>
<td>beet</td>
<td>2.6409</td>
<td>2.67</td>
<td>boat</td>
<td>2.9368</td>
</tr>
<tr>
<td>11</td>
<td>beet</td>
<td>2.6368</td>
<td>2.67</td>
<td>boat</td>
<td>2.9497</td>
</tr>
<tr>
<td>12</td>
<td>beet</td>
<td>2.6101</td>
<td>2.67</td>
<td>boat</td>
<td>2.9442</td>
</tr>
</tbody>
</table>
Table 5.2 Coefficients

<table>
<thead>
<tr>
<th></th>
<th>p</th>
<th>a1</th>
<th>a2</th>
<th>a3</th>
<th>a4</th>
<th>a5</th>
<th>a6</th>
<th>a7</th>
<th>a8</th>
<th>a9</th>
<th>a10</th>
<th>a11</th>
<th>a12</th>
</tr>
</thead>
<tbody>
<tr>
<td>bat 1</td>
<td>0.9878</td>
<td>0.9781</td>
<td>0.9373</td>
<td>1.1078</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9995</td>
<td>0.9105</td>
<td>0.8526</td>
<td>0.9287</td>
<td>0.8075</td>
<td>1.0742</td>
<td></td>
</tr>
<tr>
<td>bat 2</td>
<td>0.9781</td>
<td>0.9373</td>
<td>0.9321</td>
<td>0.9995</td>
<td>0.9105</td>
<td>0.8526</td>
<td>0.9287</td>
<td>0.8075</td>
<td>1.0742</td>
<td>0.9878</td>
<td>0.9781</td>
<td>0.9373</td>
<td></td>
</tr>
<tr>
<td>bat 3</td>
<td>0.9373</td>
<td>1.1078</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9995</td>
<td>0.9105</td>
<td>0.8526</td>
<td>0.9287</td>
<td>0.8075</td>
<td>1.0742</td>
<td>0.9781</td>
<td>0.9373</td>
<td></td>
</tr>
<tr>
<td>bat 4</td>
<td>1.1078</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9995</td>
<td>0.9105</td>
<td>0.8526</td>
<td>0.9287</td>
<td>0.8075</td>
<td>1.0742</td>
<td>0.9781</td>
<td>0.9373</td>
<td>1.1078</td>
<td></td>
</tr>
<tr>
<td>bat 5</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9995</td>
<td>0.9105</td>
<td>0.8526</td>
<td>0.9287</td>
<td>0.8075</td>
<td>1.0742</td>
<td>0.9781</td>
<td>0.9373</td>
<td></td>
</tr>
<tr>
<td>bat 6</td>
<td>0.9321</td>
<td>0.9995</td>
<td>0.9321</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9995</td>
<td>0.9105</td>
<td>0.8526</td>
<td>0.9287</td>
<td>0.8075</td>
<td>1.0742</td>
<td>0.9781</td>
<td></td>
</tr>
<tr>
<td>bat 7</td>
<td>0.9995</td>
<td>0.8526</td>
<td>0.9995</td>
<td>0.9321</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9995</td>
<td>0.8526</td>
<td></td>
</tr>
<tr>
<td>bat 8</td>
<td>0.8526</td>
<td>1.0742</td>
<td>0.8526</td>
<td>0.9321</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.8526</td>
<td>1.0742</td>
<td></td>
</tr>
<tr>
<td>bat 9</td>
<td>1.0742</td>
<td>0.9878</td>
<td>1.0742</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9995</td>
<td>0.9105</td>
<td>0.8526</td>
<td>0.9287</td>
<td>0.8075</td>
<td>1.0742</td>
<td>0.9878</td>
<td></td>
</tr>
<tr>
<td>bat 10</td>
<td>0.9878</td>
<td>0.9781</td>
<td>0.9373</td>
<td>1.1078</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9995</td>
<td>0.9105</td>
<td>0.8526</td>
<td>0.9287</td>
<td>0.8075</td>
<td>1.0742</td>
<td></td>
</tr>
<tr>
<td>bat 11</td>
<td>0.9781</td>
<td>0.9373</td>
<td>1.1078</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9995</td>
<td>0.9105</td>
<td>0.8526</td>
<td>0.9287</td>
<td>0.8075</td>
<td>1.0742</td>
<td>0.9878</td>
<td></td>
</tr>
<tr>
<td>bat 12</td>
<td>0.9373</td>
<td>1.1078</td>
<td>0.9359</td>
<td>0.9321</td>
<td>0.9995</td>
<td>0.9105</td>
<td>0.8526</td>
<td>0.9287</td>
<td>0.8075</td>
<td>1.0742</td>
<td>0.9878</td>
<td>0.9781</td>
<td></td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
5.3 Linear Parameters for Speech Recognition

In speech recognition, the basic idea is to compare the recognized word with the words in the library and find the most similar couple. The whole-word samples can be stored in a word library, and the recognized word can be compared to these words in the library sample by sample. But even for the voice format which does not have a very high quantity, say 8kHz-8bits-mono, a small library of about thousands of words will need a lot of storage place. Despite the space needed to store the word library, it will need more time to find the similar word when it is needed to compare with so many numbers of values for each word. So in the speech recognition, it is important to find the characteristic values of each word. It will not just save the space for storing the speech library, but make it reality for recognizing speech on real time.

It is known that speech waveform is more stable during 10~30ms time interval. In our test, a frames has 256 samples and the two nearest frames overlapping 56 samples. The following two figures are the coefficients for different words (Figure 5.5) and for the same words spoken by the same person at different times (Figure 5.6) when p is 5. It is easy to see that for different words the coefficients are different and for the same word the coefficients are very similar. Furthermore, because the total number of samples for each frame is 256 and there are 56 samples that overlap the nearest two frames, 5 coefficients were used to represent the speech signal instead of 200 amplitude as values before. The number of values decreased 40 times compared with using the original data.
Figure 5.5 Coefficients for different words
Figure 5.6 Coefficients for a word spoken by the same person at different time
CHAPTER 6

PEAK DETECTED PREDICTING

6.1 Method

This section discusses peak detected coding. The same as before, the speech files format used in this test is 8kHz—8bits—mono.

It is easy to notice that the speech waveform is constructed by the vibration signals along the center zero. The values of the signals sometimes are positive, sometimes are negative. The shape of the waveform in a positive or negative region is like a triangle arc. So the idea come to detect the peak and zero positions in encoding signals and use them to predict the speech signals in decoding signals.

The zero crossing positions as well as the peak position and their values are recorded as showing in table (6.1). The residual values are recorded also, which are calculated by subtracting the predicting values from the original signal values.

6.2 Data Analysis

Table 6.1 is part of the data calculated from speech file bait.wav. The first column is the zero crossing positions except the first value. When this value equals to 1, it means the peak value is first on the positive side, then on the negative side, then on the positive side, and so on: otherwise, it is first on the negative side, then on the positive side, then on the negative side, and so on. The second column is the peak positions and the third column is the peak values. One byte could be used for storing
the zero crossing and peak position data, which is that the higher 4 bits of this one byte is stored the zero crossing position value and the lower 4 bits of the same byte is stored the peak position value. Another one byte is needed for storing the peak value.

As showing in figure 6.1, figure 6.1(a) is the original waveform, figure 6.1(b) is the peak predicted waveform, figure 6.1(c) is the difference waveform of comparing the original waveform with the peak waveform, and figure 6.1(d) is the quantized difference waveform that is the difference values in figure 6.1(c) divided by 8.

In the accompany disk, these speech files were saved under the directory /samples/peak/. The quantities can be compared by listening to them. The following is the list of these files:

Bait.1.wav: original sound file
Bait.2.wav: predicted sound file
Bait.3.wav: predicted data + difference data
Bait.4.wav: predicted data + quantified difference data

Figure 6.2 are the histograms of the speech files in figure 6.1(a), (c), (d), respectively.

The histogram shows that more compression ratio can be achieved by storing the quantified difference data. By listening to the speech, the speech quantities of Bait.3.wav and Bait.4.wav are almost the same.

Table 6.2 and table 6.3 show the analysis data by using this peak detected predicting method. They are different at storing the residual data without quantization or with quantization. The encoded file should store the information as those in table 6.1 as well as the residual data. It can be noticed that the higher compression rate can be achieved after doing quantization and the compression method is word dependent.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Table 6.1 Peak detecting data segment

<table>
<thead>
<tr>
<th>zero crossing</th>
<th>peak position</th>
<th>peak value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>24</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>17</td>
<td>121</td>
</tr>
<tr>
<td>16</td>
<td>7</td>
<td>120</td>
</tr>
<tr>
<td>31</td>
<td>18</td>
<td>121</td>
</tr>
<tr>
<td>34</td>
<td>13</td>
<td>126</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>33</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>43</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>16</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>16</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>33</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>...</td>
</tr>
</tbody>
</table>

Table 6.2 Compression without quantization (8kHz. 8bits. mono)

<table>
<thead>
<tr>
<th>sample</th>
<th>size(bytes)</th>
<th>peak(bytes)</th>
<th>entropy_D</th>
<th>predicted(bytes)</th>
<th>C_Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>bait</td>
<td>1727</td>
<td>1854</td>
<td>407</td>
<td>5.7615</td>
<td>1702.25</td>
</tr>
<tr>
<td>beet</td>
<td>2887</td>
<td>3014</td>
<td>429</td>
<td>3.8685</td>
<td>1872.50</td>
</tr>
<tr>
<td>bite</td>
<td>2965</td>
<td>3092</td>
<td>643</td>
<td>4.5066</td>
<td>2496.13</td>
</tr>
<tr>
<td>boat</td>
<td>2755</td>
<td>2882</td>
<td>429</td>
<td>4.3175</td>
<td>2150.88</td>
</tr>
<tr>
<td>butte</td>
<td>3004</td>
<td>3130</td>
<td>407</td>
<td>5.1855</td>
<td>2660.00</td>
</tr>
</tbody>
</table>

Table 6.3 Compression after quantization (8kHz. 8bits. mono)

<table>
<thead>
<tr>
<th>sample</th>
<th>size(bytes)</th>
<th>peak(bytes)</th>
<th>entropy_D</th>
<th>predicted(bytes)</th>
<th>C_Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>bait</td>
<td>1727</td>
<td>1854</td>
<td>407</td>
<td>2.4213</td>
<td>1054.63</td>
</tr>
<tr>
<td>beet</td>
<td>2887</td>
<td>3014</td>
<td>429</td>
<td>0.7029</td>
<td>789.88</td>
</tr>
<tr>
<td>bite</td>
<td>2965</td>
<td>3092</td>
<td>643</td>
<td>1.2733</td>
<td>1384.25</td>
</tr>
<tr>
<td>boat</td>
<td>2755</td>
<td>2882</td>
<td>429</td>
<td>0.8965</td>
<td>773.38</td>
</tr>
<tr>
<td>butte</td>
<td>3004</td>
<td>3130</td>
<td>407</td>
<td>1.7257</td>
<td>1158.00</td>
</tr>
</tbody>
</table>

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 6.1 Waveforms for peak detected method

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 6.2 Histogram
CHAPTER 7

SPECTRUM ANALYSIS

7.1 Method

The frequency domain is another important characteristic of speech signals, as talked about in the chapter 2. By looking at the waveform of a speech segment, it can be noticed that the voiced signal is clearly periodic with a fundamental frequency, but the unvoiced (noised) signal lacks of periodicity.

The Discrete Fourier Transform (DFT) can be used to change the discrete time domain signals to the frequency domain. The Inverse Discrete Fourier Transform (IDFT) can be used to change the signal back to the time domain. As showed in the following, equation 7.1 is the definition of the DFT and equation 7.2 is the definition of the IDFT.

\[ X(k\sigma) = \sum_{n=0}^{N-1} x[nT]e^{-j2\pi nk/N} \quad k = 0, 1, ..., N - 1 \]  \hspace{1cm} (7.1)

\[ x[nT] = \frac{1}{N} \sum_{n=0}^{N-1} X(k\sigma)e^{j2\pi nk/N} \quad k = 0, 1, ..., N - 1 \]  \hspace{1cm} (7.2)

In these equations, N is the size of an array of complex numbers. Fast Fourier Transform (FFT) is a method for calculating DFT more efficiently. The time complexity of FFT is \(O(N \log N)\), and N should be a factor of 2.
7.2 Data Analysis

The same word spoken at different times, even by the same person, could have different waveforms. For example, one could be louder than another. In order to ignore these kinds of factors, the signals were normalized after doing pre-emphasis and then changed to the frequency domain by using FFT. As showed in figures 7.1 and 7.2, they illustrate two words, “bait” and “boat”, on the time domain (after doing pre-emphasis and normalization) and on the frequency domain.

It can be noticed that different words have different frequencies, and speech segments of the same word have similar frequency ranges. This property of the speech signals could be used to recognize the word. But it does not work well for a large vocabulary, because the frequency ranges are not significantly different for different words. Furthermore, some important information is averaged after doing the FFT, such as the frequency characteristics of different phonemes for a word along the time domain.

The above problem can be solved by stripping speech signals frame by frame by using the Hamming window of 256 samples per frame, and overlapping the nearest two frames with 56 samples. The FFT was done on each of these frames. Therefore, the result is an N*M matrix for a whole word. One dimension of this matrix is the frequency of the speech signal, and the other dimension of this matrix is the index of each windows which represents the characteristic of the speech along the time. The following figures (7.3 and 7.4) are the frequency spectrums (log10 values) of the same words as above.

It is very clear that for the same word, the spectrums are very similar, although they could have time alignment. In the matching step of speech recognition, this problem could be solved by using other methods as has been mentioned before.

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 7.1 The word *bait* spoken by the same person at different time
Figure 7.2 The word **boat** spoken by the same person at different time
Figure 7.3 Word "bait"'s frequency spectrum

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 7.4 Word boat's frequency spectrum

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
CHAPTER 8

CONCLUSION

This thesis describes our research work on speech processing. It focuses on finding the methods that could be used either to detect the voice signal or to represent the isolated words. These methods are important for doing future processing, such as coding, speech recognition and synthesis.

Chapter 1 and 2 discussed the purpose and general concepts of this research.

Chapter 3 discussed a statistic property of voiced, unvoiced and transitional parts of speech signals, respectively. A test based on lag1 product of the signal was introduced. This result could be used to separate the quiet period from the voiced part of the speech signals. In this research the test is administered by itself and it has a high probability of correct classification and relatively low probability of misclassification. The probability of misclassification should be reduced if in borderline cases additional tests are used.

Chapter 4 discussed methods generally used before calculating the characteristic parameters of a speech segment. Pre-emphasis is used to average the environment noise and the signal spectrum. Endpoint detection is used to find the voiced utterance. Three methods were discussed. Since sometimes it is not very accurate by using one method, two or more methods could be used together to achieve more accurately. Window function is used to strip the speech signals frame by frame. Rectangular and Hamming window functions were discussed. They were used at different cases.
Such as the rectangular window function is used at calculating the short-time energy function and the Hamming window function is used before doing FFT.

Chapter 5 discussed the Durbin’s method used to calculate the collinear parameters of a speech signal. These collinear parameters can be used in speech coding as well as speech recognition. In the coding method, 2^4 bits reducing for per sample could be achieved and the compression ratio is about 1^~3 for a speech file which has the original format as 8bits-8kHz-mono. Since this method for compression is lossless, the quantity of the voice will remain the same. In the recognition method, the co-linear parameters (predictor coefficients) were used to represent a speech, and the number of values decreased 40 times compared with using the original data.

Chapter 6 discussed a predicting method based on peak information of the predicted speech signals. Peaks carry the important information of the shape of a signal. A compression method based on this property of speech signals was tested. The residual data could be stored without quantization or with quantization and the speech quantities are almost the same. Generally, the compression ratio is 1^~2 without quantization and 1^~4 with quantization. It can be noticed that this compression method is word dependent.

The methods talked in above chapters are all based on the time domain signals.

Chapter 7 discussed a method based on the frequency domain signals. The frequency spectrums of several words were calculated by using FFT. This method could be used at speech recognition. It can achieve higher accurately, but involves more computing time compared with other methods.
8.1 Future Research

In this thesis, a lot of tests were done for detecting the voice signals as well as finding the parameters of speech signals. As discussed above, this is the first step (phase) in the speech processing. For a complete system, the second step (phase) is very important too. Problems involved in the second step are also very complexity. For examples, in speech recognition, a statistical model should be constructed in order to solve time alignment of the signals patterns. Our future research will go to this direction. And also there exist other achievements for speech processing, such as some methods based on the technique named wavelet. We should keep up with these changes because knowing different methods could help us to expand our mind to better understand the characteristics of the speech signals. I believe that our knowledge of the speech should base on better understanding of the complex relationship between the linguistic characteristic of speech signal and the discrete signals processing. And I also believe that as long as we keep on doing this research, we can make more achievements in this fields.
REFERENCES


VITA

Graduate College
University of Nevada, Las Vegas

Ying Yang

Local Address:
1600 E. University AVE. #156
Las Vegas. NV 89119

Degrees:
Bachelor of Science. Computer Software. 1993
University of Electronic Science & Technology of China. P.R.China

Thesis Title:
On Speech Processing

Thesis Examination Committee:
Chairperson. Dr. Evangelos A. Yfantis. Ph.D.
Committee Member. Dr. Berghel. Ph.D.
Committee Member. Dr. Bein. Ph.D.
Graduate Faculty Representative. Dr. Wang. Ph.D.