FUZZY SPEECH RECOGNITION

by

Patrick M. Mills

Bachelor of Science in Engineering
Swarthmore College, 1994

Submitted in Partial Fulfillment of the

Requirements for the Degree of Master of Science

in the Department of Electrical and Computer Engineering

College of Engineering

University of South Carolina

1996

Department of Electrical and Computer Engineering Director of Thesis	Department of Electrical and Computer Engineering Second Reader

Dean of the Graduate School

ACKNOWLEDGMENTS

I would like to thank my parents for their love, dedication, and support through my graduate career. Without them, I would never have been able to push myself enough to complete my goals.

I would also like to thank my advisors: John Bowles and Juan Vargas. Their criticisms and suggestions helped mold this thesis into a readable and hopefully useful work.

ABSTRACT

Speech recognition allows for extremely efficient man-machine communication. Since most computers are used to interactively input or output data, communication by speech represents the ideal computer interface. Recognition requires intelligence, and is, therefore, a much more difficult problem than speech production. While general speech recognition is a daunting task, a much simpler system would still be useful. This thesis presents a simple speech recognition system that can be implemented with a personal computer and a sound card. Once a limited system has been implemented, its capabilities can be scaled by using faster computers and specialized hardware as necessary.

Fuzzy logic allows effective decision making in the presence of uncertainty. Identifying spoken words, even in an ideal environment by a trained speaker, is a complex task filled with uncertainty. The speech waveform is nonlinear and variant, removing the possibility of simple analysis. However, by analyzing the waveform for reoccurring and semi-stable features, small segments may be classified. A fuzzy expert may then make decisions based on these features to identify the spoken word. The identification represents the decision that the chosen word is present and also that other words are not present. Furthermore, the system's confidence in its identification can be used to accept the identification or to request further information or help.

TABLE OF CONTENTS

Introduction 1

- Background 1
- Problem Definition 2
- Thesis Organization 3

The Speech Waveform 4

- Physical Manifestation 4
- Information Theory 4
- Unit of Representation 5
- Speech Features 7
- An Example 9
- Summary 11

Biological Influences 12

- Structure of the Ear 12
- Infant Language Acquisition 14
- The Brain 17
- Sound Production 18
- Summary 20

Speech Recognition Models 22

- The Early Days 22
- Computer Based Speech Recognition 23
- Speech Recognition Techniques 23
- Time Normalization 28
- Summary 30

Fuzzy Logic 32

- Background 32
- Fuzzy Sets 32
- Fuzzy Logic 35
- Fuzzy Systems 36
- Summary 37

Isolated Digit Recognition 39

- Digit Waveforms 39
- Fuzzy Speech Recognizer 40
- Symmetric Dynamic Programming 41
- Fuzzy Symmetric Dynamic Programming 47
- Summary 48

System Results 50

- System Configuration 50
- Initial Trial 51
- Additional Trials 52
- Sub-Sampling Trials 55

Summary 56
Conclusions 57
Suggestions for Future Work 57
Conclusions 57
Appendix A: Code 59
Glossary 60
Bibliography 62
Index 65

LIST OF TABLES

International Phonetic Alphabet 6
Digit Representations 40
Waveform Distance Matrix 45
Waveform Gradient Matrix 46
Initial Results with a Trained Male Speaker 51
Results with a Trained Male Speaker 53
Results with an Untrained Female Speaker 54
Crisp Template Correlation for Male Speaker 54
Sub-Sampling Results 56

LIST OF FIGURES

Amplitude Graph of the Word "recognition" 10
Spectrogram of the Word "recognition" 11
The Ear 13
Cross-section of the Human Vocal Apparatus 19
Symmetric Dynamic Programming Flowchart 29
Gradient Paths 30
A Fuzzy Set Representing *Tall* 34
A Linguistic Variable Representing *Height* 34
A General Fuzzy Controller 37
Voice Recognition System Window 41
Area Searched by Dynamic Programming Algorithm 43
Symmetric Dynamic Programming Example 44
Fuzzy Number Membership Function 48

CHAPTER 1

INTRODUCTION

1.1 Background

Speech is a human's most efficient communication modality. Beyond efficiency, humans are comfortable and familiar with speech. Other modalities require more concentration, restrict movement, and cause body strain due to unnatural positions.

In the 1950s, most computer system input was achieved through switches, and results were read off of LEDs. Later, punch cards were used. In the late 1970s, the CRT terminal became common, allowing easier, more efficient input and output. Keyboard entry is much faster than setting switches or creating punch cards, and reading characters is much faster than deciphering binary results from LEDs or punch cards. However, even with training, typing is slower than continuous speech. Furthermore, while typing or reading, the user must focus on the task of input and output. During speech, the user is able to perform other tasks allowing for multi-modal input and output.

During spontaneous speech, an average of 2.0 to 3.6 words per second can be communicated. Skilled typists can type around 1.6 to 2.5 words per second, but only when typing prepared text. A skilled typist averages only 0.3 words per second when typing spontaneous text or when problem solving. This productivity is comparable to unskilled typists who type only 0.2 to 0.4 words per second under optimal conditions. The average speed of handwriting text is around 0.4 words per second.[1] Silent reading can achieve 2.5 to 9.8 words per second, but to achieve the higher rates with high retention, the reader must concentrate only on reading.[2] These comparisons suggest that

[1] Lea, p. 6.

[2] Newell, p. 50.

an optimal man-machine interface would incorporate speech recognition as completely as possible.

With computers becoming ever present in business, government, and education, there is a tremendous market for faster, more efficient man-machine interfaces. The majority of computer processing time is used for word processing, data entry, or waiting for these inputs. By allowing humans to communicate in their most natural mode and at natural speeds, efficiency, quality, and throughput increase. Speech also offers easy and cost-effective communication over long distances and access to the disabled.

These are convincing reasons for researching and developing speech recognition. However, achieving recognition is a complex and daunting task. While humans learn language as children by exposure only, machines require complex systems to even perform the most basic of recognition tasks. Even after over forty years of research, speech recognition as a science is still in its infancy.

1.2 Problem Definition

The main problem in speech recognition, as with other complex tasks that require some form of intelligence, is the amount of information that must be examined before making a classification or decision. Speech recognition is nothing more than an extremely complex pattern matching problem. The complexity arises from the variability in speech rate, pitch, volume, and emotion. Together with the natural differences in individual human voice production systems, these factors produce a variable and nonlinear waveform. As if these challenges were not enough, a speech recognition system must also deal with non-speech sounds and environmental noise.

Before creating a general system to perform continuous recognition in a noisy environment with multiple speakers, a simpler system should be designed to allow a single trained speaker in an ideal environment. Further simplifications can be made by restricting the vocabulary and attempting to identify only isolated words. This thesis is concerned with the isolated recognition of the set of ten digits [0 - 9] through the use of digital processing techniques and the application of fuzzy techniques. After successfully completing this "simple" task, the vocabulary may be increased and multiple speakers can be incorporated.

1.3 Thesis Organization

In Chapter 2, the speech waveform is examined. Reoccurring features of the speech waveform are shown. Chapter 3 covers the human auditory system, which is the most adept and comprehensive recognition system known. The chapter examines information gained from both biological and psychological studies, as well as the speech production system. Chapter 4 reviews the history of speech recognition and current methods. The chapter focuses on methods used to locate and identify the reoccurring features of speech discussed in Chapter 2. Chapter 5 reviews fuzzy set theory and fuzzy logic. The use of fuzzy logic allows effective decision making even in the presence of uncertainty. Chapter 6 discusses the isolated digit recognition system developed for this thesis. The design tradeoffs and requirements are reviewed in depth. Chapter 7 presents the results of the system, and Chapter 8 contains conclusions and suggestions for future work.

CHAPTER 2

THE SPEECH WAVEFORM

2.1 Physical Manifestation

The speech waveform is created by a series of discrete stimuli that control the lungs, vocal tract, tongue, and mouth. The resulting sound depends not only on the current stimulus but also on past stimuli causing a coarticulation effect. While the inputs that produce the speech waveform are discrete, the waveform itself is a "two dimensional non-denumerable continuum."[3] No matter how precisely we measure the waveform at any instant in time, a finer, more precise measurement is always possible. However, the continuity of the speech waveform has little to do with recognition. Proof resides in the ear itself which discretely samples and filters the speech waveform before transmitting a representation to the brain. Based on this information, it can be inferred that the speech waveform contains much more information than is necessary for recognition.

2.2 Information Theory

Information theory provides an estimate of the information content of the speech waveform. Words can be seen as units of information. Information theory terms the average amount of information per symbol (measured in bits) as entropy:[4]

$$H(S) = \sum_{i=1}^{n} -P(S_i)\log_2 P(S_i)$$

Entropy in a word S composed of symbols S_i (i = 1, 2 … n) depends on the probability $P(S_i)$ of occurrence of each symbol S_i. For written American English, information theory defines an entropy of 4.1 bits / symbol. Depending on the unit of

[3] Lea, p. 43.

[4] De Mori, pp. 26 - 28.

information used for analyzing speech, entropies of 4.1 to 7.5 bits / symbol have been calculated. However, using a hedge to account for context, an entropy of around 4.5 bits / symbol is generally assumed. Using this information, the average information rate of speech can be estimated to be around 50 bits / sec.[5]

However, speech of telephone quality must be sampled at 10 kHz with 11 bits / sample. The average amount of information is then 110,000 bits / sec. Therefore, the amount of information contained in the speech waveform is more than 2000 times greater than the amount of information contained in speech itself. The task of a speech recognition system is to reduce the amount of information contained in the speech waveform to a more manageable and relevant level.

2.3 Unit of Representation

Speech can be broken down into several competing components: allophones, phonemes, diphones, syllables, or words.[6] Each representation has advantages and disadvantages. "Phone" denotes a minimal unit of speech sound. However, identifying an individual phone is practically impossible due to an anticipation effect which causes phones to overlap.

Researchers use allophones to represent the set of phones containing the same information content. Allophones can be identified reliably, but require complex, time consuming procedures. Also useful is the fact that allophones can be employed to identify word boundaries. However, like the phone, the allophone suffers from an anticipation or coarticulation effect. An even greater disadvantage is the large number of allophones that can be contained in a given language.

Phonemes are the collection of allophones that operate similarly in a language. Phonemes have an advantage over allophones in that the number of distinctive phonemes

[5] De Mori, pp. 26 - 28.

[6] Lea, pp. 125 - 131.

is quite small. However, phonemes are not easy to distinguish acoustically, overlap each other, and require complicated processing procedures. The 44 phonemes shown in Table 2.1 belong to the International Phonetic Alphabet (IPA). The symbol used for each phoneme is the most commonly accepted symbol, but others are sometimes used. A list of the IPA can be found in many linguistic texts or in a good dictionary.

Phoneme	Example	Phoneme	Example
æ	pat	η	sing
e	pay	o	toe
ə	about	U	book
a:	father	a	pot
ɛr	care	u	boot
b	bet	ɔ	bought
tʃ	church	aU	out
d	debt	ɔɪ	boy
ɛ	bet	p	pet
i	bee	ɹ	rent
f	fire	s	sat
g	get	ʃ	shut
h	hat	θ	thing
ɪ	bit	t	ten
aɪ	by	ð	that
ir	pier	Λ	but
ər	butter	ɜr	term
d	judge	v	vat
k	kit	w	wit
l	let	y	you
m	met	z	zoo
n	net	ʒ	azure

Table 2.1: International Phonetic Alphabet
The IPA contains 44 phonemes. An example of each is shown with the sound underlined.

A diphone is a transitional sound identified by segmenting adjacent phones at their steady-state centers. By their very nature, diphones include transitional information that can be useful in identification. Like allophones, the number of diphones in a given language can be quite large and require a unique set of phonological rules for processing.

A syllable is basically "a vowel nucleus and its functionally related neighboring consonants."[7] Syllables are relatively easy to identify in the acoustic stream. As a bonus, many rules developed for dealing with phonemes are easily extended for use with syllables. The major problem with syllables is the difficulty in identifying boundaries. Again, as with phonemes and diphones, the number of syllables in a given language can be unmanageably large.

Finally, every speaker in a language knows instinctively what a word is, yet it is difficult to define phonologically. However, a word is the smallest unit of information that communicates a complete message. Using words as the basic unit of recognition eliminates many lower levels of recognition. However, temporal differences between examples of spoken words make identification difficult, especially with larger vocabularies.

Segmenting the speech input into units reduces the amount of information that must be processed by higher level procedures. The larger the unit, the less information that must be dealt with, but the more difficult accurate identification. Identification of the units depends on identifying the sub-features they contain.

2.4 Speech Features

The speech recognition task can be broken into four processes: data acquisition, filtering, feature identification, and classification. The first two processes affect the amount of data that must be processed, but the last two processes are by far the most difficult. Feature identification involves searching the speech waveform for semi-constant patterns that reoccur under specific conditions. In order for a spoken language to

[7] Lea, p. 129.

convey information, it must consist of a finite number of distinguishable and mutually exclusive sounds.[8]

The following features can be used to identify a spoken sound regardless of the unit of identification. However, researchers have primarily used these features when identifying phonemes; Chomsky and Halle defined them for this reason in 1968. Due to the variance in the speech waveform, the classifications become less useful as the unit of identification becomes larger. The features are generally grouped into mutually exclusive groups. But as with many other parts of the sound waveform, absolute distinctions are difficult. The groups tend to overlap causing further difficulties in identification. Used together, the groups can accurately define a sound unit.

Voiced vs. Unvoiced:

> Voiced sounds are caused as air pressure pushes the vocal cords open and causes them to vibrate. The sound produced has a pitch or fundamental frequency that is directly related to the frequency of vibration. The peak amplitude of a voiced sound is much higher than an unvoiced sound. Unvoiced sounds occur when the vocal cords are held open allowing air to pass through unaffected. The two fundamental unvoiced sounds are the *plosive* and the *fricative*. A plosive is generated by a build up of air behind the lips which is rapidly expelled. A fricative is generated by a turbulent airflow through a constriction such as the teeth.[9]

Vocalic vs. Nonvocalic:

> Vocalic sounds have a sharply defined formant structure. The formant is a band of high energy concentrated in a specific frequency range. Formants are generally found in vowels. Nonvocalic sounds lack a defined formant structure.

Consonant vs. Nonconsonant:

> Consonant sounds have high total energy, while nonconsonant sounds have a lower total energy.

Tense vs. Lax:

> Tense sounds have high total energy over a relatively long period of time in a wide frequency band. Lax sounds have lower total energy concentrated in a shorter time period in a tight frequency band.

[8] Newell, p. 45.

[9] Chen, pp. 50 -51.

Nasal vs. Oral:

Nasal sounds occur when the nasal passage is used as an auxiliary acoustic tube. These sounds contain a wide frequency spread and a reduction in the intensity of formants. Oral sounds contain a more defined frequency range and usually high intensity formants.

Strident vs. Mellow:

Strident sounds contain higher intensity noise than mellow sounds.

Grave vs. Acute:

Grave sounds have a higher concentration of energy in the lower frequencies, while acute sounds have a higher concentration of energy in the upper frequencies of the spectrum.

Compact vs. Diffuse:

Compact sounds contain a high concentration of energy in a narrow region of the spectrum. Diffuse sounds spread their energy across a wider region.

Flat vs. Plain:

Flat sounds are characterized by a weakening of the higher frequency components. Plain sounds contain no such weakening.

2.5 An Example

Figures 2.1 and 2.2 show an example of the speech waveform for the word "recognition." The word was sampled at 11 kHz and contains 7,465 signed 8-bit samples. The word is easily segmented into syllables as rec•og•ni•tion, or as time periods 1-3, 4-5, 6-7, 8-10. When segmented into phonemes, the word is represented as ɹ ɛ k ə g n ɪ ʃ ə n, with each phoneme taking one time period [1, 10].

Each phoneme can be characterized by its distinctive features. The phoneme ɹ is voiced, consonant, oral, acute, and flat. The phoneme ɛ is voiced, nonconsonant, lax, acute, and compact. The phoneme k is unvoiced, nonvocalic, consonant, tense, mellow, and compact. The phoneme ə is voiced, nonconsonant, tense, acute, diffuse, and flat. The phoneme g is unvoiced, nonvocalic, consonant, lax, mellow, and compact. The phoneme n is voiced, consonant, nasal, acute, and diffuse. The phoneme ɪ is voiced, nonconsonant, lax, acute, and diffuse. The phoneme ʃ is unvoiced, vocalic, consonant, tense, acute, and compact.

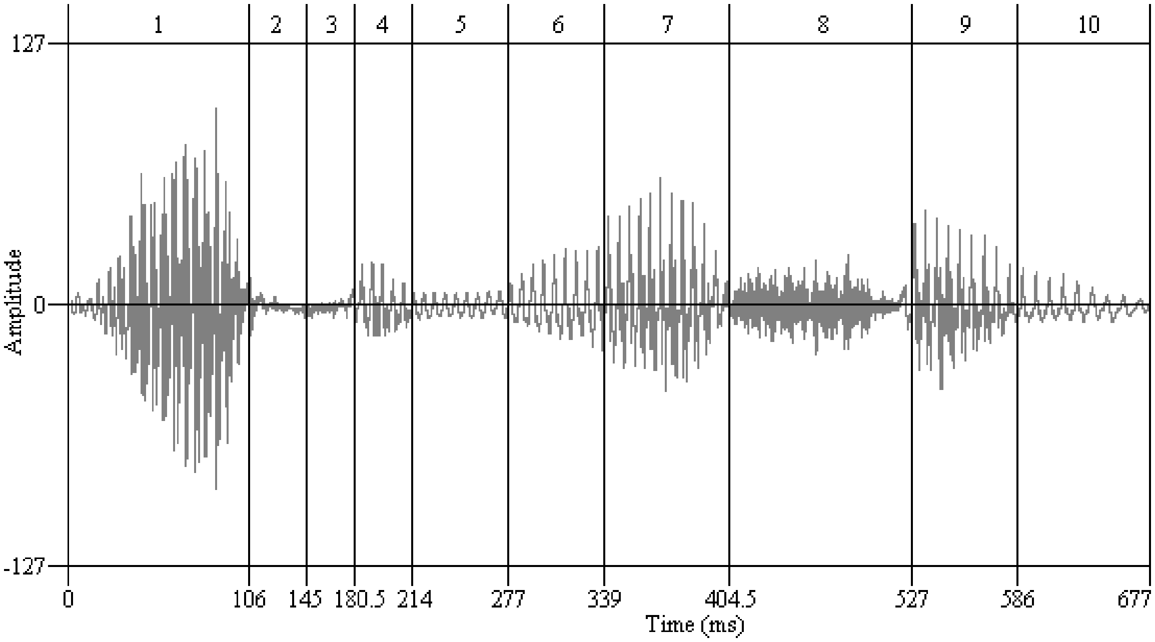


Figure 2.1: Amplitude Graph of the Word "recognition"
The time based graph shows the word represented as 8-bit signed values.
The word has been segmented into phonemes (1 - 10). Phonemes 4 and 9 are
the same, as are 6 and 10. Note the similarities and differences between them.

There are two ə phonemes in the word (segments 4 and 9 in figure 2.1 and denoted p_4 and p_9), and the comparison is enlightening. Notice the amplitude difference; p_4 has a lower total energy content than p_9. This is a prosodic effect caused by the natural stress positions in the word. Also p_9 contains more high frequency components than p_4. The ə phoneme is a flat sound; therefore, the higher frequency components in p_9 are actually carried over from the preceding ʃ phoneme. This carryover is called a coarticulation effect. There are also two n phonemes (denoted p_6 and p_{10}). Again, the amplitude difference is caused by prosodic effects. Notice that the frequency components are quite similar, since the preceding phonemes are either unvoiced or contain similar frequency components.

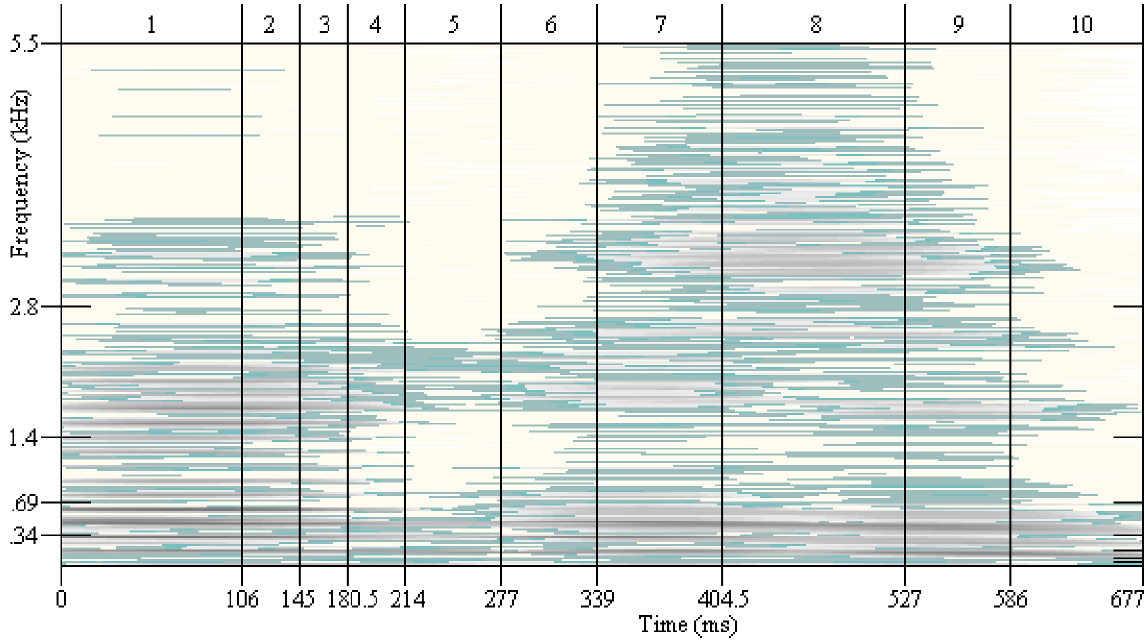


Figure 2.2: Spectrogram of the Word "recognition"
The time based spectrogram shows frequency intensity represented by gray shades; the darker the shade, the higher the intensity. The word has been segmented into phonemes (1 - 10). Phonemes 4 and 9 are the same, as are 6 and 10. Note the similarities and differences between them.

2.6 Summary

Even with the features of speech sounds defined, the task of speech recognition is not complete, as has been shown in countless attempts at speech recognition over the past forty years. There is still too much information in the speech waveform, and the reliable detection of information units is too slow. Furthermore, the effects of prosody, coarticulation, and anticipation cause nonlinear disturbances in the waveform. The next chapter provides biological (and psychological) background that can be used to eliminate unnecessary information from the speech waveform.

CHAPTER 3

BIOLOGICAL INFLUENCES

3.1 Structure of the Ear

The ear is a mechanical transducer that converts pressure into electrical impulses. The human ear consists of three distinct parts: the outer ear, the middle ear, and the inner ear. The outer ear is responsible for collecting sound waves and concentrating them onto the eardrum. The middle ear acts as a mechanical amplifier. Due to the compressibility of air, only a small force is necessary to cause the eardrum to vibrate. As the eardrum vibrates, the bones of the middle ear move, exerting over thirty times as much pressure on the fluid filled inner ear. The main organ of hearing is housed inside the cochlea. The cochlea is a spiral shaped cone that is divided into two main sections by the basilar membrane. The bottom half is called the scala tympani and ends at the round window. The top half, called the scala vestibuli, is connected to the ossicles, the bones in the middle ear, by the oval window. The scala vestibuli is further divided into two parts; the lower half, the scala media, contains the primary organ of hearing, the organ of Corti. The organ of Corti contains 3,000 inner cilia that raise it above the basilar membrane and 20,000 outer cilia that support the tectorial membrane. At the base of the hair cells are nerve fibers which form the cochlear nerve. The connection between nerve fibers and hair cells is not one-to-one. Figure 3.1 shows the three parts of the ear and a view of each section of the cochlea.

Pressure exerted by the ossicles on the oval window causes the basilar membrane to bulge into the scala tympani. The organ of Corti is supported by two stiff rods in a triangular configuration that cause it to rotate towards the center of the scala vestibuli rather than bulging into the scala tympani. This rotation causes the cilia lining the organ of Corti to be bent by the shearing force of the tectorial membrane. The nerves connected to the base of the cilia, excited by this bending motion, send impulses to the brain.

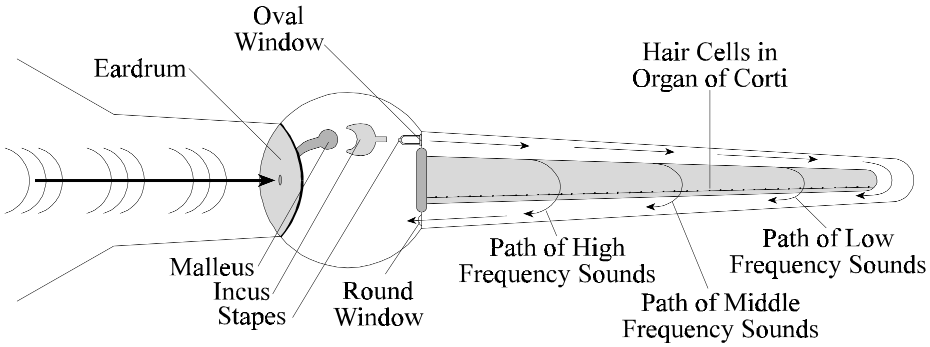


Figure 3.1: The Ear
The path of sound waves as they enter the ear, are transmitted through the middle ear, and pass through the cochlea (shown uncoiled).

The impulses sent to the brain encode both amplitude and frequency information. Amplitude is detected by the amount of bending in the cilia, while frequency is determined by which cilia move. The basilar membrane changes stiffness from one end of the cochlea to the other. Therefore, high frequencies tend to cause vibration only in the front end of the basilar membrane, middle frequencies in the front and middle (stronger), and low frequencies all over, with the strongest occurring towards the tip. Using this information, the brain can determine a sound's frequency. The human ear is generally capable of detecting frequencies in the range of 16 to 20,000 Hertz, although sensitivity is logarithmically distributed.

Experiments have shown that humans are largely phase insensitive. The basilar membrane is only deformed when the stapes pushes on the oval window, and thus very little information is available to the brain to determine a waveform's phase. Applying this fact to speech recognition algorithms can reduce the amount of data in the encoded waveform by half.

At the initial onset of a sound, the firing rate of auditory nerves is highest. However, as time passes, the discharge rate will decay to a steady-state, an effect called adaptation. Initial adaptation, known as rapid adaptation, is caused by a refractory property of the nerve fibers.[10] The slower adaptation is caused by a depletion of neurotransmitters. For low frequency sounds, the nerves tend to fire in phase; however, for high frequencies, they fire in integral multiples of the period. This difference occurs because a neuron can only fire several hundred times per second, since it uses a chemical process to rebuild the electrical differential after firing. After adaptation, a change in fundamental frequency will only be noticeable if the amount of change is at least ±0.5% of the fundamental frequency. Changes in formant frequency are noticeable only if they are at least ±5% of the formant frequency.

3.2 Infant Language Acquisition

Infants are able to acquire language without training. They build a lexicon through exposure only. This ability is of extreme importance to machine speech recognition. Through this feat alone, the ability to perform speech recognition can be proven. If infants were born with a lexicon, it might not be possible for a machine to recognize speech. The inborn lexicon could be used in a top-down process, using semantic and linguistic information to predict the speech waveform. If such information were required for speech recognition, then isolated word recognition would not be possible since no such information would be present. This is not to say that such information is not used, but rather that the information is used in an auxiliary process to help recognize more difficult waveforms and to provide context to a deciphered waveform. Since the lexicon is acquired through repeated exposure, the infant must use a bottom-up strategy.

[10] Waibel, pp. 101 - 102.

At as early as four days old, infants are able to differentiate between speech and music. By four months, infants can segment and recognize speech. The key to lexicon acquisition is the inborn ability to segment the speech waveform. Segmentation is the key feature of the bottom-up processing model. Segmentation involves identifying word onset, dividing the word into syllables as they are presented, and identifying word conclusion. This type of serial segmentation is incredibly powerful, and is perhaps the optimal solution to one of speech recognition's fundamental and more difficult problems. Once segmented, prosodic and syllabic analysis are performed. Prosody aids in segmentation and provides context sensitive information through prominence, intonation, and melody. Note that syllabic was chosen over other segmentation units. Psychology has shown that infants segment speech sounds into units that are basically syllabic in nature.[11]

Prosody comprises all the attributes of speech that are not part of the syllabic features.[12] At this point in time prosody is a uniquely human feature; attempts have been made to model prosody, but much research remains before accurate models can be created. The distinction between prosodic and syllabic features is quite important. If a child had to learn a new word for each type of intonation, the normal vocabulary would explode from 20,000 commonly used words to infinite variations. In order to acquire the lexicon, the child must separate prosodic information, identifying the word using only syllabic information. This means that speech can be recognized (not necessarily understood) without prosodic information. There may be cases where mistakes are made due to lack of context, but for the majority of cases, isolated word recognition is quite reliable. Prosodic effects are nonlinear and affect the speech waveform on many levels. They are involved in the anticipation and coarticulation effects as well as in emotional and stress-related effects. However, it is possible to remove them as infants have proven.

[11] Altmann, pp. 240 - 244.

[12] ARPA, p. 315.

Such a removal would drastically simplify the speech waveform and reduce the amount of information a speech recognition system must process.

At this stage, word matching is possible against entries already present in the lexicon. Assuming that the word exists in the lexicon, a match will help to reinforce or verify the segmentation and can also help identify word boundaries. Matches also help reinforce entries in the lexicon. Entries that have not been accessed for long periods of time take longer to find, if they are found at all. The ability to forget is a concept that has not been captured by many computer algorithms but is uniquely suited to dynamic language acquisition. If a match is not found, either the speech waveform was improperly segmented or the word is new. When building a new lexicon, most unknown words are new and should be added to the lexicon. After a certain age (around 8), the segmentation process has matured and excludes certain structures. This effect is manifested by difficulties in learning new languages which involve different syllabic units. Another effect is an increasing resistance to learning new words.

While the ability to segment speech waveforms is present at birth, the process is modified by experience. The question of genetics vs. environment in this case is clear. Humans are genetically endowed to segment incoming speech waveforms, but if these are not provided in the environment, the ability is lost. This environmental adaptation aids in increasing the efficiency and reliability of the segmentation process. If a syllabic unit never occurs in the environment, time is wasted searching. More importantly, if the unit is similar to other "neighboring" units, possible confusion can be avoided, increasing the probability of correct identification even in the presence of noise.

Finally, prosodic information, semantics, and syntax are used to remove ambiguities. However, it is important to note that this information is not vital to speech recognition, only to speech understanding. Semantics and syntax are misused in everyday situations, but in most cases, the spoken words are still understood. Meaning often takes longer to sort out.

3.3 The Brain

The brain is the most complex organ in the body. It contains billions of neurons, each of which is capable of connecting to multitudes of other neurons through synapses. A synapse serves as a connection between the axon of one neuron and the dendrite of another. The brain remains the biggest mystery in human physiology. How the lexicon is stored in the brain is unknown. Some theories maintain that words are stored in static locations, while other theories hypothesize that memory, in general, is dynamic. While how memory is stored and accessed is unknown, it is known that language processing occurs in a specific region of the brain. This region of the brain accesses the lexicon in parallel.

Recent studies have shown that word identification is performed on a syllable by syllable basis.[13] When the initial syllable is presented to the language center, a parallel search is initiated for any words with matching initial syllables. Matching words are activated by the search and further comparison is allowed; words that do not match are excluded from further analysis. Studies have shown that even rhyming words that differ only in the first phoneme are excluded from further analysis. It is currently unclear whether subsequent syllables must match or if mistakes are allowed. However, recent evidence suggests that only exact matches are allowed and that higher level processing involving semantics and syntax is used to correct mistakes made in initial segmentation.

Cognitive psychology provides further information about the brain through indirect experimentation. To test the effect of word priming, subjects are instructed to listen carefully to words presented to the right ear and to totally disregard the left ear. Subjects are further instructed to write down words presented to the right ear. During the coarse of the experiment, the subject is presented with several ambiguous words such as "sea" and

[13] Altmann, pp. 150 - 151.

"see". Just before these words are presented to the right ear, a primer is presented to the left ear such as "ocean." While subjects report that they did not hear the primer, they repeatedly choose the associated word.[14] These results reveal the inner workings of prosody and other higher order processing going on in the brain's language center. While a word can be identified without additional information, the context provided can make a difference in continuous speech recognition and speech understanding.

3.4 Sound Production

The human speech production system consists of a primary acoustic tube, the vocal tract, and sometimes an auxiliary acoustic tube, the nasal cavity. The acoustic tube starts with the vocal cords and ends with the lips. The shape of the tube is determined by the position of the velum, tongue, jaw, and lips.[15] A cross-section of the human vocal apparatus is shown in Figure 3.2. The shape of the vocal tract at any given time determines the resonant frequencies or formants that create an envelope around the resultant sound. "For an adult male speaker, the fundamental frequency typically ranges from 100 to 160 Hz … for an adult female, the average fundamental frequency is nearly twice as great."[16]

Sound is produced by exciting the acoustic tube with an appropriate source. For voiced sounds, the vocal cords vibrate, producing quasi-periodic pulses of air pressure. Unvoiced sounds do not use the vocal cords and are best characterized as random noise. For fricatives, the vocal tract is constricted at some point, causing a turbulent air flow which creates the sound. Plosives rely on a build up of air at the lips, which is released in a rush.

[14] Massey.
[15] Oppenheim, pp. 119 - 120.
[16] Cooper, p. 7.

Before the first sound is uttered, the concept being expressed is coded into language. Generally, speech is a continuous operation with multiple words spoken in a single breath. As the brain stimulates the vocal system, it anticipates word boundaries, merging or slurring them together. Depending on many factors including emotion and rate of speech, the boundaries between words can be lost altogether. In some cases, phonemes or syllables within a word are merged; this phenomenon is called coarticulation.

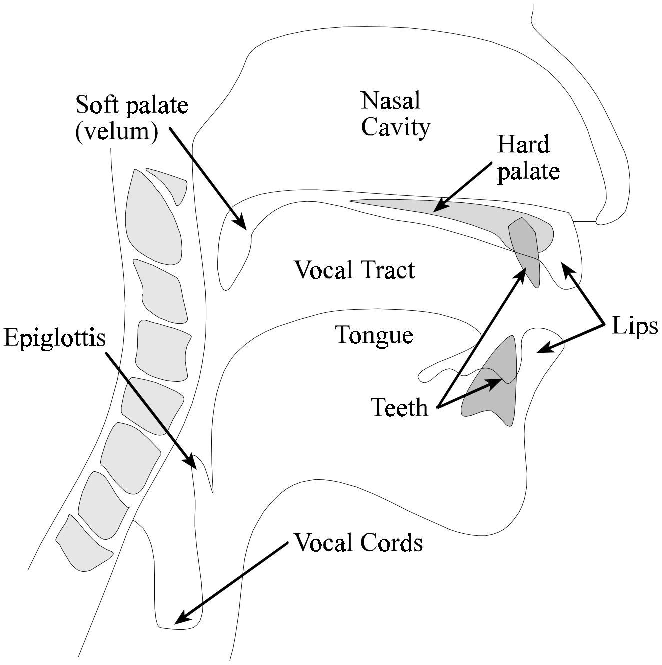


Figure 3.2: Cross-section of the Human Vocal Apparatus
The vocal tract can be viewed as an acoustic tube starting at the vocal cords and terminating at the lips. The nasal cavity provides an auxiliary tube in some cases.

Not all sounds are appropriate for language. The vocal tract is a nonlinear system. For many sounds, slight changes in the vocal tract and tongue produce large changes in acoustic output. One of the fundamental properties discovered by speech researchers over the years is that no two speech waveforms are exactly the same. This is true even when the same person is talking. The problem is that a speaker makes small errors in reaching articulatory targets. Language psychology hypothesizes that languages tend to select areas of acoustic stability, called plateau regions, for sounds. In plateau regions, small errors produce minimal differences in the resultant sound.[17] This hypothesis has been verified for vowels.[18] Future studies could provide important information for the development of speech recognition systems.

3.5 Summary

Information gained from human physiology and psychology can aid in the development of machine speech recognizers. Studies of the ear shows that the brain receives information about the speech waveform in both time-amplitude and time-frequency forms. The ear is also phase insensitive so the speech recognition system can reduce the amount of data needed to encode the speech waveform by half. Small changes in frequency are not noticeable; furthermore, speech tends to concentrate in plateau regions where small errors in production have minimal effects on the resultant sounds. These facts add credence to the use of fuzzy logic to perform speech recognition, as will be further explained in Chapters 5 and 6.

Studies of the brain show that word processing is a serial process, while word access is a parallel process. Again, the fuzzy paradigm seems appropriate since fuzzy rules are evaluated in parallel, while word processing would always be a serial process. Furthermore, initial syllable identification is paramount, as it directs the entire search. This information is useful for continuous speech recognition systems, which can perform similar searches. However, for isolated speech recognition, prosodic information can be ignored. Without prosodic information, the recognizer has no way to correct mistakes made in initial syllable identification. Therefore, the recognizer must check words that differ in initial syllable.

[17] Lea, p. 46.

[18] Cooper, pp. 9 - 10.

By examining the human auditory system, several insights have been gained that allow substantial reductions in the size of the encoded waveform. For example, the human auditory system can hear sounds from 16 Hz to 20 kHz, but most of the information necessary for intelligible speech is contained below 3 kHz. Therefore, speech can be encoded with a minimum sampling rate of 6 kHz. These insights also aid in simplifying the algorithms used to implement a speech recognition system. The next chapter reviews several classical speech recognition systems and presents several algorithms that extract features from the speech waveform.

CHAPTER 4
SPEECH RECOGNITION MODELS

4.1 The Early Days

The following history is a summary of Wayne Lea's "History of Speech Recognition."[19] The earliest known speech recognition system was a toy dog, "Radio Rex." The dog would jump when it detected its name being spoken. The detection system was extremely simple, and the dog would jump for many other words. The development of the spectrograph in the late 1930s prompted more serious attempts. In 1950, Dreyfus-Graf developed an analog filter-based system which separated a waveform into six frequency bands. The resulting signals controlled a device similar to an oscilloscope. As the signals changed, a pattern was drawn on the display.

The first complete recognition system was developed at Bell Laboratories in 1952. The system was capable of recognizing the ten digits [0, 9]. This was an analog system that performed simple pattern matching against templates for each of the digits. Matching was performed based on two inputs: a frequency cut and a fundamental frequency estimation. The frequency cut was performed by separating the frequency spectrum into two bands, above and below 900 Hz. The fundamental frequency was estimated by counting the number of zero-crossings. The system boasted an impressive 97% accuracy for a trained speaker. The first phoneme based recognition system, Audrey, was developed in 1958. An incoming signal was segmented into phonetic units by searching for specific stored spectral patterns. This system also boasted near perfect accuracy for a trained speaker.

[19] Lea, pp. 59 - 77.

4.2 Computer Based Speech Recognition

The first use of a digital computer for speech recognition occurred in 1960. Denes and Matthews introduced the concept of time normalization. Time normalization allows waveforms of differing lengths to be automatically compressed or expanded to match the stored template. The use of time normalization brought speech recognition one step closer to speaker independence. The use of digital computers also allowed researchers to experiment with larger vocabularies and more difficult recognition algorithms.

One of the biggest movements in speech recognition occurred in 1971 when ARPA commissioned the Speech Understanding Research (SUR) project. The project specified the development of a system that could accurately (over 90%) recognize continuous speech from multiple speakers. Speech was to be constrained to a specific grammar using a 1000 word vocabulary. When the project was completed in 1976, one system, HARPY, developed at Carnegie-Mellon University, met or exceeded all the specifications. HARPY accepted sentences from 3 male and 2 female speakers, had a vocabulary of 1011 words, required only 20 training sentences per speaker, and sustained 95% semantic accuracy even in relatively noisy environments. Developments from the ARPA SUR project spawned a multitude of new techniques and brought serious attention to the field of speech recognition.

4.3 Speech Recognition Techniques

Models of speech recognition can be divided into four primary categories: acoustic signals, speech production, sensory reception, and speech perception. The first category views the speech waveform as a general acoustic signal. Recognition is achieved by applying digital signal processing techniques to identify the input. Speech production models the human vocal system in order to determine how the signal was produced and thereby identify the input. The sensory reception viewpoint suggests recognition through biological models of the ear, nervous system, and auditory sections of the brain. Finally, the speech perception model relies on psychological information to extract important features and to categorize the input. Each model comes from a particular scientific discipline.[20] The most successful speech recognition systems employ techniques derived from all four models. However, the majority of current techniques originate from the acoustic signal and speech production models.

The speech waveform is nonlinear and variant. However, over short periods of time (≈ 10 to 30 ms), the waveform remains roughly invariant. Techniques for dealing with the waveform over short periods of time can be broken into two types: time domain and frequency domain. Time domain analyses view the speech waveform as a function of time and amplitude. Frequency domain analyses view the speech waveform as a function of time and frequency. Using these two representations, the relevant features in the speech waveform can be isolated and the word identified.

The amplitude of the speech waveform for voiced speech is much greater than for unvoiced speech. Therefore, the maximum peak amplitude during an interval can be used to discern between voiced and unvoiced speech. The maximum peak amplitude is often taken as a simple indication of the amplitude of the entire sample. The time between corresponding peaks is equal to the fundamental period (or pitch period) for voiced speech. Using this value, an estimate for the fundamental frequency or pitch of the interval can be computed.

The problem with peak measurements is that there may be several similar peaks in a given interval. A better estimate for the fundamental period of voiced speech can be obtained by counting zero crossings. A zero crossing occurs when:

$$\text{sign}\,[x(n)] \neq \text{sign}\,[x(n+1)]$$

[20] Lea, p. 42.

where x(n) and x(n+1) are samples in the current interval. After counting the number of zero crossings in a given interval, the fundamental frequency f_o can be estimated by:

$$f_{\circ} = \frac{n_c}{2t}$$

where n_c is the number of zero crossings and t is the time interval in seconds. For an interval of 10 ms the minimum frequency that can be detected is 50 Hz, for an interval of 30 ms the minimum frequency is approximately 16 Hz. Thus, the interval period determines the accuracy of a given estimate; however, the signal sampling rate determines the ultimate resolution of the zero crossing measurements.

Zero crossings can also be used to discern between voiced and unvoiced speech. Voiced speech tends to be concentrated below 3 kHz, while unvoiced speech, especially fricatives, are generally above 3 kHz. Therefore, when the zero crossing rate is high, the implication is unvoiced; if the zero crossing rate is low, the implication is voiced.

Zero crossing measurements have been used in many speech recognition applications due to their simple implementation and speed. However, zero crossing is especially susceptible to noise, dc offset, and 60 Hz hum. Low pass filtering is usually necessary to remove frequencies below 60 Hz before counting zero crossings.[21] The fundamental frequency of voiced sounds is affected by prosodic information such as stress and intonation as well as syntactic information. The effects of this additional information can cause problems with the estimate derived by zero crossing measurements.

Using an energy representation eliminates many of the problems introduced by prosodic information. For a varying signal such as speech, the energy of the signal can be defined as the convolution:

$$E(n) = \sum_{m=0}^{N-1} [w(m)x(n-m)]^2$$

$$or$$

$$E(n) = \sum_{m=0}^{N\text{-}1} \left| w(m)x(n\text{-}m) \right|$$

[21] Waibel, p. 54.

where w(m) is a window function, the current interval contains N samples, and x(n) denotes the sample value at interval n. The window function weights the samples so that past samples have less importance than more recent samples. Picking an appropriate window function helps reduce variance in the energy calculation over time. The second form of the equation reduces the measurement's sensitivity to amplitude levels.

Like peak and zero crossing measurements, energy measurements can be used to separate voiced speech from unvoiced speech. High values imply voiced speech while low values imply unvoiced speech. Also, when the quality of the speech waveform is high, the energy measurement can be used to separate unvoiced speech from silence.

Short-time autocorrelation analysis can be used to show structure in the speech waveform and also to estimate the pitch period. Short-time autocorrelation is defined as:[22]

$$\varphi_l(m) = \frac{1}{N} \sum_{n=0}^{N'-1} x(n+l)x(n+m+l)$$

where *l* is the beginning of the interval, m denotes the lag ($0 \le m \le M_0$), and N' is either N if data outside the segment is to be used or N-m if only data within the segment is to be used. In the latter case, a weighting function is often used to smooth the interval ends to zero. Autocorrelation identifies periodicity in the speech waveform; therefore, the interval period should be increased to at least twice the period used for other techniques (20 ms to 60 ms).

When estimating the pitch period, recent research has shown that removing the middle of the waveform helps reduce the sample-to-sample correlation of the signal. This nonlinear technique, called clipping, can be efficiently performed by shifting the samples to remove lower amplitude information.[23]

The most common frequency domain techniques include the Fast Fourier Transform (FFT) and Linear Predictive Coding (LPC). The Fourier Transform converts a time-

[22] Waibel, pp. 54 - 55.

[23] Waibel, pp. 54 - 55.

amplitude signal to a time-frequency signal. The FFT is a computationally efficient implementation of the discrete Fourier Transform. While more computationally efficient than a naive implementation, even the best FFT implementations require n $\log_2(n)$ computations.

LPC is based on a model of the vocal tract. LPC assumes "that a sample of speech can be approximated by a linear combination of the past *p* speech samples."[24] By minimizing the difference between the actual signal and the predicted values, the coefficients of the predictor can be adjusted to accurately reflect the sample. If the signal x_n is approximated by:

$$x'_n = \sum_{k=1}^{N} a'_k x_{n-k}$$

then the total squared error is given by:

$$E = \sum_{n=N}^{n-1} (x_n - x'_n)^2$$

The autocorrelation function is given by:

$$R_n = \frac{1}{N} \sum_{i=1}^{N-|n|} x_i x_{i+|n|}$$

The linear prediction coefficients can be computed using the recursive algorithm shown below.[25] Note that the algorithm requires N^2 computations.

[24] Waibel, p. 60.

[25] Chen, pp. 94 - 95.

$$E_\circ = R_\circ$$

$$\alpha_n = \frac{R_n - \sum_{i=1}^{n-1} a_i^{(n-1)} R_{|n-i|}}{E_{n-1}} \qquad \text{for } 1 \leq n \leq N$$

$$A_n^{(n)} = \alpha_n$$

$$A_i^{(n)} = A_i^{(n-1)} - \alpha_n A_{n-i}^{(n-1)} \qquad \text{for } 1 \leq i \leq n-1$$

$$E_n = (1-\alpha_n^2) E_{n-1}$$

$$a_n = a_n^N \qquad 1 \leq n \leq N$$

Once the spectrum has been obtained, it can be analyzed to determine the formant frequencies and the pitch of the speech waveform. While analyses performed in the frequency domain are more accurate than those performed in the time domain, the computational overhead is quite high even by today's standards. Real-time speech processing is not possible on a personal computer without specialized hardware. However, in the coming years, it will be possible to take advantage of these techniques with virtually no performance degradation.

4.4 Time Normalization

Given two representations of the same spoken word by the same speaker under similar conditions, it is highly probable that they will be of different lengths. The main problem is that variations in speaking rate cause nonlinear changes on the time axis. Dynamic Programming (DP) is one technique that attempts to optimally eliminate timing differences between two waveforms. The algorithm works by warping one waveform onto the axis of the other. However, rather than merely stretching or compressing the waveform, the algorithm attempts to match the waveforms so that similarities are maintained and time aligned.

In the classic algorithm, one of the waveforms is warped onto the time axis of the other. However, recent research has shown that mapping both waveforms onto a new common axis performs much better. The flowchart for symmetric dynamic programming is shown in Figure 4.1 for two waveforms, A and B, having I and J samples respectively.[26]

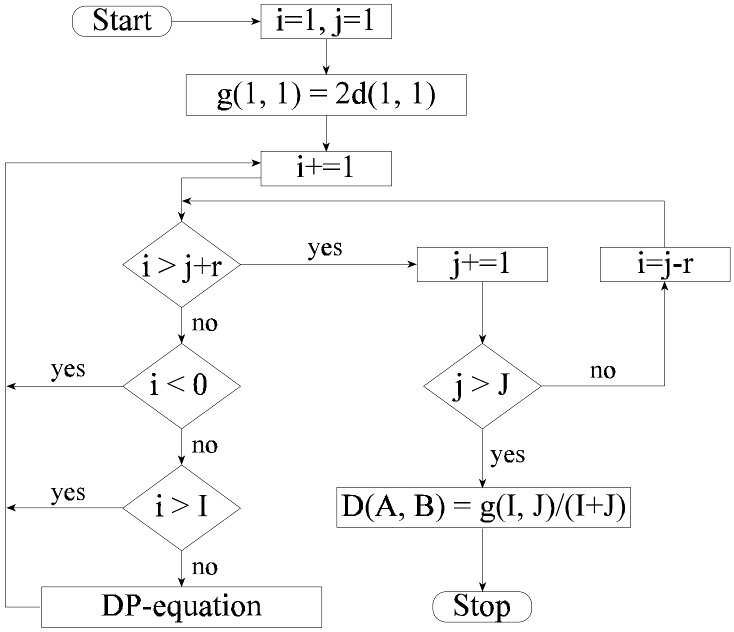


Figure 4.1: Symmetric Dynamic Programming Flowchart
Symmetric dynamic programming creates a common time axis
for two waveforms with differing time axes.

The performance of the DP-equation for symmetric dynamic programming depends on the chosen slope restriction. The slope condition ensures that the warping function has an even gradient. The slope restriction also ensures that the algorithm does not focus on similarities that are outside a window of usefulness. Experimental results show a slope restriction of one to be optimal. Therefore, the DP-equation is:[27]

[26] Waibel, pp. 159 - 163.

[27] Waibel, pp. 163.

$$g(i,j) = \min \begin{bmatrix} g(i-1,j-2) + 2d(i,j-1) + d(i,j) \\ g(i-1,j-2) + 2d(i,j) \\ g(i-2,j-1) + 2d(i-1,j) + d(i,j) \end{bmatrix}$$

where the distance between any two samples is given by:

$$d(i,j) = \left\| a_i - b_j \right\|$$

The DP-equation g(i, j) involves picking the path with the lowest distance and can therefore be viewed as a gradient. With a slope restriction of one, three paths must be examined as shown in Figure 4.2. The slope restriction requires that the function step at most once in the horizontal or vertical directions before stepping orthogonally or diagonally.

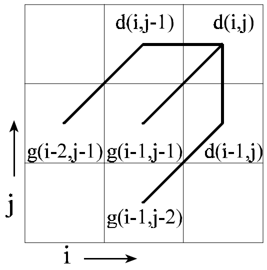


Figure 4.2: Gradient Paths
When the slope restriction is set to one, three paths
must be examined to find the minimum gradient.

The total distance between the waveforms, A and B, is given at the end of the algorithm as D (A, B). This distance can be used directly as a comparison between the input waveform and a stored template. Alternatively, the common axis representation of the waveform after warping can be used for further processing, such as feature extraction.

4.5 Summary

The choice of representation for the speech waveform depends on three factors: processing complexity, information rate, and flexibility. Time domain based representations are simple, fast, and fairly flexible, but lack accuracy for processes not involving time-normalized template matching. Frequency domain representations are complex and slow but extremely flexible and accurate. As the size of a recognizer's vocabulary grows, the potential for confusion increases. Therefore, large vocabulary recognizers may require frequency domain representations to achieve adequate accuracy. To achieve fast results, these systems need specialized hardware to quickly perform the requisite frequency transforms and to handle the increased processing demands of a larger vocabulary. Restricting analysis to the time domain allows speech processing to be achieved in real time on comparatively simple hardware.

CHAPTER 5

FUZZY LOGIC

5.1 Background

Dealing with uncertainty has caused a paradigm shift in science. In the old paradigm, uncertainty was considered unfavorable; science strove for precision and accuracy. However, with the development of molecular physics, uncertainty became an issue that could not be removed. The ability to measure an event or object had a limit that could not be overcome. Therefore, it became necessary to develop new techniques to deal with uncertainty. As these techniques were developed, their usefulness in other fields became apparent.

Fuzzy set theory and fuzzy logic were conceived in 1965 by Lofti Zadeh as a way of allowing uncertainty or vagueness to be represented mathematically. Fuzzy sets are not the first attempt to deal with uncertainty mathematically. Probability theory is capable of dealing with statistical uncertainty, and Aristotelian two-valued logic has been extended to multi-valued logics; however, fuzzy logic is perhaps the most flexible.[28]

5.2 Fuzzy Sets

Fuzzy sets are a super-set of classical sets. Each element in a fuzzy set is associated with a real number which represents the degree of membership of the element in the set. Fuzzy sets are usually expressed as a set of elements having degrees of membership or truth values in the closed unit interval [0, 1]. Fuzzy sets violate several key axioms of Aristotelian sets: the law of the excluded middle and the law of contradiction. This

[28] Klir, 217 - 220.

means that an element of a fuzzy set can simultaneously be both a member and a nonmember of the set. When all elements in a set have either complete membership or complete nonmembership, the fuzzy set reduces to a classical or crisp set.

Fuzzy sets are quite different from statistical models. Probabilities represent the likelihood of a certain outcome given a distribution of past events. Given a probable outcome, there is still a chance that the opposite will occur. Furthermore, after the event has occurred, its probability changes to either 100% or 0% depending on the observed outcome. The elements of a fuzzy set, on the other hand, represent the applicability of the element to the set. While the element may not be totally representative of the set, it at least has some similarity to the concept the set represents. Furthermore, this relationship does not change with time; fuzziness is an intrinsic property of the element. The uncertainty in probability comes from the randomness of the system being analyzed, while in a fuzzy set, it comes from the information source or the inability to completely specify a process or model.

Every fuzzy set consists of three parts: a horizontal domain axis which specifies the set's population, a vertical membership axis which specifies each element's degree of membership, and the surface itself which provides a one-to-one connection between each element and a degree of membership. Fuzzy sets also have a context that indicates how they are meant to be interpreted and utilized. For example, Figure 5.1 shows a simple fuzzy set for the concept *Tall*. This set would not be a good representation if the context was *Skyscrapers*, but is a good representation of *American Women over 20*. Women 4 feet or less have no membership in the set *Tall* while women over 6 feet have total membership. To determine a specific membership, the person's height is found on the horizontal axis, followed to the surface function, and then the degree of membership, μ (x), is read from the vertical axis.

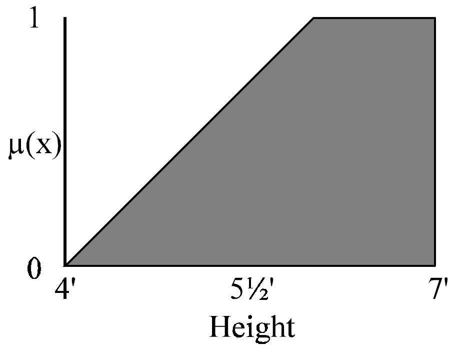


Figure 5.1: A Fuzzy Set Representing *Tall*
The concept tall represented as a fuzzy set for a specific context. Anyone below 4 feet tall has no membership in the set tall, while anyone over 6 feet has total membership. Heights in between are proportionally distributed.

The idea of a fuzzy set representing a concept and having a context is further expanded by linguistic variables. A linguistic variable is assigned to a fuzzy region, a set of fuzzy sets, that represents a complete concept. Figure 5.2 shows an example for the concept *Height*. The variable consists of three fuzzy sets: *Short*, *Medium*, and *Tall*. As with fuzzy sets, context plays a large part in the interpretation of a linguistic variable. The horizontal axis specifies the base variable which is a crisp interval. For a given base value, the degree of membership in each fuzzy set can be determined. This process is called fuzzification.

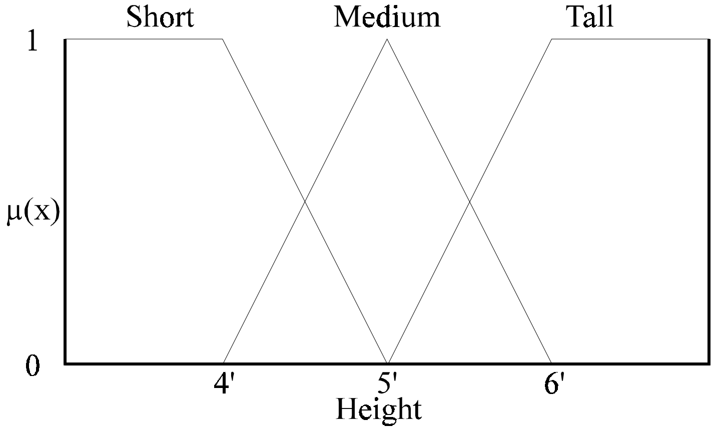


Figure 5.2: A Linguistic Variable Representing *Height*
Linguistic variables are assigned to fuzzy sets representing a single concept.

5.3 Fuzzy Logic

Fuzzy logic is a super-set of classical logic. "Logic is the study of the methods and principles of reasoning in all its possible forms. Classical logic deals with propositions that are required to be either true or false."[29] Fuzzy logic extends the membership of propositions to include a graded membership between no membership, false, and complete membership, true. In the case where a proposition has either no membership or complete membership, fuzzy logic reduces to classical logic.

Fuzzy propositions are created using individual fuzzy sets or groups of fuzzy sets. The power of fuzzy logic comes from its ability to represent concepts lacking clearly defined boundaries. These concepts are generally expressed as linguistic variables. Instead of developing propositions which depend on the mathematical variables, fuzzy propositions use linguistic variables to express the relationship between concepts. Fuzzy propositions allow systems to be created that reason in a more human fashion and are, therefore, easier to understand and maintain. Due to the direct correlation between a linguistic expression and the system being controlled, rules can be elicited from an expert in vague linguistic terms that might be impossible to represent in mathematical form. For example, a rule in a system to control temperature might read:

```
If TEMPERATURE is WARM and HUMIDITY is HIGH
       then AIR_CONDITIONING is HIGH
```

This rule is quite easy to understand in the context of a home cooling system. A system of such rules might be used to maintain the temperature in a structure while minimizing the amount of energy used. Changes in temperature and humidity cause gradual changes in the amount of air conditioning that must be used to maintain the desired room temperature. Also, it is not necessary to fully understand the theory of heat transfer or to know much about the physical structure of the room being controlled.

[29] Klir, p. 212.

Fuzzy logic is capable of dealing with highly nonlinear systems, time varying processes, or noisy environments. The ultimate goal of fuzzy logic is to provide a framework for reasoning about imprecise propositions. Such reasoning is called approximate reasoning and is used to create fuzzy control systems.

5.4 Fuzzy Systems

A fuzzy system consists of a fuzzification subsystem, a fuzzy inference engine, a fuzzy rule base, and a defuzzification subsystem. Figure 5.3 shows a general fuzzy control system.[30] Given a set of crisp inputs that represent the process' current state, the fuzzification subsystem converts them into appropriate fuzzy sets and determines their degree of membership in those sets. Note that a given input may simultaneously be a member of more than one set within a single fuzzy region. The fuzzy inputs are then used by the inference engine to determine the fuzzy outputs. The inference engine interacts with the rule base and uses the inputs to determine which rules are applicable. The rules are independent, and, therefore, may be evaluated in parallel. The outputs are a set of fuzzy sets defined on the universe of possible outputs. These fuzzy outputs are defuzzified to generate the crisp outputs used to control the process. Many methods exist for defuzzification, but the most popular include: centroid, center of maxima, and mean of maxima.

[30] Klir, p. 331.

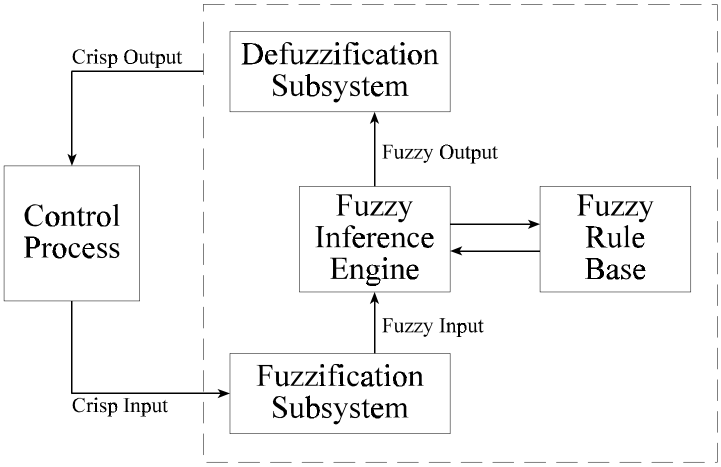


Figure 5.3: A General Fuzzy Controller
Every fuzzy controller consists of at least four major subsystems.
Notice that the Fuzzy Inference Engine is the only subsystem to interact
with the Fuzzy Rule Base. This interaction can occur in parallel.

5.5 Summary

Fuzzy logic allows highly nonlinear, poorly understood, or mathematically complex systems to be modeled reliably and efficiently. Furthermore, fuzzy logic deals well with uncertain and noisy data. These characteristics suggest that fuzzy logic might be an effective tool for speech recognition.

One of the strengths of fuzzy systems is their ability to express a confidence in the reasoning results. Recent studies have shown that fuzzy systems and neural networks are both part of a class of universal approximators of continuous functions. This similarity means that a fuzzy system can be replaced by some form of neural network and vice versa.[31] While much success has been gained by using neural networks in speech recognition, they are unable to express a confidence in the results or explain how the results were generated. Neural networks are generally viewed as black boxes, although current research is attempting to generate techniques for tracing the reasoning process. Fuzzy systems, on the other hand, express a level of confidence in the output by the

[31] Klir, pp. 344 - 349.

degree of membership of the fuzzy outputs. The reasoning that generated the outputs is also available by examining the active rules. The next chapter describes a fuzzy based system that takes advantage of the characteristics of fuzzy systems to perform speech recognition.

CHAPTER 6

ISOLATED DIGIT RECOGNITION

6.1 Digit Waveforms

The majority of isolated word recognizers use a process called template matching to perform recognition. Template matching involves gathering waveforms for each word in the vocabulary. Each template is associated with its name. When an unknown waveform is presented to the system, it is compared to each template to find the best match. If a match is found, that template's name is assigned to the unknown waveform.

Although the recognition system designed for this thesis can perform general word recognition, it was designed primarily to perform isolated digit recognition. Each digit can be categorized by the features discussed in sections 2.3 and 2.4. Since the digits are relatively short, the unit of recognition is a word. However, examining the phonemes and syllables can provide important information.

Table 6.1 shows the phonemes and syllables in each digit. Several possible recognition problems are apparent from the phoneme representation. Digits six and seven and digits four and five start with the same phoneme which increases the probability of confusion between the waveforms. The digits two and three start with a similar phoneme; and since two is quite short, it may be easily confused with three. The digits one, seven, and nine end with the same phoneme which may make them likely candidates for confusion. The digits zero and seven are the only multi-syllable words. Since the primary stress is on the initial syllable, they should both be easier to distinguish. Digits zero and eight are the only truly unique words in the vocabulary and should therefore have a much higher recognition rate. However, eight is an extremely short word, which may cause problems.

Number	Phoneme	Syllable
zero	z i′ r o	ze • ro
one	w Λ n	one
two	t u	two
three	θ r i	three
four	f ɔ r	four
five	f aɪ v	five
six	s ɪ k s	six
seven	s ɛ v′ ə n	sev • en
eight	e t	eight
nine	n aɪ n	nine

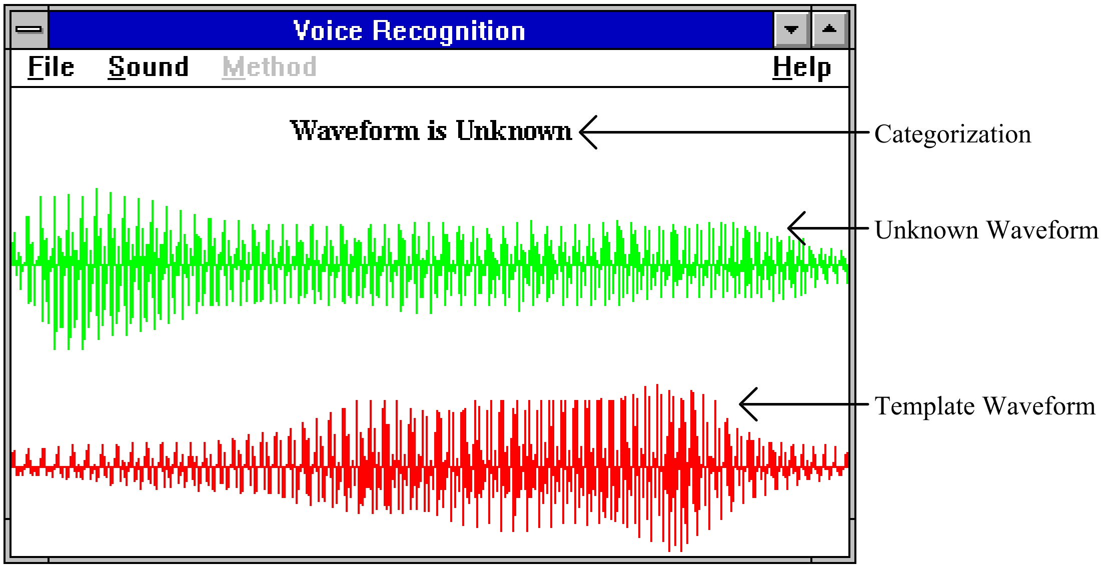
Table 6.1: Digit Representations
The digits 0 - 9 represented as phonemes and syllables.

6.2 Fuzzy Speech Recognizer

The system developed for this thesis uses a personal computer optionally equipped with an 8-bit sound card. Waveforms are stored in the WAVE format using pulse coded modulation (PCM). 8-bit PCM is a signed format with the minimum at 0h, midpoint at 80h, and maximum at FFh. The system reads in the templates for each digit at program startup. After a digit is read, it is normalized so that its minimum is 0h and its maximum is FFh. Normalization helps alleviate the effects of volume variation between samples. Additional modifications, including thresholding, phase removal, and sub-sampling, will be discussed in chapter 7.

Samples for recognition can be either read from a pre-recorded WAVE file or recorded live using the sound card. To ensure consistent frequencies, samples are recorded at the same sampling frequency as the templates. After recording is complete, the system performs segmentation to isolate the word's beginning and ending. At this point, the modified waveform is stored in memory and may be played back to the speaker to ensure that the sample has been correctly recorded and segmented.

After a sample has been loaded or recorded, its time-amplitude graph is displayed at the top of the window. During the analysis process, the unknown waveform is displayed at the top of the window while the template it is currently being compared against is displayed at the bottom, as shown in Figure 6.1. A status bar shows the progress for each template. After the analysis is complete, the result (categorization), is displayed above the unknown waveform.



Waveform is Unknown

Figure 6.1: Voice Recognition System Window
The main window for the voice recognition system during analysis.
The status window has been hidden to completely show both waveforms.

6.3 Symmetric Dynamic Programming

As discussed in section 4.4, some form of time normalization is necessary for general word recognition. Symmetric dynamic programming was chosen due to is past success. While finding the best time alignment between a template and an unknown, dynamic programming also isolates and matches features. The implementation employed for this thesis used dynamic programming on time-amplitude values only; no spectral analysis was performed. While neglecting spectral information limits the overall accuracy of the recognition process, dynamic programming without specialized hardware is too time consuming to perform additional complex analysis such as Fourier analysis or Linear Predictive Coding.

Dynamic programming has a complexity of $O(n^2)$, which can cause unreasonable demands on both processing time and system memory. Fortunately, several constraints can be employed to reduce the complexity. By providing a slope constraint, the search area can be limited. Based on work by Sakoe and Chiba, an optimal slope constraint of one was chosen.[32] Dynamic programming can be viewed graphically as a plot of the template waveform (a) vs. the unknown waveform (b). The warping function defines the optimal path from the starting point, which is always (0, 0), to the ending point, (A-1, B-1) where A and B are the lengths of the template and the unknown waveforms respectively. When the two waveforms are exactly the same, the warping function becomes the diagonal line b = a.

Since the starting and ending points are fixed and the slope is constrained, it is possible to define the area of all possible solutions. The top and left sides are bounded by:

$$y = \frac{x}{2} + (B-1) - \frac{(A-2)}{2} \qquad \text{or} \qquad x = 2(y-(B-1)) + (A-2)$$

$$y = 2x + 1$$

The bottom and right sides are bounded by:

$$y = 2x + (B-2) - 2(A-1) \qquad \text{or} \qquad x = \frac{y-(B-2)}{2} + (A-1)$$

$$y = \frac{(x-1)}{2}$$

Figure 6.2 shows the possible paths that may be taken by the warping function. However, due to the recursive nature of the dynamic programming algorithm, the left and bottom boundaries cannot be utilized since values outside their boundaries may be needed

[32] Waibel, pp. 159 - 165.

for calculating values within the area of possible solutions. Notice that the top and right boundaries correspond to starting at the end point and following the lines of maximum and minimum slope towards the start point. Similarly, the bottom and left boundaries correspond to following the lines of maximum and minimum slope from the starting point and moving towards the end point.

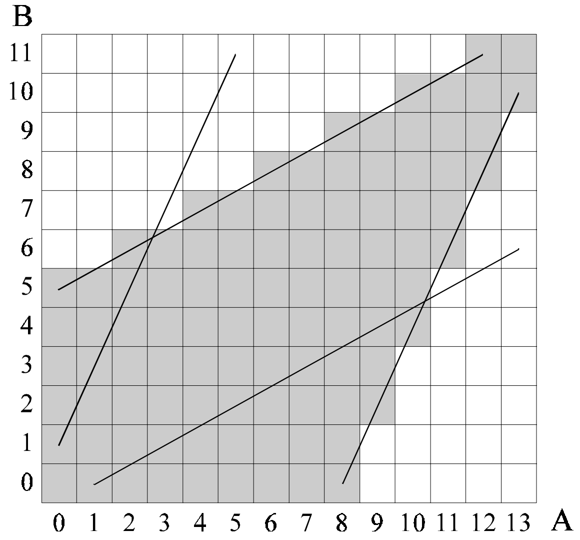


Figure 6.2: Area Searched by Dynamic Programming Algorithm
The four lines bound the region of all possible solutions. The gray region shows the area that must be searched for possible solutions.

The DP-algorithm is perhaps best explained with an example. Figure 6.3 shows the two waveforms to be normalized; waveform A has a length of 10, and waveform B has a length of 15.

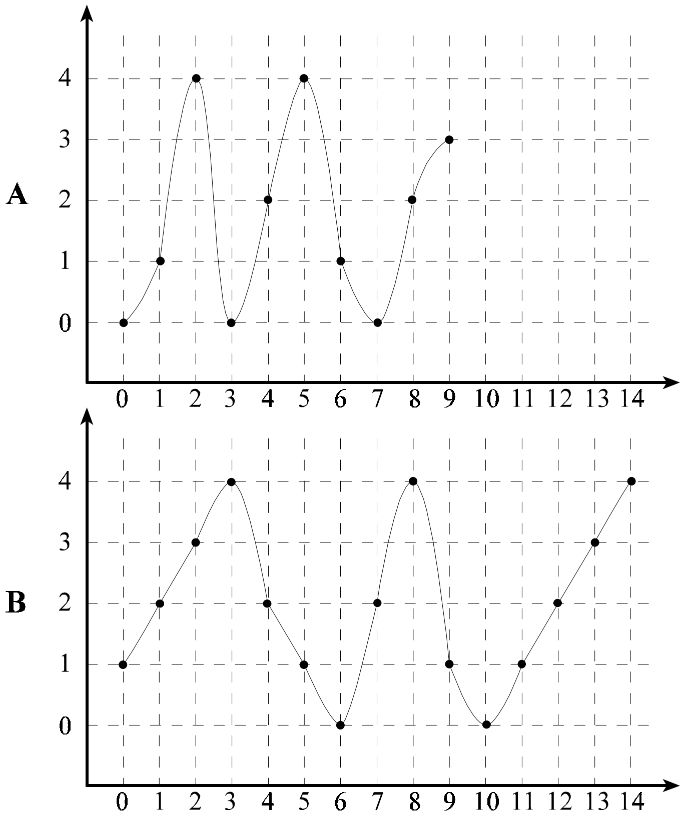


Figure 6.3: Symmetric Dynamic Programming Example
The two waveforms A and B used in the example are shown.

The first step is to calculate the distance from each point in waveform A to each point in waveform B. Table 6.2 shows these distance calculations.

B \ A	0	1	2	3	4	5	6	7	8	9
0	1	0	3							
1	2	1	2	2						
2	3	2	1	3	1					
3	4	3	0	4	2					
4	2	1	2	2	0	2				
5	1	0	3	1	1	3				
6	0	1	4	0	2	4	1			
7	2	1	2	2	0	2	1			
8	4	3	0	4	2	0	3	4		
9	1	0	3	1	1	3	0	1		
10	0	1	4	0	2	4	1	0	2	
11		0	3	1	1	3	0	1	1	
12				2	0	2	1	2	0	1
13						1	2	3	1	0
14								4	2	1

Table 6.2: Waveform Distance Matrix
Each cell contains the distance from a point on
waveform A to a point on waveform B.

The next step is to calculate the recursive gradients. Table 6.3 shows the gradient calculations.

B \ A	0	1	2	3	4	5	6	7	8	9
0	2	0	6							
1	4	4	4	6						
2	6	5	5	9						
3	8	11	6	10	10					
4	4	10	8	10	10					
5	2	4	7	10	11	15				
6	0	4	8	7	9	18				
7	4	2	4	9	7	9	14			
8	8	5	2	6	9	7	10			
9	2	8	5	4	5	10	9	8		
10	0	4	8	4	7	12	8	9		
11			3	6	6	9	12	8	9	
12					6	8	11	12	8	
13							10	15	9	8
14									16	9

Table 6.3: Waveform Gradient Matrix
Each cell contains the minimum gradient from a
point on waveform A to a point on waveform B.

The final step is to normalize the gradient at (A-1, B-1) = (9, 14) by dividing by A + B = 25. The final result, 0.36, is the best distance between the two waveforms.

The symmetric dynamic programming algorithm has been implemented in two forms, a classical crisp form and a fuzzy form. The crisp form calculates the distance at each step by taking the absolute magnitude of the difference between the amplitudes of the two waveforms. The total distance measure between the two waveforms is then given by the sum of these differences divided by the sum of the lengths of the waveforms. This final normalization is necessary so that short waveforms and long waveforms have a common distance measure. The best match is chosen by selecting the waveform with the shortest total distance. The next section discusses the fuzzy implementation.

6.4 Fuzzy Symmetric Dynamic Programming

The fuzzy implementation assumes that all waveforms contain uncertainty. This uncertainty comes from speaker variation, waveform quantization, noise, and the inability to completely specify the process of speech recognition. Each amplitude is therefore represented as a fuzzy number. A fuzzy number may be viewed as a set of numbers around a certain interval. For example, integers close to *x* can be represented by a fuzzy set in which the closer a number is to *x*, the higher its membership in the fuzzy set. The fuzzy set must be normal, since *x* must have maximum membership.

During the distance calculation, amplitudes are assumed to vary by some specified amount from the measured values. Various values were tested, but a maximum of ±8 was experimentally determined to give the best results. Therefore, when the absolute magnitude of the distance between two points is computed, it is adjusted to be a maximum of 16 points closer. The minimum is 0. The membership function, shown in Figure 6.4, is:

$$DOM = \left\{ \begin{array}{ll} \dfrac{16-d}{32}+0.5 & 0 \le d \le 15 \\ \dfrac{255-d}{510} & 15 < d \le 255 \end{array} \right\}$$

For distances below the fuzzy distance threshold, in this case 16, the degree of membership is computed on the interval [0.5, 1]. For distances above the threshold, the degree of membership linearly decreases to zero as the distance approaches the maximum.

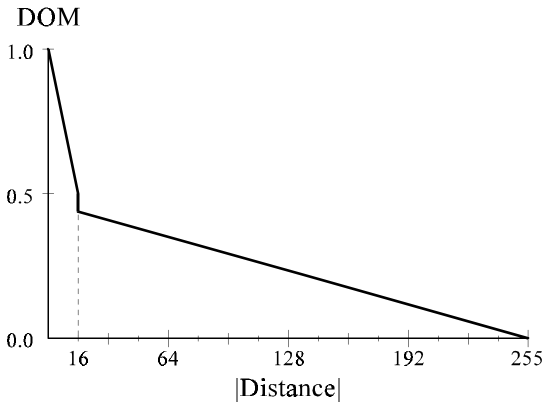


Figure 6.4: Fuzzy Number Membership Function
Values closer to the actual value have a higher degree of
membership, while values further away have lower membership.

The membership of the total distance is the average of the degree of membership of every point along the optimal path. The result of comparing each template to the unknown waveform is a fuzzy set consisting of the set of templates and their degree of membership, which represents the template's similarity to the unknown waveform. To defuzzify the result and select the most likely category, the maximum degree of membership is identified. Then the element from the set of maximum membership with the minimum distance is selected as the best match.

6.4 Summary

Representing amplitude values as fuzzy numbers allows the speech recognition system to deal with the uncertainty inherent in complex systems. In addition to improved accuracy, the credibility of the solution can be quantified. The degree to which the solution should be believed is expressed by its degree of membership. Choosing the solution with a maximum degree of membership ensures that the solution is the best possible. Additional requirements can be imposed to ignore solutions below a certain threshold of credibility.

The next chapter presents the results of tests using the system described in this chapter. Tests under varying conditions, multiple speakers, and utilizing various thresholds and values are presented. The source code for the system is presented in Appendix A.

CHAPTER 7
SYSTEM RESULTS

7.1 System Configuration

Multiple tests were conducted to determine the system's performance under varying conditions. In each test, templates and samples were recorded using a low-cost unidirectional microphone in a noisy environment. The sampling frequency was 6 kHz with 8 bits per sample.

Initial tests were conducted to test the usefulness of removing a waveform's phase by inverting all negative samples. While this procedure makes sense from a biological point of view, the recognition system suffered decreased discrimination ability. Therefore, in all further tests the waveform's phase was not removed.

Clipping procedures were also tested. Clipping is a nonlinear technique whereby samples below a specified value are set to zero. Various values were tested up to a quarter of the maximum amplitude, 32. Clipping was predicted to improve accuracy by removing small low-volume variations in the waveform such as background noise and 60 Hz hum. However, results showed that the system again suffered from reduced discrimination, so no further clipping was performed.

After recording a template or unknown waveform, the system performs segmentation to isolate the word. Various threshold values used by the segmentation routine were tested. The *threshold maximum* determines how loud a sample must be to indicate the start or end of the word. When processing the waveform, the segmentation routine searches for the first sample that exceeds the threshold. The *threshold minimum* determines how loud a sample must be to considered part of the word. Once the threshold maximum is found, the routine searches backward until a sample below the minimum is found. The isolated word is then normalized and stored in memory.

7.2 Initial Trial

The first full scale trial was performed using a trained male speaker. The *threshold maximum* was set to 16, and the *threshold minimum* was set to 4. Template waveforms for each of the ten digits were recorded and saved first. Then for each digit, ten sample waveforms were recorded and saved. The system was then setup to analyze all 100 "unknown" waveforms using first the crisp classification algorithm and then the fuzzy classification algorithm. The results are shown in Table 7.1.

	Crisp			Fuzzy		
Digit	% Correct	Error	Chosen Digits	% Correct	Error	Chosen Digits
0	0	3.41	6	80	0.005/0.74	6
1	20	1.21	1, 5, 6	40	0.011/0.25	7, 9
2	20	1.88	6	90	0.003/0.17	7
3	0	3.80	6	30	0.013/0.71	2, 6
4	0	3.28	6	90	0.001/0.34	6
5	0	4.68	6	40	0.005/0.80	6, 7, 9
6	0	4.24	0, 4, 5, 6	100	0.0/0.0	
7	90	0.05	5	100	0.0/0.0	
8	0	3.34	0, 6	20	0.009/1.08	0, 2, 3
9	10	1.49	6	100	0.0/0.0	
Total	14	2.75		69	0.005/0.41	

Table 7.1: Initial Results with a Trained Male Speaker
Ten unknown samples for each digit were classified. The system accuracy for each digit and the total accuracy are shown for both crisp and fuzzy techniques. The Chosen Digits column shows the results of misclassifications.

Error values represent the average distance from the correct classification to the actual classification. Crisp error values were calculated by averaging the absolute magnitude of the differences between the distance from the correct classification and the distance from the actual classification:

$$E_i = \frac{\sum_{n=0}^{9} \left| d'_{i,n} - d_{i,n} \right|}{10}$$

where E_i is the Error for digit i, d' is the distance for the correct classification for digit i and d is the distance for the actual classification made by the system for digit i. For the results using the fuzzy algorithm, the error is expressed as the degree of membership error and the distance error. Both errors are computed the same way as the crisp error.

Analysis of the crisp results reveals that when the system misclassified a waveform, it classified it as the waveform *Six* 83% of the time. Examination of the templates revealed that the waveform for *Six* was on average 2.25 times shorter than the other templates. Even though distance calculations are normalized to reduce the effects of template length, an extremely short template, relative to the lengths of the other templates, will predispose the system to that classification. The fuzzy results show a more even distribution among misclassifications.

The waveform for the template *Six* was much shorter than the other templates because it begins and ends with the unvoiced phoneme s. Unvoiced phonemes have a much lower relative volume than voiced phonemes. Since the *threshold maximum* was set to a relatively high value, the beginning and ending phoneme were almost completely removed from the template. In order to remove the classification bias, a much lower *threshold maximum* needs to be used.

7.3 Additional Trials

Using the same trained male speaker as in the initial trial, a new set of template and sample waveforms were recorded. The *threshold maximum* was set to 8, and the *threshold minimum* was set to 0. The results are shown in Table 7.2.

	Crisp			Fuzzy		
Digit	% Correct	Error	Chosen Digits	% Correct	Error	Chosen Digits
0	100	0.0		100	0.0/0.0	
1	60	0.29	5	100	0.0/0.0	
2	40	0.55	3	100	0.0/0.0	
3	0	1.32	2	40	0.008/0.52	2
4	10	1.56	0	30	0.016/1.27	0
5	20	1.66	8	70	0.01/0.28	6, 8
6	100	0.0		100	0.0/0.0	
7	0	2.81	8	100	0.0/0.0	
8	100	0.0		100	0.0/0.0	
9	20	1.10	5, 8	80	0.002/0.14	1, 6
Total	45	0.93		82	0.004/0.22	

Table 7.2: Results with a Trained Male Speaker
Ten unknown samples for each digit were classified. Each template and unknown were segmented using a lower *threshold maximum* to remove classification bias.

The results show a dramatic increase in accuracy for the crisp algorithm. Classification accuracy of all waveforms except *Three* and *Seven* increased. For the crisp algorithm, the system confused the waveform *Two* with the waveform *Three* and visa versa. The system also confused *Zero* with *Four*, *Eight* with *Five*, and *Five* and *Eight* with *Nine*. Interestingly, recognition for *Seven* dropped from 90% to 0%; this result is quite unexpected since seven is a two syllable word and has one of the most unique and consistent waveforms. For the fuzzy algorithm, the system most frequently confused *Two* with *Three* and *Zero* with *Four*. It is interesting to note that the template for *Zero* was longer than the template for *Four* (*Two* was the shortest template).

Another trial was conducted with an untrained female speaker. The segmentation thresholds were the same as with the previous male speaker. The results are shown in Table 7.3. The system correctly classified all 100 waveforms using both the crisp and fuzzy algorithms.

	Crisp			Fuzzy		
Digit	% Correct	Error	Chosen Digits	% Correct	Error	Chosen Digits
0	100	0.0		100	0.0/0.0	
1	100	0.0		100	0.0/0.0	
2	100	0.0		100	0.0/0.0	
3	100	0.0		100	0.0/0.0	
4	100	0.0		100	0.0/0.0	
5	100	0.0		100	0.0/0.0	
6	100	0.0		100	0.0/0.0	
7	100	0.0		100	0.0/0.0	
8	100	0.0		100	0.0/0.0	
9	100	0.0		100	0.0/0.0	
Total	100	0.0		100	0.0/0.0	

Table 7.3: Results with an Untrained Female Speaker
Ten unknown samples for each digit were classified. Each template and unknown were segmented using a lower *threshold maximum* to remove classification bias.

In an attempt to explain the perfect results of the female speaker, crisp template correlations for each speaker were calculated. The results for the male speaker are shown in Table 7.4; the results for the female speaker are shown in Table 7.5. The template correlations are symmetric.

Digit	0	1	2	3	4	5	6	7	8	9
0	0	11.73	14.78	13.15	12.03	13.11	15.57	13.40	14.13	16.04
1		0	13.89	11.20	11.69	10.68	14.31	9.88	13.30	12.54
2			0	12.61	15.17	12.49	15.46	15.21	13.73	12.59
3				0	10.99	10.14	13.43	12.09	10.14	11.38
4					0	10.95	12.13	10.27	9.40	14.54
5						0	10.02	10.75	9.01	12.47
6							0	11.73	7.48	16.49
7								0	11.86	12.97
8									0	10.81
9										0

Table 7.4: Crisp Template Correlation for Male Speaker
Each cell shows the distance between a pair of templates.

Digit	0	1	2	3	4	5	6	7	8	9
0	0	9.99	8.54	9.17	11.38	7.55	10.30	8.32	11.81	10.42
1		0	9.43	9.50	14.08	6.95	7.22	6.51	10.09	8.14
2			0	5.88	13.85	8.43	9.07	9.65	11.80	12.16
3				0	13.44	12.61	10.20	11.88	12.60	15.99
4					0	13.26	14.68	14.28	18.28	17.61
5						0	11.46	10.71	8.71	8.93
6							0	6.46	6.52	11.60
7								0	7.91	9.36
8									0	10.35
9										0

Table 7.5: Crisp Template Correlation for Female Speaker
Each cell shows the distance between a pair of templates.

The correlation results, while interesting, do not provide any revealing answers to the question of why the female speaker's results were so impressive. On average, the male speaker’s templates were more diverse, which should indicate that they would be better for classification. One possible answer is that the female speaker was British, which may predispose her to speak clearly and accurately. More trials are needed to determine which factors most prominently affect the system.

7.4 Sub-Sampling Trials

Using the templates and a subset of the sample waveforms from the initial trial set, sub-sampling trials were performed to determine how well the crisp and fuzzy algorithms respond to information loss. Given a waveform recorded at a specific sampling frequency, sub-sampling involves finding the maximum (or peak) value from a group of samples within a specific time period. For example, the template and sample waveforms were sampled at 6 kHz. Sub-sampling at 1000 Hz involves finding the maximum value from every 6 samples; sub-sampling at 100 Hz means taking the maximum value from every 60 samples. Table 7.6 shows the results of sub-sampling.

The crisp system tends to be erratic and the results are not consistent as the degree of information loss increases. The fuzzy system is much more tolerant to information loss, and degrades well as the degree of information loss increases. These results were expected, and confirm the assertion that fuzzy systems are able to work effectively even in the presence of uncertainty.

Sub-Sampling	Crisp	Fuzzy
Frequency (Hz)	% Correct	% Correct
1000	20	75
400	10	50
200	35	45
100	35	25

Table 7.6: Sub-Sampling Results
The Table shows the results of sub-sampling a various frequencies.

7.5 Summary

The accuracy of the fuzzy speech recognition system is guaranteed to be at least as good as the crisp system; however, as the results show, much better accuracy can generally be achieved. As the results with the female speaker showed, some speech factors can affect recognition accuracy dramatically. Isolating these features and training speakers to exploit them will aid future recognition systems.

CHAPTER 8
CONCLUSIONS

8.1 Suggestions for Future Work

Most importantly, more trials should be conducted. Various segmentation threshold values should be tested for each speaker to determine optimal values for general recognition. Testing recognition using languages other than English could also provide useful insights. Tests with increased vocabulary can be performed to determine the system's ability to scale gracefully.

The ultimate goal of speech recognition is the design of a system capable of recognizing continuous speech from multiple speakers from a large vocabulary. Testing speakers using templates from other speakers should provide results that will aid in extending the system’s ability to recognize speech from multiple speakers. A series of templates could be used in a clustering algorithm to determine an optimal template for each word in the vocabulary. Also, the clustering algorithm could be used to determine the uncertainty inherent in the system and build fuzzy membership functions for optimal recognition.

If higher accuracy is required, spectral analysis can be added to select the most likely classification. To limit the number of spectral analyses that must be performed, only the most likely candidates should be examined.

8.2 Conclusions

This thesis examined the issues involved in designing a simple speech recognition system. The speech waveform was analyzed and its features discussed. Information gained from biological and psychological studies was used to simplify the speech waveform. The highlights of speech recognition history were reviewed and recognition techniques outlined. Basic fuzzy set theory was presented and related to the problem of speech recognition.

The last half of the thesis was concerned with the design of a fuzzy speech recognition system. Analysis of the recognition technique reveals that the use of fuzzy logic can only improve the system's performance. For the male speaker, the fuzzy system performed considerably better than the crisp system. The results also showed that the fuzzy system degraded better than the crisp system as the information's uncertainty increased.

APPENDIX A

CODE

```
[The code for this Thesis has been removed; for more
information please contact the author]
```

GLOSSARY

allophone
A set of phones within a language that contain the same information.

diphone
A transitional sound identified by segmenting adjacent phones at their steady-state centers.

dynamic programming
A pattern matching technique with a nonlinear time normalization effect. The algorithm attempts to optimally eliminate timing differences by warping one waveform onto the axis of another while aligning similarities.

formant
A band of high energy concentrated in a specific frequency range. Multiple formants may exist at harmonics of the base formant. Formants are the result of the natural resonances of the human vocal tract. Formants are generally found in vowels.

fuzzy number
A set of numbers around a certain interval. To qualify as a fuzzy number, the set must satisfy:

1.) The set must be a normal fuzzy set.
2.) ${}^{\alpha}A$ must be a closed interval for every $\alpha \in (0, 1]$ where
    ${}^{\alpha}A = \{x \mid A(x) \geq \alpha\}$.
3.) The support of A, ${}^{0+}A$, must be bounded.

fuzzy set
A set of elements which may have a graded degree of membership between no membership and complete membership.

linguistic variable
A fuzzy set or collection of fuzzy sets which represent a single concept in a particular context.

phone
Minimal unit of speech sound. Physically unidentifiable due to coarticulation and anticipation effects.

phoneme
The total collection of allophones that operate similarly.

syllable
A unit of spoken language containing a vowel, dipthong, and / or a consonant.

time normalization
A technique used to eliminate nonlinear timing differences between two waveforms.

vocalic sound — A sound containing a highly defined formant structure.

voiced sound — A sound in which the vocal cords vibrate. Examples include the vowels and "m", "n", and "r". Sounds like "s", "k", and "p" are unvoiced.

word — A sound or collection of sounds that communicates a complete message.

BIBLIOGRAPHY

Abse, D. Wilfred, MD. Speech and Reason: Language Disorder in Mental Disease. The University Press of Virginia, 1971.

Altmann, Gerry T. M., ed. Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives. The MIT Press, 1990.

Aminzadeh, Fred and Mohammad Jamshidi, editors. Soft Computing: Fuzzy Logic, Neural Networks, and Distributed Artificial Intelligence. Prentice Hall, 1994.

ARPA, sponsor. Human Language Technology. Morgan Kaufmann, 1993.

Bezdek, James C. Analysis of Fuzzy Information: Volume I, Mathematics and Logic. CRC, 1987.

Bezdek, James C. Analysis of Fuzzy Information: Volume II, Artificial Intelligence. CRC, 1987.

Bezdek, James C. Analysis of Fuzzy Information: Volume III, Applications. CRC, 1987.

Bezdek, James C. "Fuzzy Models – What Are They, and Why?" *IEEE Transactions on Fuzzy Systems*, (Volume 1, Number 1, February, 1993).

Blahut, Richard E. Fast Algorithms for Digital Signal Processing. Addison-Wesley, 1985.

Burrus, C. S. and T. W. Parks. DFT/FFT and Convolution Algorithm: Theory and Implementation. John Wiley & Sons, 1985.

Chen, C. H., ed. Digital Waveform Processing and Recognition. CRC Press, 1982.

Clippinger, John Henry, Jr. Meaning and Discourse: A Computer Model of Psychoanalytic Speech and Cognition. The Johns Hopkins University Press, 1977.

Cooper, William E. Speech Perception and Production: Studies in Selective Adaptation. Ablex Publishing, 1979.

Cox, Earl. "Fuzzy Fundamentals." *IEEE Spectrum*, (October 1992).

Cox, Earl. The Fuzzy Systems Handbook. Academic Press, 1994.

Cullingford, Richard E. Natural Language Processing: A Knowledge-Engineering Approach. Rowman & Littlefield, 1986.

De Mori, Renato. Computer Models of Speech Using Fuzzy Algorithms. Plenum Press, 1983.

Friedman, David H. Detection of Signals by Template Matching. The Johns Hopkins Press, 1968.

Kandel, Abraham. Fuzzy Mathematical Techniques with Applications. Addison-Wesley, 1986.

Karmiloff-Smith, Annette. A Functional Approach to Child Language. Cambridge University Press, 1979.

Klir, George J., and Bo Yuan. Fuzzy Sets and Fuzzy Logic: Theory and Applications. Prentice Hall, 1995.

Lea, Wayne A., ed. Trends in Speech Recognition. Prentice-Hall, 1980.

Ludeman, Lonnie C. Fundamentals of Digital Signal Processing. John Wiley & Sons, 1986.

Massey, Christine. "Cognitive Psychology." Swarthmore College. Fall 1991.

Newell, Allen, et al. Speech Understanding Systems: Final Report of a Study Group. North-Holland Publishing, 1973.

Niemann, H., ed. Recent Advances in Speech Understanding and Dialog Systems. Springer-Verlag, 1987.

Nguyen, Hung T., Michio Sugeno, Richard Tong, and Ronald R. Yager. Theoretical Aspects of Fuzzy Control. John Wiley & Sons, 1995.

Oppenheim, Alan V., ed. Applications of Digital Signal Processing. Prentice-Hall, 1978.

Ross, Timothy J. Fuzzy Logic with Engineering Applications. McGraw-Hill, 1995.

Tatman, Gunhan H. "Real-Time Neural Networks for Speech Recognition." Theis. University of South Carolina, 1992.

Waibel, Alex and Kai-Fu Lee, ed. Readings in Speech Recognition. Morgan Kaufmann, 1990.

Wang, Lih. "Speaker Normalization Expert." Thesis. University of South Carolina, 1985.

INDEX

A
adaptation, 14, 16
allophone, 5, 60
ARPA, 23
autocorrelation, 26
B
brain, 17–18
C
coarticulation, 4, 5, 10, 11, 15, 19, 60
D
diphone, 6, 60
dynamic programming, 28–30, 41–46, 60
    fuzzy, 47–48
E
ear, 12–14, 21
energy measurements, 25
F
FFT, 26
fuzzy logic, 35–36
fuzzy sets, 32–34
fuzzy systems, 36–37
H
HARPY, 23
I
infant language acquisition, 14–16
information theory, 4–5, 11
L
LPC, 26
P
peak measurements, 24
phase, 13, 20, 40, 50
phoneme, 5, 9–10, 19, 39, 60
plateau regions, 20
prosody, 11, 15–16
S
segmentation, 15, 40, 50
speech features, 8–9, 25, 61
speech waveform, 4, 7, 23, 39
sub-sampling, 55–56
syllable, 7, 9, 19, 20, 39, 60
T
threshold, 40, 47, 48, 50
time normalization, 28, 60
W
word, 7
word priming, 17
Z
zero crossings, 25