Voice Authenticationa Study Of Polynomial Representation Of Speech Signals

2005

John Strange
University of Central Florida

Voice Authentication
A Study of Polynomial Representation of Speech Signals

By

JOHN E. STRANGE
B. Ch.E. Georgia Institute of Technology, 1976

A thesis submitted in partial fulfillment of the requirements
for the degree of Master of Science
in the Department of Mathematics
in the College of Arts and Sciences
at the University of Central Florida
Orlando, Florida

Spring Term
2005

ABSTRACT

A subset of speech recognition is the use of speech recognition techniques for voice authentication. Voice authentication is an alternative security application to the other biometric security measures such as the use of fingerprints or iris scans. Voice authentication has advantages over the other biometric measures in that it can be utilized remotely, via a device like a telephone. However, voice authentication has disadvantages in that the authentication system typically requires a large memory and processing time than do fingerprint or iris scanning systems. Also, voice authentication research has yet to provide an authentication system as reliable as the other biometric measures.

Most voice recognition systems use Hidden Markov Models (HMMs) as their basic probabilistic framework. Also, most voice recognition systems use a frame based approach to analyze the voice features. An example of research which has been shown to provide more accurate results is the use of a segment based model. The HMMs impose a requirement that each frame has conditional independence from the next. However, at a fixed frame rate, typically 10 ms., the adjacent feature vectors might span the same phonetic segment and often exhibit smooth dynamics and are highly correlated. The relationship between features of different phonetic segments is much weaker. Therefore, the segment based approach makes fewer conditional independence assumptions which are also violated to a lesser degree than for the frame based approach. Thus, the HMMs using segmental based approaches are more accurate.

The speech polynomials (feature vectors) used in the segmental model have been shown to be Chebychev polynomials. Use of the properties of these polynomials has made it possible to reduce the computation time for speech recognition systems. Also, representing the spoken word

waveform as a Chebychev polynomial allows for the recognition system to easily extract useful and repeatable features from the waveform allowing for a more accurate identification of the speaker.

This thesis describes the segmental approach to speech recognition and addresses in detail the use of Chebychev polynomials in the representation of spoken words, specifically in the area of speaker recognition. .

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1 : INTRODUCTION

Speech recognition is a field of study involving the accurate interpretation of spoken words. A number of applications include the transcription of spoken words into text, the synthesis of speech, and security. Voice authentication (alternately known as speaker recognition) is the task of identifying speaker based on his spoken words, is a branch of speech recognition applicable to the field of security, and is a biometric alternative to fingerprinting and iris scanning. Voice authentication has advantages over the other biometric measures in that it can be utilized remotely, via a telephone for example. However, voice authentication has disadvantages in that the authentication system typically requires much more memory and processing time than do fingerprint or iris scanning systems. Also, voice authentication research has yet to provide an authentication system as accurate as the other biometric measures.

The purpose of this thesis is to provide research on one aspect of speaker recognition, segmental modeling as opposed to a frame based modeling of the voice. Though this thesis will concentrate on a specific segmental probability approach to voice authentication, a brief introduction into the subjects of voice authentication and speech recognition, a brief overview of the basic mathematical probability principals of speech recognition, and an overview of signal processing in terms of feature extraction is provided in Chapter 2. Chapter 2 also provides a brief description of a frame based probabilistic model for voice authentication. The purposes of Chapter 2 are to give a novice reader an overview of the physical and mathematical complexities associated with speech recognition and of how a typical voice authentication system might work.

Chapter 3 provides an overview of orthogonal polynomials. Chapter 4 provides a detailed discussion of a segmental probabilistic model for speaker recognition. Chapter 5 shows that the polynomials representing features of the voiceprint utilized in the segmental model are in fact Chebychev polynomials. In showing that the polynomials are Chebychev polynomials, many useful characteristics of orthogonal polynomials can be used to create more efficient computational algorithms and thus overcome some of the disadvantages of voice authentication compared to other biometric authentication measures. One such characteristic of orthogonal polynomials is the spectral moments of speech polynomials described in Chapter 5.

CHAPTER 2 : VOICE AUTHENTICATION/SPEECH RECOGNITION OVERVIEW

The overall subject of speech recognition, though a seemingly simple concept, is actually quite complex and is deeply rooted in mathematics. The complexity arises from the almost infinite variety of tonal inflections, accentuations, pronunciations, and volume when comparing one speaker with another speaker. Also presenting problems are words that sound the same but have different spelling. Further complicating the process is the variety of equipment used to record the speech at the front end of converting the speech to text. Complexities added due to the equipment include noise introduced from imperfect microphones and differences in sampling rates when converting the inherently analog voice signal to a digital signal.

Early speech recognitions systems were based on discrete data with restricted syntax and small vocabularies. These systems relied on words being spoken slowly, with pauses between words. Also, with small vocabulary and restricted syntax, the system knew which words were legal and made the job of interpretation much easier. An example of this is a speech recognition project undertaken in the late 1960's by Raj Reddy at the Stanford University [Jelinek -1]. Reddy decided to develop a system to recognize spoken chess moves. The system he developed would look for the closest word match and compare to the legal syntax (legal chess move) and if inappropriate would reject that choice.

A voice authentication system, as with any speech recognition system, begins with equipment to capture and digitize the speakers' voice. The basic equipment is a microphone or telephone to input speech, an analog-to-digital converter, a computer, and a database to store voice characteristics. Typically, these systems match the features of a voice (harmonic and

resonant frequencies, as well as the way the speaker pronounces phonemes – a language's smallest distinctive sounds) against an authorized user's digital voiceprint. The voiceprint is created when an authorized user enrolls in the authentication system and is stored as a digital file in a database. The system calculates a score that indicates how closely the spoken voice matches the stored voiceprint for the person the speaker claims to be. The score is based on probabilities that the spoken word is that which is stored.

The basic probability of speech recognition utilizes a statistical model called the Hidden Markov Model (HMM) to calculate overall probability of matching speech. The HMM utilizes small segments of speech, called frames, with each segment having an associated probability density function. This probability model is discussed in the next section titled "Basic Mathematics of Speech Recognition".

Basic Mathematics of Speech Recognition

Because of these complexities in speech recognition, when considering continuous speech instead of discrete speech and when considering very large vocabularies necessary to interpret every-day conversational language, mathematical models form a very important research tool of the system developers.

The speech signal is inherently a non-stationary random process. A stationary random process is one in which the joint probability distribution function (pdf) of random variables or vectors at an instant of time t equals that at time $t + \Delta t$. However, for very short periods of time, the speech signal is stationary. Because in these short periods of time, commonly referred to as frames, the signal is stationary, the joint pdf is considered to remain unchanged. This allows the

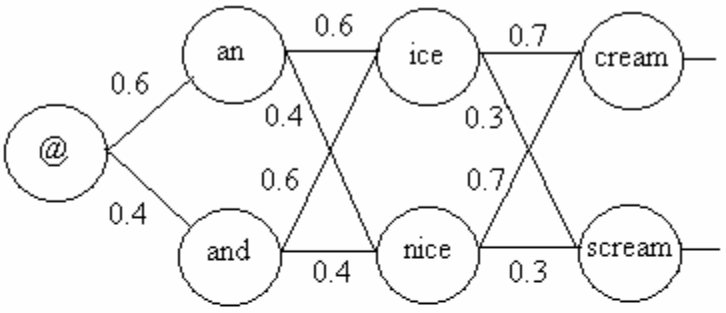
calculation of features such as autocorrelation, covariance, and mean amplitude or frequency of the samples belonging to the frame with the help of that pdf.

A given frame usually composes only a part of any given phoneme. Therefore, the frames need to be combined into phonemes. Later in this chapter and in the section entitled "Frame Based Probabilistic Model", we will describe the frame based probabilistic model for combining frames into phonemes and the combining phonemes into words. As discussed in that section, these frames are not statistically independent. However, frame based models have been developed for combining frames assuming independence. These models provide good results [Matsui & Furui, -14; Tseng, Soong & Rosenberg, .-15]. The section titled "Signal Processing and Analysis" describes an alternate segment based approach to combing frames with even better results.

Next, phonemes must be combined to form words and strings of words. In considering how the phonemes are combined to form words, note that speech recognition is complicated by the fact that even for a single speaker, no two utterances of the same word are the same. Particularly, one utterance may last longer than the other. In early speech recognition research, waveforms were "time warped", that is, the waveforms were modified in the time domain to closely match features of a given frame. It was then an easy exercise to compare features (sometimes in the form of polynomial coefficients) and construct words from the phonemes recognized. However, time warping was a very time consuming and a resource demanding procedure. Later, and most commonly used today, Hidden Markov Models (HMMs) were found to work well and alleviate the need to time warp the waveforms. [See Deller et.al,-2, for a complete discussion of time warping.] Deller et.al. have also discussed HMMs extensively.

In the HMM based speech recognition system, many different pathways through various frames of speech are analyzed with the highest probability path being the one taken to represent the spoken word. Although the next utterance of the same word may not exactly match a set of features, when considering a threshold probability, the word is considered to match even though not an exact match in terms of all features compared. In other words, each path may be represented by a combination of different feature vectors or polynomials containing features of some words with the HMM analyzing which word or path most likely represents the spoken word. For a more detailed description of the Hidden Markov Model for speech recognition one can refer to "Statistical Methods for Speech Recognition" by Frederick Jelinek [Jelinek -1].

.



Possible Probability Paths

an ice cream : 0.6 + 0.6 + 0.7 = 1.9
an ice scream: 0.6 + 0.6 + 0.3 = 1.5
an nice cream: 0.6 + 0.4 + 0.7 = 1.7
an nice scream: 0.6 + 0.4 + 0.3 = 1.3

and ice cream: 0.4 + 0.6 + 0.7 = 1.7
and ice scream: 0.4 + 0.6 + 0.3 = 1.3
and nice cream: 0.4 + 0.4 + 0.7 = 1.5
and nice scream: 0.4 + 0.4 + 0.3 = 1.1

Figure 2-1 Word Based Hidden Markov Model

To illustrate how a HMM works, consider Figure 2-1 [reproduced from Gazdar – 17]. This figure illustrates a word based HMM. Words are determined by a phoneme based HMM, similar to the word based HMM shown in Figure 2-1, only instead of analyzing path through all possible words, the paths analyzed are through all possible phonemes. Likewise, phonemes are determined by a frame based HMM where the paths analyzed are through all of the feature vectors extracted from the frames.

At the most basic level, the mathematical formulation of combining phonemes into words and strings of words can be described in the following statistical terms [see, Jelinek-1]:
Let *A* represent a sequence of symbols taken from some alphabet. Each symbol might represent a unique phoneme. This sequence represents the translation of the waveforms into some intermediate form by the translation device. In essence it is the acoustic evidence provided by the translation device. Let *W* represent a string of spoken words, each belonging to a known vocabulary. The speech recognition problem can then be stated as finding the maximum probability that the translated text, $\tilde{W}$, is the same as the spoken words *W* given the evidence *A*.

$$\tilde{W} = \max_{W} P(W \mid A) \tag{2-1}$$

Using Bayes' formula, the right side of the probability above can be rewritten as

$$P(W \mid A) = \frac{P(W)P(A \mid W)}{P(A)} \tag{2-2}$$

P(W) is the probability that the word string W will be spoken. $P(A|W)$ is the probability that when the speaker says W the acoustic evidence A will be observed. $P(A)$ is the average probability that A will be observed. Because the maximization is carried out with the variable A fixed (only the given acoustical data is considered), the aim of the translation device is to maximize the product of $P(W)P(A|W)$, or

$$\widetilde{W} = \max_{W} P(W)P(A|W) \qquad (2\text{-}3)$$

With the problem of speech recognition expressed in these general probability terms, the problem now reduces to the determination of the probabilities when analyzing a given voiceprint. Generally, the probabilities are obtained by experimentation. That is, a speech recognition system first has to be trained by having the speaker speak specific text with the input words stored in the vocabulary as feature vectors or polynomials. Then probabilities are assigned to each feature or polynomial. The probabilities are based on the principal that each speech segment has a statistical distribution of parameters and can be modeled by some distribution function.

Signal Processing and Analysis

As a preface to this section, a paper originally written in 1971 needs to be mentioned as a prerequisite for complete understanding of speech analysis. The paper titled "Analysis of Fundamental Frequency Contours" by H. Levitt and L. R. Rabiner [Levitt and Rabiner – 11]

presents much of the early work in the subject of speech analysis. The paper discusses the problem of variability in spoken words and presents solutions and test results based on the solutions. Generally, Levitt and Rabiner showed that the coefficients of polynomials obtained in consecutive time windows are useful in solving speech recognition problems. The paper also discusses time normalization, sliding window analysis, between window differences, and approximating families of frequency contours. This paper should be read for a detailed understanding of the subject, however, many of the concepts are generally captured in the following discussion.

All models of speech recognition require some form of digital representation of the input signal waveform. As there are many ways to address the overall speech recognition problem, there are many ways to address the problem of how best to convert the analog waveform input signal to a digital form. The purpose behind signal processing is to obtain a form of the input where certain characteristic features of spoken words is easily obtained. The "speech recognizer" must be able to compare the waveforms representing the spoken words to some library of words regardless of which model is used to process them.

How is a spoken word processed? First the spoken word is converted to an analog signal by a microphone. The microphone generates an electrical signal called a waveform. Then the waveform is filtered to reduce noise and converted to some digital form for comparison.

Filtering is necessary to remove extraneous noise which, if not removed, would hinder the speech recognition systems ability to accurately predict the speaker or his words. Speech waveforms and the noise associated with them are characteristically non-linear. One area of non-linear signal processing (filtering) is known as polynomial signal processing [Mathews et.al.-3]. Polynomial signal processing utilizes some of the same mathematical principals used in

speech recognition to determine what is noise and what is related to the word spoken. After filtering, the conversion to a digital signal involves sampling the analog waveform signal at specific and constant rate.

Signal analysis involves the extraction of features [Paulus et.al.-4] from the digital representation of the raw signal that may contain repeatable patterns every time the same word is spoken. There are two ways signals are processed for extraction of features. The waveform can be processed in the time domain or the signal can be converted by Fourier Transform to the frequency domain. Multiple features are extracted, in either the time or frequency domain, and are organized as feature vectors.

One method of signal analysis is to convert the feature vectors into polynomials. The polynomial coefficients contain information on the features of the waveform that is not as readily apparent in the original waveform. For example, for a polynomial created from frequency data, the zero degree polynomial gives the average frequency of the waveform, the first degree polynomial gives the average slope of the waveform, and the second degree polynomial gives the average quadratic curvature of the waveform. By comparing coefficients instead of the raw digitized data, more accurate conclusions can be made. Also, the comparison of coefficients of the polynomial is easier than comparison of raw data.

To obtain the feature vectors, the raw voice waveform is typically processed in frames of approximately 25 ms in length, advanced 10 ms at a time. Then, features are extracted from each frame. Features of the waveform include, but are not limited to, autocorrelation, covariance, the number of times the frequency crosses the zero axis, the average slope at the points of the zero crossings and the time indexed frequency of the spoken words [Deller et.al.-2]. Feature vectors are a set of features extracted from an individual window. This process can be looked upon as a

form of data compression. First, the analog signal is converted to thousands of digital data points. Then within 25 ms frames, all the digital data points are analyzed and a few features are extracted. For example, a small word may result in a raw count of 5,000 digital data points. This input may then be split into 5 windows with 3 features extracted per window for a total of 15 pieces of data requiring saving compared to the original 5,000 data points.

In summary, the idea behind good signal processing and signal analysis is to obtain the ability to extract useful information from the data increasing the probability that the word spoken was the same as what was translated or, in the case of speaker recognition, the word spoken was spoken by the individual determined by the system. As we know people have different accents, different inflections, different points of accentuations, etc. The current technology of speech recognition involves "training" where the speaker trains the computer by reciting predetermined text. This method is required for speaker identification but is useless in word transcription where it is desired for the transcription to be accurate independent of the speaker. The new research into recognition independent of speaker identity involves extraction of features that are common to all speakers' utterances of the same word. Thus signal processing and the subsequent signal analysis constitute a fundamental building block of speech recognition.

Frame Based Probabilistic Model

Liu and Wang [Liu & Wang, – 5] describe two ways that the input waveforms are processed. These two methods are the frame based method and the segment based method, both illustrated in Figure 2-2. For a frame based HMM system, the signal is processed in fixed intervals, typically about 25 ms, advanced by 10 ms at a time, and the observation probability is

obtained by a frame-based probability density function. Then a fixed number of these frames are combined and used for comparison purposes. The fixed size of the combined frames may or may not correspond to a complete phoneme. Therefore, several of these frame combinations may need to be combined again to capture a single phoneme. In the frame based approach, the statistics of each frame is assumed to be independent.

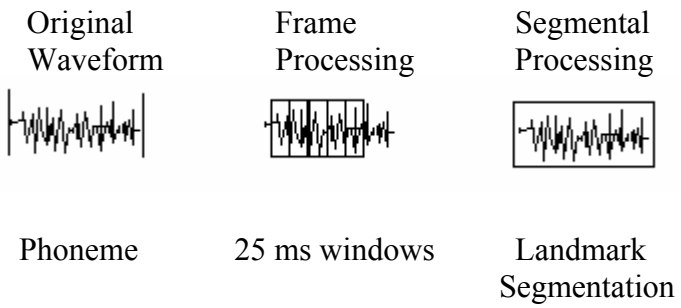


Figure 2-2 Illustration of Frame Based and Segmental Based Signal Processing
(The waveform illustrated does not represent any particular utterance.)

The Frame Based Probabilistic Model (FBPM) is parameterized by a finite mixture of Gaussian probability density functions (pdfs). For a given vector *x*, the likelihood for the FBPM is expressed by

$$P(x \mid \Lambda) = \sum_{m=1}^{M} c_m P(x \mid \mu_m, U_m) \tag{2-4}$$

where $P(x \mid \Lambda)$ is the likelihood for x, $P(x \mid \mu_m, U_m)$ is the m[th] pdf kernel, M is the total number of pdf kernels, μ_m is the mean vector, U_m is the covariance matrix and c_m is the mixture weighting factor.

In the training phase, a set of training feature vectors, $X = \{x(1)\ldots,x(t),\ldots,x(T)\}$ for a speaker is given to find the parameters of the FBPM, $\Lambda = \{ c_m, \mu_m, U_m \mid m = 1,\ldots,M\}$, such that the likelihood score $P(X \mid \Lambda)$ is a maximum. That is it is desired to find Λ such that

$$P(X \mid \Lambda) = \max_{\Lambda'} P(X \mid \Lambda') \qquad (2\text{-}5)$$

Equation 2-5 is difficult to solve. One possible algorithm useful in solving Equation 2-5 is the EM algorithm [Dempster, Laird and Rubin – 16]. This algorithm will not be discussed in this thesis.

CHAPTER 3 : ORTHOGONAL POLYNOMIALS

Many speech recognition models, included the Frame Based Probabilistic Model described in Chapter 2, use orthogonal polynomials in the extraction of speaker-dependent features from speech waves. The Liu-Wang Segmental Probabilistic Model, described later in Chapter 4, also uses orthogonal polynomials containing the features extracted from the speech waves. Orthogonal polynomials are used as a compression technique which allows for the independent compression of different aspects of the speech spectrum. Each polynomial corresponds to a different feature of the short-term speech spectrum, for example, the polynomials of the first and second degrees correspond to the average slope and quadratic curvature of the spectrum.

Because this thesis will be discussing in detail the Liu-Wang Segmental Probabilistic Model as an alternative to the Frame Based Probabilistic Model, and because this paper will then show that the polynomials found in the Liu-Wang model are in fact Chebychev polynomials, we shall first review some basic facts about orthogonal polynomials. Thus, the next two sections of this chapter will give a brief overview of orthogonal polynomials and some specifics of discrete orthogonal polynomials.

Orthogonal Polynomial Overview

Polynomials of order *n* are analytic functions that can be written in the form

$$p_n(x) = a_0 + a_1 x + a_2 x^2 ..... + a_n x^n \tag{3-1}$$

They can be differentiated and integrated for any value of x, and are fully determined by the $n+1$ coefficients a_i, $i = 0 \ldots n$. Polynomials are often used to approximate more complicated or unknown functions such as the function that represents a speech waveform. Normally, the order n_i is defined by the quality of the approximation desired.

Using polynomials as defined in Equation 3-1 tends to lead into numerical difficulties when determining the a_i, even for small values of n. Therefore, it is more practical to stabilize numerical results by using orthogonal polynomials over an interval [a,b]. With $W(x)$ defined as a weight function with $W(x) > 0$, orthogonal polynomials obey the orthogonality relationship

$$\int_a^b p_n(x)\, p_m(x) W(x) dx = 0 \quad for \quad n \neq m \tag{3-2}$$

Then, define a scalar product of two real functions *f* and *g* with respect to a weight function $W(x)$ by

$$\langle f, g \rangle = \int_a^b f(x) g(x) W(x) dx \tag{3-3}$$

From Equations 3-2 and 3-3, starting with a basis of {1, x, x^2, …} and using the Gram-Schmidt orthogonalization process leads to a set of orthogonal polynomials. Specially named polynomials result with specific weight functions $W(x)$ applied on the interval [-1,1]. If the weight function is given as $W(x) = 1$, then the resulting polynomials are Legendre polynomials. If $W(x) = (1-x^2)^{-1/2}$, then the resulting polynomials are Chebychev polynomials.

The Gram-Schmidt orthogonalization process for $W(x) = 1$ on the interval [-1,1] (yielding Legendre polynomials) is demonstrated as follows:

Let $x_0, x_1, x_2, ...x_n = 1, x, x^2, ...x^n$ and define $p_0(x) = x_0 = 1$, then

$$p_n(x) = x_n - \sum_{j=0}^{n-1} \frac{\langle x_n, p_j(x)\rangle p_j(x)}{\| p_j(x) \|^2} \tag{3-4}$$

With the inner product defined as

$$\langle x_n, p_n(x)\rangle = \int_{-1}^{1} x_n p_n(x) w(x) dx \tag{3-5}$$

and the norm defined by

$$\| p_n(x) \|^2 = \langle p_n(x), p_n(x)\rangle = \int_{-1}^{1} p_n^2(x) w(x) dx \tag{3-6}$$

With w(x)= 1 (and not showing the functions as functions of x)

$$p_1 = x_1 - \frac{\langle x_1, p_0 \rangle p_0}{\| p_0 \|} = x - \frac{\langle x,1 \rangle 1}{\| 1 \|} = x - \frac{(\int_{-1}^{1} x dx)}{\int_{-1}^{1} 1 dx} = x \tag{3-7}$$

$$p_2 = x_2 - \frac{\langle x_2, p_0 \rangle p_0}{\| p_0 \|} - \frac{\langle x_2, p_1 \rangle p_1}{\| p_1 \|} = x^2 - \frac{\langle x^2,1 \rangle 1}{\| 1 \|} - \frac{\langle x^2, x \rangle x}{\| x \|}$$

$$= x^2 - \frac{(\int_{-1}^{1} x^2 dx) 1}{\int_{-1}^{1} 1 dx} - \frac{(\int_{-1}^{1} x^3 dx) x}{\int_{-1}^{1} x^2 dx} = x^2 - \frac{1}{3} \tag{3-8}$$

Similarly, we can find

$$p_3 = x^3 - \frac{3}{5} x \tag{3-9}$$

From the above values of p_n, we can see that orthogonal polynomials of successive orders can be expressed by a recurrence relation. The three term recurrence relation for Legendre orthogonal polynomials is given by:

$$p_{n+1} = \frac{(2n+1)}{(n+1)} x p_n - \frac{n}{n+1} p_{n-1} \tag{3-10}$$

Specifics of Discrete Orthogonal Polynomials

Observe that the derivations of orthogonal polynomials shown above concern orthogonal polynomials of continuous variables. The Liu-Wang model concerns discrete variable speech polynomials. Therefore, we must provide a formulation of the orthogonality relationship and three-term recurrence relation for discrete orthogonal polynomials. Classical orthogonal polynomials of discrete variables are well developed in the text "Special Functions of Mathematical Physics" [Nikiforov et al – 12]. The following is a derivation of these relations using information from Nikiforov [12]:

Consider an orthogonal polynomial of the form

$$y_n(x) = a_n x^n + b_n x^{n-1} + ... \tag{3-11}$$

From page 113 of the Nikiforov [12], all discrete orthogonal polynomials $y_n(x)$ satisfy the following orthogonality relationship, up to a normalizing factor,

$$\sum_{x_i=a}^{b-1} y_n(x_i) y_m(x_i) \rho(x_i) = d_n^2 \delta_{nm} \qquad \text{on interval (a,b)} \tag{3-12}$$

where $\rho(x_i)$ is the discrete weight function with $\rho(x_i) > 0 \quad for \quad a \le x_i \le b-1$, d_n^2 is the squared norm, and δ_{nm} is the Kronecker symbol.

Then from page 128 of [12], the three-term recurrence relation is given by

$$xy_n(x) = \alpha_n y_{n+1}(x) + \beta_n y_n(x) + \gamma_n y_{n-1}(x), \qquad \text{or} \tag{3-13}$$

$$y_{n+1}(x) = \frac{(x-\beta_n)y_n(x) - \gamma_n y_{n-1}(x)}{\alpha_n} \tag{3-14}$$

$$\alpha_n = a_n / a_{n+1},\ \beta_n = b_n / a_n - b_{n+1} / a_{n+1}, \text{ and } \gamma_n = (a_{n-1} / a_n) d_n^2 / d_{n-1}^2. \tag{3-15}$$

Next, we will develop the coefficients α_n, β_n and γ_n for orthogonal Chebychev polynomials in the interval (0,N). On the interval (0,N), the Chebychev polynomial coefficients a_n and b_n and the squared norm d_n^2 are given in Table 3, page 129 of Nikiforov [12] as

$$a_n = \frac{1}{n!}(n+1)_n \tag{3-16}$$

$$b_n = -\frac{N-1}{(n-1)!}(n)_n \tag{3-17}$$

$$d_n^2 = \frac{(N+n)!}{(2n+1)(N-n-1)!} \tag{3-18}$$

where $(n)_n$ denotes the usual Pochhammer symbol defined as

$$(n)_n = n(n+1)(n+2)...(2n-2)(2n-1). \tag{3-19}$$

As can be seen in equation 3-15, we need to also have equations for a when n = n+1 and n = n-1, for b when $n = n+1$ and for d^2 when $n = n-1$. Therefore, the following additional equations will be needed to find values for α_n, β_n, and γ_n. These equations are derived from equations 3-16 through 3-19 by substituting $n = n+1$ and $n = n-1$ as appropriate.

$$a_{n+1} = \frac{1}{(n+1)!}(n+2)_{n+1} \tag{3-20}$$

$$a_{n-1} = \frac{1}{(n-1)!}(n)_{n-1} \tag{3-21}$$

$$b_{n+1} = -\frac{N-1}{(n)!}(n+1)_{n+1} \tag{3-22}$$

$$d^2_{n-1} = \frac{(N+n-1)!}{(2n-1)(N-n)!} \tag{3-23}$$

$$(n+1)_n = (n+1)(n+2)(n+3)...(2n-1)(2n) \tag{3-24}$$

$$(n+1)_{n+1} = (n+1)(n+2)(n+3)...(2n)(2n+1) \tag{3-25}$$

$$(n+2)_{n+1} = (n+2)(n+3)(n+4)...(2n+1)(2n+2) \tag{3-26}$$

$$(n)_{n-1} = n(n+1)(n+2)...(2n-3)(2n-2) \tag{3-27}$$

Now, finding values for α_n, β_n, and γ_n using equations 3-15 through 3-27:

$$\alpha_n = \frac{a_n}{a_{n+1}} = \frac{\frac{1}{n!}(n+1)_n}{\frac{1}{(n+1)!}(n+2)_{n+1}} = \frac{(n+1)![(n+1)(n+2)...(2n-1)(2n)]}{n![(n+2)(n+3)...(2n+1)(2n+2)]}$$

$$= \frac{(n+1)(n+1)}{(2n+1)(2n+2)} = \frac{(n+1)}{2(2n+1)} \tag{3-28}$$

$$\beta_n = \frac{b_n}{a_n} - \frac{b_{n+1}}{a_{n+1}} = \frac{-\frac{N-1}{(n-1)!}(n)_n}{\frac{1}{n!}(n+1)_n} - \frac{-\frac{N-1}{(n)!}(n+1)_{n+1}}{\frac{1}{(n+1)!}(n+2)_{n+1}}$$

$$= -\frac{n!(N-1)[n(n+1)...(2n-2)(2n-1)]}{(n-1)![(n+1)(n+2)...(2n-1)(2n)]} + \frac{(n+1)!(N-1)[(n+1)(n+2)...(2n)(2n+1)]}{n![(n+2)(n+3)...(2n+1)(2n+2)]}$$

$$= -\frac{n(N-1)n}{(n+1)(2n)} + \frac{(n+1)(N-1)(n+1)}{2n} = (N-1)\left[\frac{(n+1)(n+1)}{2(n+1)} - \frac{n^2}{2n}\right]$$

$$= (N-1)\left(\frac{n+1}{2} - \frac{n}{2}\right) = (N-1)\left(\frac{n+1-n}{2}\right) = \frac{N-1}{2} \tag{3-29}$$

$$\gamma_n = \frac{a_{n-1}d_n^2}{a_n d_{n-1}^2} = \frac{\left(\frac{1}{(n-1)!}(n)_{n-1}\right)\left(\frac{(N+n)!}{(2n+1)(N-n-1)!}\right)}{\left(\frac{1}{n!}(n+1)_n\right)\left(\frac{(N+n-1)!}{(2n-1)(N-n)!}\right)}$$

$$= \frac{n!(n)_{n-1}(N+n)!(N-n)!(2n-1)}{(n-1)!(n+1)_n(N+n-1)!(N-n-1)!(2n+1)} = \frac{n(n)_{n-1}(N+n)(N-n)(2n-1)}{(n+1)_n(2n+1)}$$

$$= \frac{n(N+n)(N-n)(2n-1)[n(n+1)...(2n-3)(2n-2)]}{(2n+1)[(n+1)(n+2)...(2n-1)(2n)]} = \frac{(N^2-n^2)n^2(2n-1)}{(2n+1)(2n-1)(2n)}$$

$$= \frac{n(N^2-n^2)}{2(2n+1)} \qquad \text{(3-30)}$$

The values for α_n, β_n, and γ_n shown above in equations 3-28, 3-29, and 3-30 agree exactly with that shown in Table 3 on page 129 of Nikiforov [12].

The paper by G. Carballo, R. Alvarez-Nodarse and J. S. Dehesa, [Carballo et.al.-10] normalizes the orthogonality relationship and thus provides a different looking three-term recurrence relation with different coefficients than shown above in equations 3-12 through 3-14. (The differences in these equations will be shown later after the Carballo equations are given. In the Carballo paper, the following was provided as the basis for discrete variable orthogonal polynomials using the same Gram-Schmidt orthogonalization process for discrete functions described in the earlier section titled “Orthogonal Polynomial Overview” as well as some well-known facts from the general theory of orthogonal polynomials.

Let *u(x)* be a non-constant and non-decreasing function in [a,b]. Then orthogonal polynomials must obey an orthogonalilty relationship given in equation 3-2 except we will replace *W(x)* with *u(x)* and polynomials with functions *f* and *g*.

$$\int_a^b f(x)g(x)u(x)dx = 0 \tag{3-31}$$

Next, let's define a new scalar product of two real functions *f* and *g* in terms of *u(x)*.

$$\langle f, g\rangle = \int_a^b f(x)g(x)du(x) \tag{3-32}$$

Equation 3-32 is therefore similar to the definition of an inner product shown in equation 3-3. A particular case corresponds to the ones when *u(x)* is a step function with jumps at *N* finite number of points. By taking *u(x)* as a step function we will show how the continuous variable procedure present earlier in this chapter is discretized. In this case equation 3-31 becomes

$$\langle f, g\rangle = \sum_{i=1}^{N} f(x_i)g(x_i)\rho(x_i) \ , \quad \rho(x) > 0 \ , \ \forall x \in [a,b] \tag{3-33}$$

where ρ is a discrete weight function. Given a sequence of linearly independent functions, it is always possible to obtain an orthogonal sequence. If we denote the determinant $D_n = \| u_{i+j} \|$, $i, j = 0$ to n, where

$$u_k = \int_a^b x^k du(x), k = 1,2,.... \tag{3-34}$$

are the moments associated with u , then using the Gram-Schmidt orthogonalization process illustrated for continuous variables, shown earlier, will yield a set of orthogonal polynomials for discrete variables. When u has N finite points of increase, n is finite with $n \le N$, then the following theorem holds.

Theorem 1: Given a distribution function u with moments u_k, $k = 1,2,...$, there exists a uniquely determined up to a constant multiplicative factor sequence of orthogonal polynomials $\{p_n\}$, each of which have degree exactly equal to n, providing that $D_n > 0$ for all $n \ge 0$. Moreover, if $\{p_n\}$, $n = 0,1,2,.....$, is a monic orthogonal polynomial sequence with respect to a weight function $\rho(x)$, then the polynomials p_n satisfy a three-term recurrence relation of the form

$$p_n(x) = (x - c_n) p_{n-1}(x) - \lambda p_{n-2}(x), \quad p_{-1}(x) = 0, p_0(x) = 1, n \ge 1 \tag{3-35}$$

where c_n and λ_n for $n = 0,1,2,.....$ are given by

$$c_n = \langle x p_{n-1}, p_{n-1} \rangle / \langle p_{n-1}, p_{n-1} \rangle, n \ge 1 \tag{3-36}$$

$$\lambda_n = \langle x p_{n-1}, p_{n-2} \rangle / \langle p_{n-2}, p_{n-2} \rangle, n \ge 2 \tag{3-37}$$

The paper by G. Carballo, R. Alvarez-Nodarse and J. S. Dehesa, [Carballo et.al.-10] shows that the classical discrete Chebychev monic polynomials are in fact a subclass of the Hahn polynomials with α = β = 0 and satisfy an orthogonality relationship of the form

$$\sum_{x=0}^{N-1} t_n(x,N)t_m(x,N) = \delta_{nm}\frac{n!^2\,(N+n)!}{(2n+1)(N-n-1)!(n+1)_n^2} \qquad x = 0,1,...,N-1 \qquad (3\text{-}38)$$

where δ_{nm} is the Kronecker symbol ($\delta_{nm}=1$ if $n=m$ and 0 otherwise) and $(a)_n = a(a+1)(a+2)...(a+n-1)$ denotes the Pochhammer symbol ($a = n+1$ in equation 3-19). Therefore, the three-term recurrence relation satisfying equation 3-35 has coefficients given by

$$c_n = \frac{N-1}{2} \qquad \text{and} \quad \lambda_n = \frac{(n-1)^2[N^2-(n-1)^2]}{4[4(n-1)^2-1]} \qquad (3\text{-}39)$$

In comparing the orthogonality relation and the three-term recurrence relation developed from the method provided in the Nikiforov text to those developed in the Carballo paper, one can see some difference attributable to the Carballo method utilizing a normalization technique. The equivalent equations (with originally assigned equation numbers) are shown next for easier comparison.

Orthogonality Relationship Differences

$$\sum_{x_i=a}^{b-1} y_n(x_i) y_m(x_i) \rho(x_i) = d_n^2 \delta_{nm} \tag{3-12}$$

$$\sum_{x=0}^{N-1} t_n(x,N) t_m(x,N) = \delta_{nm} \frac{n!^2 \, (N+n)!}{(2n+1)(N-n-1)!(n+1)_n^2} \tag{3-38}$$

Note that the two orthogonality relations above are actually quite similar. Considering the interval (0,N) instead of (a,b) for equation 3-12 gives the same summation limits as in equation 3-38. Also, per Table 3 on page 129 of Nikiforov [12]

$$\frac{n!^2 \, (N+n)!}{(2n+1)(N-n-1)!(n+1)_n^2} = \frac{d_n^2}{a_n^2} \tag{3-40}$$

which makes the right side of equation 3-12 similar to the right side of equation 3-38, the only difference being a normalization factor of $1/a_n^2$.

Three-Term Recurrence Relation Differences

$$y_{n+1}(x) = \frac{(x-\beta_n)y_n(x) - \gamma_n y_{n-1}(x)}{\alpha_n} \tag{3-14}$$

$$\alpha_n = \frac{(n+1)}{2(2n+1)} \quad (3\text{-}28) \qquad \beta_n = \frac{N-1}{2} \quad (3\text{-}29) \qquad \gamma_n = \frac{n(N^2-n^2)}{2(2n+1)} \quad (3\text{-}30)$$

$$p_n(x) = (x-c_n)\,p_{n-1}(x) - \lambda p_{n-2}(x), \quad p_{-1}(x) = 0,\, p_0(x) = 1, n \geq 1 \tag{3-35}$$

$$c_n = \frac{N-1}{2} \quad \text{and} \quad \lambda_n = \frac{(n-1)^2[N^2-(n-1)^2]}{4[4(n-1)^2-1]} \tag{3-39}$$

One difference between equations 3-30 and 3-39, when comparing γ_n and λ_n, is due to the difference in indexing. Equation 3-30 is base on the highest polynomial degree $n+1$ whereas equation 3-35 is based on the highest polynomial of degree n. However, when adjusting for the indexing difference, the equations are still different, though much more similar. The remaining differences in the equations above are attributable to Carballo using a normalization technique. In other words the remaining differences in the equation are attributable to the treatment of polynomial coefficients a_n and b_n, noting that α_n and β_n contain a_n, a_{n+1}, a_{n-1} , b_n, and b_{n-1} (see equation 3-15).

CHAPTER 4 : THE SEGMENTAL PROBABILISTIC MODEL

In Chapter 2, a basic assumption of the Frame Based Probabilistic Model is that each frame is independent of the next. This assumption is not valid for speech signals since the frames that might compose a single phoneme are in fact dependent. To overcome this dependency, Liu and Wang [Liu, Wang – 5] developed a segmental approach of polynomial representation of the waveform. In a segment based HMM system, an entire phoneme is usually contained in a single segment with a unique probability distribution. The segment length is determined by certain types of "landmarks" such as significant changes in amplitude. Thus, the segment is much more likely to completely represent a phoneme. This also overcomes the false assumption that the individual frames composing a single phoneme are independent. Then, in combining phonemes, there is greater independence between the segments and signal variations are preserved.

In the Liu-Wang Segmental Probabilistic Model (SPM), several frames are concatenated into a single segment which typically corresponds to a complete phoneme (see Figure 2-2). Then, the feature vectors for each frame are combined in the form of orthogonal polynomials for analysis. For example, if a given phoneme is processed as four frames, they will be concatenated into a single segment and the feature vectors are combined as follows. If the feature vector has three features, there will be three orthogonal polynomials computed, each containing the four similar features from the four vectors. These orthogonal polynomials have been shown to play a relevant role in speech recognition [Levitt and Rabiner- 10], particularly, in research on speaker dependent features of speech waveforms. This is the case where Legendre polynomials have

been used for speech recognition, speech enhancement and speaker adaptation [Deng et.al.-6, Fukada et.al-7, Holmes et.al.-8, and Gish et.al.- 9].

The Liu-Wang method, whose aim is to verify the identity of a claimed speaker, has a training phase and a verification phase. For a given speech signal of a specific speaker, the number of segments, the length of each segment, and appropriate segmental probabilistic model are determined in the training phase. The performance of the model depends on the mixture number (i.e. the number of acoustic segments used for modeling the speaker's voice characteristics) and the degree of the orthogonal polynomials. This degree affects the accuracy of the model. Liu and Wang showed by experimentation that the best accuracy was obtained with polynomials of degree 3.

The Orthogonal Polynomial Function

The Liu and Wang Segmental Probabilistic Model is a model where the parameters are mapped into a time sequence of feature vectors. This time sequence of feature vectors is where features are extracted from each frame with the features organized in one vector per frame and then stored in time sequence. Then calculation of the likelihood between the time sequence of given vectors and mapping vectors is possible. Given a set of orthogonal coefficients, A = $\{a_0, a_1, a_2, \ldots a_r, \ldots a_R\}$, the following formula can be used to regenerate a time sequence of L-length feature vectors, $X_l = \{x(1),...x(l),...x(L)\}$. The mapping formula is given by

$X_l = F \quad (A;L;R)$, (4-1)

with the column vector x(*l*) equal to

$$x(l) = \sum_{r=0}^{R} a_r \Phi_r^L(l) \text{ , for } l = 1,\ldots L, \quad (4\text{-}2)$$

where F is the orthogonal polynomial function whose input arguments are a set of orthogonal coefficients A, L is the segment length (i.e. the number of frames composing the segment) and R is the degree of the polynomial function. $\Phi_r^L(l)$ is a polynomial of degree r. The dimension of a feature vector x(*l*) is assumed to be *d*. Note that the segment length L determines the degree of the orthogonal polynomial. For an orthogonal polynomial of degree r, the smallest length L of the segment is (r+1).

The orthogonal polynomials $\{\Phi_0^L(l),....,\Phi_r^L(l),....,\Phi_R^L(l)\}$ in the interval [1,L] satisfy the following orthogonal conditions:

$$(1)\ \sum_{l=1}^{L} \Phi_r^L(l) = 0 \text{ , for r = 1,…,R.} \quad (4\text{-}3)$$

$$(2)\ \sum_{l=1}^{L} \Phi_r^L(l)\Phi_k^L(l) = 0 \text{ , for r} \neq \text{k and r, k = 1,….R.} \quad (4\text{-}4)$$

where r is the degree of the polynomial $\Phi_r^L(l)$ and $\Phi_0^L(l) =$ constant for $1 \le l \le L$. The orthogonal polynomial of degree r in the interval [1,L] can be derived from the lower-degree orthogonal polynomials $\Phi_{r-1}^L(l)$ and $\Phi_{r-2}^L(l)$ by the recurrence relation

$$\Phi_r^L(l) = (\Phi_1^L(l) + \alpha)\Phi_{r-1}^L(l) + \beta\Phi_{r-2}^L(l) \tag{4-5}$$

with

$$\alpha = -\frac{\sum_{i=1}^{L} \Phi_{r-1}^L(l)\Phi_{r-1}^L(l)\Phi_1^L(l)}{\sum_{i=1}^{L} \Phi_{r-1}^L(l)\Phi_{r-1}^L(l)} \tag{4-6}$$

$$\beta = -\frac{\sum_{i=1}^{L} \Phi_{r-1}^L(l)\Phi_{r-2}^L(l)\Phi_1^L(l)}{\sum_{i=1}^{L} \Phi_{r-2}^L(l)\Phi_{r-2}^L(l)} \tag{4-7}$$

$$\Phi_1^L(l) = l - \frac{1}{L}\sum_{m=1}^{L} m \tag{4-8}$$

$$\Phi_0^L(l) = \text{constant} \tag{4-9}$$

Equations 4-5 through 4-9 will be used in Chapter 5 to shown that the speech polynomials used in the Segmental Probabilistic model are in fact Chebychev polynomials.

Formulation of the SPM

Each orthogonal polynomial has a probability associated with it which will be determined by experimentation described earlier. The voice authentication system then compares polynomial coefficients for a stored set of speakers' speech polynomials and determines which speaker is the closest match. As there are many polynomials associated with each speaker and there may be many speakers, the best match is determined by finding the highest probability match through all possible combinations of segments represented by the polynomials. The HMM is used as the method of analysis to find the highest probability match.

A SPM is represented by $\Lambda = \{c_m, A_{R,m}, U_m \mid m = 1,...m\}$ where c_m is the mixture weight, and $A_{R,m} = \{a_{0,m},...a_{r,m},...a_{R,m}\}$ is a set of orthogonal coefficients which are used to generate segment mean according to the equations above, $a_{r,m}$ is an orthogonal coefficient vector for an orthogonal polynomial of degree r, U_m is a d x d dimensional covariance matrix where d is the dimension of a feature vector $x(l)$, and M is the total number of mixtures. For a set of signal feature vectors, $X = \{x(1),...x(t),...x(T)\}$, the log-likelihood for this signal X is given by

$$\log P(X \mid \Lambda) = \max_{B} \sum_{j=0}^{J-1} \log P(X_j \mid \Lambda, B), \tag{4-10}$$

where B is a possible segment boundary in the set $\{b_0,...b_j,...b_J \mid b_j \in [b_{j-1}+1, T]$, for $j = 1,...J$ with $b_0 = 0$ and $b_J = T\}$, J is the number of partitioned segments in accordance with B, $X_j = \{x(b_j+1),...x(b_{j+1})\}$ is the j[th] segment, $\log P(X_i \mid \Lambda, B)$ is defined as

$$\log P(X_j \mid \Lambda, B) = \log \sum_{m=0}^{M} c_m P(X_j \mid A_{R,m}, U_m, B) \tag{4-11}$$

$$P(X_j \mid A_{R,m}, U_m, B) = \prod_{t=b_j+1}^{b_{j+1}} (2\pi)^{-d/2} \mid U_m \mid^{-1/2} \times \exp[-\frac{1}{2} o_m^j(t)^T U_m^{-1} o_m^j(t)], \tag{4-12}$$

and

$$o_m^j(t) = \left(x(t) - \sum_{r=0}^{R} a_{r,m} \phi_r^{(b_{j+1}-b_j)}(t - b_j) \right). \tag{4-13}$$

Equation 4-10 illustrates that the log-likelihood $\log P(X \mid \Lambda)$ is obtained by choosing the optimal segment boundary from all possible segment boundaries. Algorithms are available which can find the optimal segment boundary for any given speech boundary.

In the training phase, the task is to find the optimal segment boundary B_{op} and the model parameters Λ_{op} such that the log-likelihood $\log P(X \mid \Lambda_{op}, B_{op})$ is a maximum for the given training feature vectors $X = \{x(1),...x(t),...x(T)\}$. $LogP(X \mid \Lambda_{op}, B_{op})$ is defined as

$$\log P(X \mid \Lambda_{op}, B_{op}) = \max_{B,\Lambda} \sum_{j=0}^{J-1} \log P(X_j \mid \Lambda, B) \tag{4-14}$$

where B is a possible segment boundary, $\log P(X_i \mid \Lambda, B)$ is defined in Equation 4-11, and J is the number of partitioned segments.

As equation 4-14 is difficult to solve, Liu and Wang went on to develop an iterative algorithm for solving Equation 4-14. This algorithm can be found in Liu and Wang [5].

Speaker Verification

The purpose of speaker verification is to verify the identity of a claimed speaker. This is accomplished first by the input of speaker training data. This data is processed by the Segmental Probabilistic Model where the log likelihood of equation 4-10 is calculated and stored. Then given the utterance of the claimed speaker designated by $Y = \{y(1)...y(T)\}$, the log likelihood of the utterance is computed by

$$\log P(Y \mid \Lambda) = \max_{B} \sum_{j=0}^{J-1} \log P(Y_j \mid \Lambda, B) \qquad (4\text{-}15)$$

Where B is a possible segment boundary and J is the possible number of segments in accordance with B. The training log likelihood and the new utterance log likelihood are then compared, and based on a predetermined threshold value, the speaker is either verified as claimed or rejected.

In experimental tests of the Segmental Probabilistic Model, Liu and Wang found that for a larger mixture number (i.e. larger number of frames composing the segment), a higher spectral resolution for the speaker model was obtained. In other words, as the mixture number increases, the error rate decreases. Also, Liu and Wang found that as the degree of the orthogonal polynomial was increased from 1, accuracy improved though improvements were ever smaller for each increase in degree. In fact, beyond a polynomial of degree 3, they observed that the accuracy began to decrease. In addition, higher degree polynomials require larger training databases and require additional computation time. Finally, it was found that the SPM model

was at least 50% more accurate than the FBPM. Improvements beyond 50% were noted as each model was reduced in terms of the number of mixtures [Liu, Wang - 5].

CHAPTER 5 : CHEBYCHEV POLYNOMIALS

G. Carballo, R. Alvarez-Nodarse and J. S. Dehesa, [Carballo et.al.-10] show that the speech polynomials in the Liu Wang segmental model are shifted Chebychev polynomials. In doing so, the authors observed that some mathematical tools used in speaker recognition methods could be considerably reduced which implies a big reduction and simplification in the algorithms inherent to these methods. This section will show that the polynomials in the Liu Wang model are in fact Chebychev polynomials. We also provide a discussion of how the special characteristics of orthogonal polynomials simplify the tasks inherent to speaker identification.

Speech Polynomials as Chebychev Polynomials

In the Liu-Wang model, a sequence of orthogonal polynomials were shown to satisfy the orthogonality relationship given in equation 4-4 which corresponds to the discrete scalar product given in equation 3-33. Thus, we are allowed to use Theorem 1 to claim that the polynomials given by the Liu-Wang model, $\Phi_n^L(l)$ are uniquely determined up to a constant factor. Also, since the distribution function *u(x)* is a step function with L jumps at points $x = 1,2,...L$, the family $\Phi_n^L(l)$ is finite and in terms of equation 4-4, $R \le L-1$. Now, in comparing equation 4-4 with equation 3-38, we can see that the speech polynomials $\Phi_n^L(l)$ are proportional to the Chebychev polynomials $t_n(l-1, L)$. Making a change of variable $x = l-1$ in equation 3-38 yields

$$\sum_{x=0}^{N-1} t_n(x,N)t_k(x,N) = \sum_{l=1}^{L} t_n(l-1,L)t_k(l-1,L) = 0 \qquad \text{n} \neq \text{k and r, k=0,1.2…L-1} \tag{5-1}$$

Equation 5-1 now corresponds to equation 4-4. Also, the square norm of $\Phi_n^L(l)$ has the explicit form

$$\sum_{l=1}^{L} \Phi_n^L(l)\Phi_n^L(l) = \frac{n!^2\,(L+n)!}{(2n+1)(L-n-1)!(n+1)_n^2} \tag{5-2}$$

Considering monic polynomials and using the three-term recurrence relation for Chebychev polynomials given in equations 3-35 and 3-39, we can give the speech polynomials $\Phi_n^L(l)$ in the form

$$\Phi_n^L(l) = (l - \frac{L+1}{2})\Phi_{n-1}^L(l) - \frac{(n-1)^2[L^2-(n-1)^2]}{4[4(n-1)^2-1]}\Phi_{n-2}^L(l) \tag{5-3}$$

Equation 5-3 is the same relation given by Liu-Wang in equations 4-5 through 4-9 where

$$\alpha = 0 \text{ and } \beta = -\frac{(n-1)^2[L^2-(n-1)^2]}{4[4(n-1)^2-1]}. \tag{5-4}$$

Spectral Moments of Speech Polynomials

In the section above, it was shown that the speech polynomials given in the Liu-Wang Segmental Probabilistic Model are in fact Chebychev polynomials. By knowing that the speech polynomials are Chebychev polynomials many useful properties can be obtained. One very useful property of the speech polynomial concerns the moments of their zeros. These moments are defined by

$$\mu_0 = 1 \text{ and } \mu_m^{'(n)} = 1/n\sum_{k=1}^{n} x_{k,n}^m \text{ , } m = 1,2,...L-1, n \le L-1 \tag{5-5}$$

Where $x_{k,n}$, k = 1, 2, …n denotes the zeros of the polynomial $\Phi_n^L(l)$. Using methods provided by Alvarez-Nordarse [Alvarez-Nordarse et al – 13] the first few spectral moments of $\Phi_n^L(l)$ have expressions

$$\mu_1^{'(n)} = \frac{L-1}{2}, \tag{5-6}$$

$$\mu_2^{'(n)} = \frac{(1+3L(1+L))+(2+3L)^2 n+2n^2-n^3}{24n-12} \tag{5-7}$$

$$\mu_3^{'(n)} = \frac{(1+L)(-2L-4L^2+4Ln+5L^2 n+2n^2-n^3}{16n-8} \tag{5-8}$$

These measures give different dispersion measures of the zeros of the speech polynomials. As each speaker has different characteristic speech in terms of the number of zero crossings and the space between these zeros, this measure of the dispersion is an important and useful tool in identification of the speaker. Also useful is the centroid, $\mu_1^{'(n)}$, of the distribution of zeros and the variance, σ^2, given by

$$\sigma^2 = \mu_2^{'(n)} - (\mu_1^{'(n)})^2 = \frac{(n-1)(3L^2 + n - n^2 - 1)}{24n - 12} \tag{5-9}$$

CHAPTER 6 : CONCLUSIONS

As stated in the Introduction, identification of speech is a difficult problem. A narrower problem of speech recognition is speaker recognition. Research has provided several models to solve this problem. Two methods for solving speaker recognition problems discussed in this paper involve the Frame Based Probabilistic Model and the Segmental Probabilistic Model. Because of assumptions of independence between frames being faulty, the segmental approach was shown to be more effective. The segmental model utilized orthogonal polynomials to represent the speech waveform. It was then shown that the speech polynomials were in fact Chebychev polynomials. In showing that the speech polynomials were Chebychev polynomials, many useful and well know properties of Chebychev polynomials can be used to compare various speech waveforms. With these extra properties at the disposal of the speech recognizer, the accuracy of recognition should be increased.

Now that the speech polynomials have been identified as Chebychev polynomials, future research can be conducted into how to use the characteristics of Chebychev polynomials. The moments of zeros of the speech polynomial were mentioned above. Other possible characteristics that may be investigated in the future in terms of the Chebychev polynomial include, but not limited to, autocorrelation, covariance, the average slope at the points of the zero crossings and the time indexed frequency of the spoken words.

Today, speaker verification systems are more accurate because they are concerned only with individual speakers. These systems compare a speaker's utterances to his own previously recorded speech. Also, current speech recognition systems are very accurate for systems that have a training phase and require recognition for a specific speaker. However, speech

recognition is not very accurate for speaker independent applications because of the differences between speakers. This inaccuracy for speaker independent applications may be improved by knowledge that the speech polynomials are Chebychev polynomials. This is an area of ongoing research.

LIST OF REFERENCES

1) F.Jelinek, Statistical Methods for Speech Recognition, The MIT Press, 1998

2) J. Deller, J. Hansen, J Proakis, Discrete-Time Processing of Speech Signals, Macmillan, 1993

3) V.J. Mathews, G.L. Sicuranza, Polynomial Signal Processing, Wiley, 2000

4) D. Paulus, J. Hornegger, Applied Pattern Recognition, Vieweg, 3rd edition 2001

5) C. Liu, H. Wang, A Segmental Probabilistic Model of Speech Using an Orthogonal Polynomial Representation – Application to Text-independent Speaker Verification, Speech Communication, Vol 18, pp 291-304, 1996

6) L. Deng, M. Askamovic, X. Sun and J. Wu, Speech Recognition using Hidden Markov Models with Polynomial Regression Functions as Non-stationary States, IEEE Trans on Speech and Audio Processing 2 (4), pp 507-520, 1994

7) T. Fukada, Y. Sagisaka and K. Pahwal, Model Parameter Estimation Mixture Density Polynomial Segment Models, Proc ICASSP 2, pp 1403-1407 , 1997

8) W.J. Holmes and M.J. Russell, Linear Trajectory Segmental HMMs, IEEE Signal Processing Letters 4 (3), pp 72-74, 1997

9) H. Gish and K. Ng, Parametric Trajectory Models for Speech Recognition, Proc ICSLP, Volume 1, pp 446-469, Philadelphia, Pa., October 1997

10) G. Carballo, R. Alvarez-Nodarse and J. S. Dehesa, Chebychev Polynomials in a Speech Recognition Model, Applied Mathematics Letters 14 (2001) pp. 581-585

11) H. Levitt and L.R. Rabiner, Analysis of Fundamental Frequency Contours in Speech, The Journal of the Acoustical Society of America, Volume 49 Number 2 (1971) pp. 569-582

12) A Nikiforov and V. Uvarov, Special Functions of Mathematical Physics, (1988) pp. 106-129

13) R. Alvarez-Nordarse and Js Dehesa, Zero Distribution of Discrete and Continuous Polynomials From Their Recurrence Relations, Applied Math Comput.)

14) T. Matusi and S. Furui, Comparison of text-independent Speaker Recognition Methods using VQ-Distortion and Discrete/Continuous HMMs, Proc. International Conference Acoustical Speech Signal Processing-1992, Volume 2, pp 157-160

15) B.L. Tseng, F.K. Soong and A.E. Rosengerg, Continuous Probabilistic Acoustic Map for Speaker Recognition, Proc. International Conference Acoustical Speech Signal Processing-1992, Volume 2, pp 161-164

16) A.P. Dempster, N.M. Laird and D.B Rubin, Maximum Likelihood from Incomplete Data Via the EM Algorithm, J. Roy Statistical Society, Volume 39, pp 1-38, (1977)

17) Gerald Gazdar, from a course web page updated 25 March 1999
www.informatics.susx.ac.uk/research/nlp/gazdar/teach/nlp/nlpnode49.html