MULTICLASSIFIER FUSION OF AN ULTRASONIC LIP READER IN AUTOMATIC SPEECH RECOGNITION

THESIS

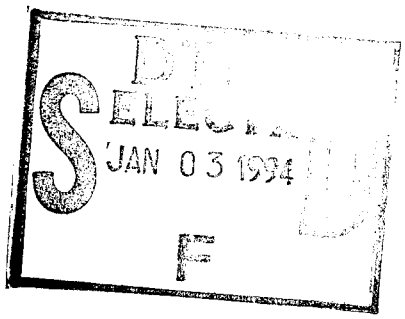
David L. Jennings
Captain, USAF

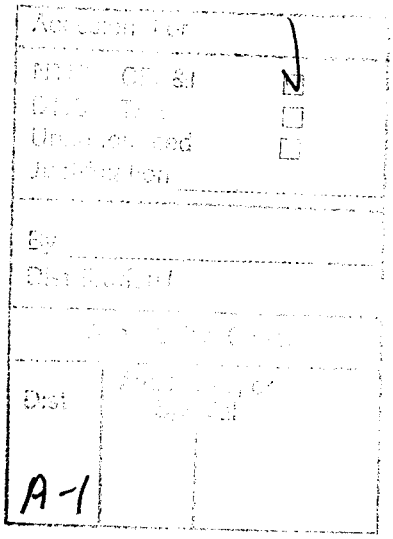
AFIT/GE/ENG/94D-16

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT/GE/ENG/94D-16





MULTICLASSIFIER FUSION OF AN ULTRASONIC LIP READER IN AUTOMATIC SPEECH RECOGNITION

THESIS
David L. Jennings
Captain, USAF

AFIT/GE/ENG/94D-16

Approved for public release; distribution unlimited

AFIT/GE/ENG/94D-16

MULTICLASSIFIER FUSION OF AN ULTRASONIC LIP READER IN AUTOMATIC SPEECH RECOGNITION

THESIS

Presented to the Faculty of the School of Engineering
of the Air Force Institute of Technology
Air University
In Partial Fulfillment of the
Requirements for the Degree of
Master of Science in Electrical Engineering

David L. Jennings, B.S.E.E.
Captain, USAF

December, 1994

Acknowledgements

I want thank my thesis advisor Captain Dennis Ruck, his guidance and leadership were an integral part of this research project. I also want to thank Dr. Marty Desimio for the numerous conversations and discussions we had, they were beneficial to my education and my thesis. I also want to express my thanks to Col. Paul Morton and John Schnurer of Armstrong Lab, who provided the devices and technical support for this thesis, without them this research would not have taken place. I also wish to thank my wife Colleen, who now knows more about lip reading and automatic speech recognition than most engineers. Her patient listening to my research ideas and problems were key to my successful completion of this project. Finally, I want to say thanks to my children, Breanna, Larissa, Chantal, Bethany, Aaron and Joshua, for the time they lost with their father, which allowed me to finish this work.

David L. Jennings

Table of Contents

List of Figures

List of Tables

Abstract

This thesis investigates the use of two active ultrasonic devices in collecting lip information for performing and enhancing automatic speech recognition. The two devices explored are called the "Ultrasonic Mike" and the "Lip Lock Loop." The devices are tested in a speaker dependent isolated word recognition task with a vocabulary consisting of the spoken digits from zero to nine. Two automatic lip readers are designed and tested based on the output of the ultrasonic devices. The automatic lip readers use template matching and dynamic time warping to determine the best candidate for a given test utterance. The automatic lip readers alone achieve accuracies of 65-89%, depending on the number of reference templates used. Next the automatic lip reader is combined with a conventional automatic speech recognizer. Both classifier level fusion and feature level fusion are investigated. Feature fusion is based on combining the feature vectors prior to dynamic time warping. Classifier fusion is based on a pseudo probability mass function derived from the dynamic time warping distances. The combined systems are tested with various levels of acoustic noise added. In one typical test, at a signal to noise ratio of 0dB, the acoustic recognizer's accuracy alone was 78%, the automatic lip reader's accuracy was 69%, but the combined accuracy was 93%. This experiment demonstrates that a simple ultrasonic lip motion detector, that has an output data rate 12,500 times less than a typical video camera, can significantly improve the accuracy of automatic speech recognition in noise.

MULTICLASSIFIER FUSION OF AN ULTRASONIC LIP READER IN AUTOMATIC SPEECH RECOGNITION

I. Introduction

For many years man has dreamed of a computer that he could converse with, such as the computer "HAL" in "2001: A Space Odyssey" or the computer on "Star Trek: the Next Generation." Although we are still a long way from these science fiction speech understanding systems, automatic speech recognition has made tremendous advances in commercial applications. Applications, such as operator assistance systems and personal computer systems, have become possible partly due to the tremendous advances in the processing speed of computers over the past decade and partly by reducing the scope of the speech recognition problem. By limiting the vocabulary, requiring training for each new speaker, or controlling the speakers environment, high accuracy automatic speech recognition is possible today. For example, some labs have reported accuracies greater than 95% with large vocabularies in fluent sentences (28), under low noise conditions. The recent successes of automatic speech recognition have inspired researchers to look for even more applications for the future. One barrier that stands in the way of many applications is noise. Noise is a many faceted problem and already a lot of research has been devoted to dealing with noise (27), but only with limited success. The research of this thesis is specifically directed at the problem of automatic speech recognition in a varied noise environment. This chapter starts off by giving a brief background, leading to the development of the problem statement for this thesis. Then the scope of the research is defined setting the boundaries and goals of the research. Finally a brief description of the unique approach of this thesis is given.

1.1 Background

If the problem of noise in automatic speech recognition could be overcome many applications in adverse noise environments would become possible, such as in the cockpit of an aircraft, in noisy factories and offices, and in cars. At first the tasks would probably be limited in vocabulary and require training for each user, but even with these limitations automatic speech recognition could free the users hands for more important tasks. Unfortunately accuracy in automatic speech recognition today is highly dependent on a consistent noise environment. If the training environment differs from the actual environment results quickly fall off. Figure 1.1 demonstrates the typical performance of automatic speech recognition in noise. These results were based on isolated word recognition with a vocabulary consisting of the digits. As can be seen from the figure, there is some improvement if the reference template has a signal to noise level equal to that of the environment. Therefore, one approach to reducing the impact of noise is to add an estimate of the environment noise to a clean template. Even though using a matching signal to noise ratio offers some improvement, accuracy still falls to an unacceptable level as the signal to noise ratio approaches 0dB; therefore, alternative methods of dealing with noise need to be found. It is well known that humans when faced with such an adverse environment resort to knowledge sources such as syntax and context to maintain communication. Another knowledge source that humans use to improve communication is lip reading. Based on these observations if an automatic speech recognition system could also include these knowledge sources high accuracy results in adverse environments should be possible. The goal of this research is to demonstrate that a simple device can provide lip information necessary for both stand alone automatic recognition or enhance an acoustic recognizer in a varied noise environment.

1.2 Problem Statement

In past efforts to include lip reading in an automatic speech recognition, the primary source of information has been a video image of the talker's mouth (15, 23, 31, 32, 35, 33). In all of these past efforts, improved accuracy was demonstrated by including information

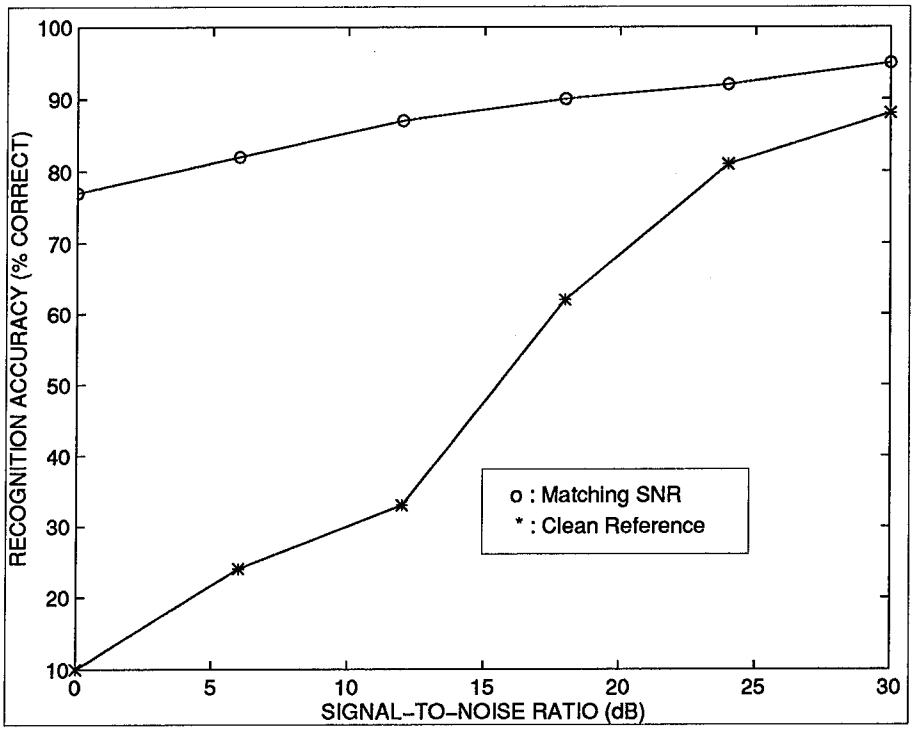


Figure 1.1 Speech recognition performance in noisy conditions, o: Training and testing have matching SNR, *: Only clean reference is used and abscissa indicates test SNR. (27:306)

derived from the video images of the mouth but at a cost. The video images were difficult to acquire and computationally expensive to process. This research specifically investigates two devices developed by Armstrong Lab, Wright-Patterson AFB (18). These devices both use an active ultrasound signal reflected off of the talker's mouth to acquire lip motion information. The benefit of these devices over using video is that they both output a one dimensional signal similar to the output of an acoustic microphone, therefore decreasing the data as much as 12500 to 1 from that of video. Another problem found in all of the past research is a good method to fuse the lip information into the system so as to optimize the performance of the system. In many cases a rule based system or voting scheme was used for integration. This thesis also investigates methods of fusing the lip data both at a feature and classifier level. In one classifier fusion test, at a signal to noise ratio of 0dB, the acoustic recognizer's accuracy was 78%, the lip reader accuracy was 69%, but the combined accuracy was 93%. In classifier fusion the goal is to have a combined system with accuracy as good as or better than the best results of either the automatic speech recognizer or automatic lip reader alone. Considering the background of automatic speech recognition in varied noise and given the equipment from Armstrong Lab, the problem statement of this research can be defined.

> Specifically, the goal of this research is to demonstrate that these simple devices can provide information from the lips of the talker that can be used to conduct or enhance automatic speech recognition. The research therefore considers the design of a both a separate automatic lip reader based solely on the output of the ultrasonic devices and considers fusing this information with a normal acoustic recognizer in varied noise situations. The research considers both feature and classifier level fusion, searching for the highest accuracy design.

1.3 Scope

The system is tested as a speaker dependent system with isolated words. The vocabulary consists of the spoken digits from zero to nine. The system is also tested using basic pattern recognition techniques of template matching and dynamic programming. If the results of this research are successful then subsequent research can extend into harder automatic speech

recognition tasks and employ more advanced pattern recognition techniques using hidden Markov models or neural networks.

1.4 Approach and Methodology

The approach of the thesis is fairly straight forward. First, a basic automatic speech recognizer is designed. This system will be tested under various signal to noise ratios. Second, two automatic lip readers are designed based on the output of the two ultrasonic devices. Finally, two fusion techniques are investigated based on feature level and classifier level fusion. The goal in fusion being to achieve the highest overall accuracy.

1.5 Conclusion

Clearly there are many benefits to an automatic speech recognition system that can operate in a varied noise environment. Even though researchers may never be able to achieve a system that can operate as well as humans, the advantages of even a basic speaker dependent, limited vocabulary system would be tremendous especially in an aircraft or in a car. The approach of this thesis research to this problem is unique both in the way the talker's mouth movement information is collected and in the fusion techniques employed to include this information into the system. The results of this thesis demonstrate that a simple device can provide information to enhance automatic speech recognition and that fusion can improve accuracy in varied noise situations.

1.6 Organization

The rest of this thesis is organized as follows. Chapter 2 presents a detailed literature review including methods to deal with noise, basic lip reading science, past efforts to include lip reading in automatic speech recognition, and ultrasound basics. Chapter 2 also discusses fusion techniques and the hardware and the software used in this project. Chapter 3 describes the design of the automatic lip reader, the automatic speech recognizer, and the combined

systems. Chapter 4 presents the results of tests on the various systems. Finally, Chapter 5 summarizes the results with recommendations for future work in this area.

II. Literature Review

This chapter will review the past efforts in automatic lip reading and the techniques used to extract and fuse the lip information into a traditional, acoustic based, automatic speech recognition system. The first section will discuss the problem of noise in automatic speech recognition in general specifically focusing on the problem of an environment with changing noise. The next section gives some necessary background information on lip reading science. Key definitions are defined and results from studies with humans are presented. Then a detailed review of all the past efforts to include lip reading in automatic speech recognition is given, emphasizing methodology, problems, and results. Next a review of ultrasound principles and dynamics is covered to help explain how the two devices investigated in this thesis function. After that a review of classifier and feature level fusion techniques is given. Finally, a review of the supporting hardware and software, used in the research, is given.

2.1 Dealing with Noise

Noise is a particularly difficult problem in automatic speech recognition. One method of dealing with noise is to use specialized equipment or techniques such as noise canceling microphones or by using two microphones and using signal processing to reduce the noise. These methods do provide some help but accuracy still falls off drastically as the operating environment differs from the training environment. Typical results are depicted in Figure 1.1. These results are based on isolated digit recognition and are similar to the results achieved during this research, using the acoustic classifier alone. Another way to deal with noise is to add an estimate of the noise environment to the clean reference templates before attempting classification. As noted in Figure 1.1 this can improve accuracy, but as the signal to noise ratio falls to 0dB the error rate still becomes unacceptable. No matter the method tried to compensate for noise the accuracy falls off drastically as the SNR approaches 0dB, but we know that humans are capable of much higher accuracy in this environment. To improve accuracy in high noise situations humans typically use knowledge sources such as context

or syntax to maintain communication. Another knowledge source that humans use if it is available is lip reading. In a study done with partially deaf people, Walden (38) demonstrates the benefit humans derive from lip reading. In a consonant recognition task with audio only, the accuracy was approximately 48% and with visual only, around 45%, but with both audio and visual accuracy was near 85%. Other research has shown that seeing the talker's face can be equivalent to increasing the SNR by about 15dB (36). Obviously a 15dB increase in SNR would improve the accuracy in automatic speech recognition. These research efforts demonstrate that humans use lip reading to enhance communication and lip reading can provide information that is, in some cases, not available in the acoustic signal alone.

2.2 *Lip Reading Science*

To clearly understand the best way to acquire lip information for automatic speech recognition, a basic review of lip reading science is necessary. This section will define some basic terms in lip reading science and then present some results of studies with humans. Although the results presented are based on phoneme level recognition, the key points are important in designing and testing an automatic lip reader. Finally this section will present some research into how humans integrate and fuse the visual information in speech recognition.

2.2.1 Definitions and Key Concepts. There are a number of unique terms associated with lip reading. The viseme, or visual phoneme, is the basic unit in lip reading. The relationship of phoneme to viseme is many to one. For example the phonemes B, P, and M are all the same viseme. The exact number of consonant visemes in English, just like the exact number of phonemes, depends on which expert is asked. The number ranges from as few as 4 to as many as 12. All vowels are considered to be separate visemes, but the boundaries between vowel visemes are not clearly defined. At the word level, words that look alike on the lips are called homophenes. For example the words "bat" and "pat" are homophenes. It is estimated that 40-60% of words in English are homophenous, and therefore the maximum accuracy of an automatic lip reader with a large vocabulary would be 40-60% (3), but on a specific limited vocabulary the results could be much higher. Another important finding in

speech reading science is that humans can produce up to 13 sounds a second, but can visually perceive only 8-10 movements a second (10). Therefore an automatic lip reader may be able to acquire important lip information that a human can not.

2.2.2 Consonant Confusion Tables and the Vowel Triangle. Many studies have been done with humans to form confusion matrices of visemes. One of the most detailed studies was done by Berger (3). The results of his work are presented in Tables 2.1 and 2.2. Berger considers 12 unique visemes for beginning of words and 8 for endings of words. All the phonemes are identified using ARPABET as shown in Table 2.3. Clearly the results of Berger's study indicates that some viseme groups are very distinguishable on the lips such as (P,B,M) whereas others such as (TH,DH) are much more difficult to identify. Even though the (TH,DH) group only achieved 39% accurate, this is still significantly better than chance. There have also been numerous studies into vowel visemes. Generally all the researchers agree that each of the vowels represent separate viseme, although the degree of separability is based, as it is in acoustics, on the vowel triangle. Table 2.4 is a confusion table derived on vowels and Figure 2.1 is the vowel triangle. The vowel triangle is labeled with some key visual differences. In vowel recognition the most significant features are vertical lip separation and lip extension or roundness. In another study researchers compared visual recognition rates against nine measurements of the mouth. Their research indicated that two most important features of the mouth in discrimination were indeed vertical lip separation and the traditional lip extended-rounded feature (9).

2.2.3 Human Audio-Visual Integration. There have been numerous tests investigating audio-visual fusion in humans (6, 16, 17, 37). In one of the first fusion experiments a listener was presented with an audio "ba-ba" and a visual "ga-ga" and asked what they heard. In the 18-40 year old group 98% of the time, the listener responded with "da-da (17)." This phenomenon has become known as the McGurk effect. The discovery of the McGurk effect prompted a great deal of human factors research to try and explain this fusion phenomenon. Summerfield considers five different approaches to explain the fusion (37). One of the key

Table 2.1 Percent of Visual Identifications and Confusions of Initial Consonants (3: 94). Blank entries indicate less than 5%.

	P B M	W	F V	TH DH	T D N	L	S Z	SH CH ZH JH	R	Y	K G	HH
P B M	91											
W		91										
F V			91									
TH DH				39	16	14	5				10	8
T D N					51	5	8	6		6	6	6
L				5	16	47			7		6	9
S Z					25	8	46	9				
SH CH ZH JH							5	79		5	6	
R				5	15	5		8	53			
Y					26		9	16	8	13	16	
K G					24	11		5		8	29	10
HH				5	14	7			7	9	16	29

Table 2.2 Percent of Visual Identifications and Confusions of Final Consonants (3: 95). Blank entries indicate less than 5%.

	P B M	F V	TH DH	T D N	L	S Z	SH CH ZH JH	K G NX
P B M	90							
F V		92						
TH DH			38	17	13	16		10
T D N			6	45	12	18	7	10
L				35	36	15		10
S Z			5	21	7	32	27	
SH CH ZH JH						8	82	
K G NX				25	9	7	11	46

Table 2.3 A condensed list of phonetic symbols for American English (27: 24).

ARPABET	Example	ARPABET	Example
IY	beat	NX	sing
IH	bit	P	pet
EY	bait	T	ten
EH	bet	K	kit
AE	bat	B	bet
AA	Bob	D	debt
AH	but	H	get
AO	bought	HH	hat
OW	boat	F	fat
UH	book	TH	thing
UW	boot	S	sat
AX	about	SH	shut
IX	roses	V	vat
ER	bird	DH	that
AXR	butter	Z	zoo
AW	down	ZH	azure
AY	buy	CH	church
OY	boy	JH	judge
Y	you	WH	which
W	wit	EL	battle
R	rent	EM	bottom
L	let	EN	button
M	met	DX	batter
N	net	Q	(glottal stop)

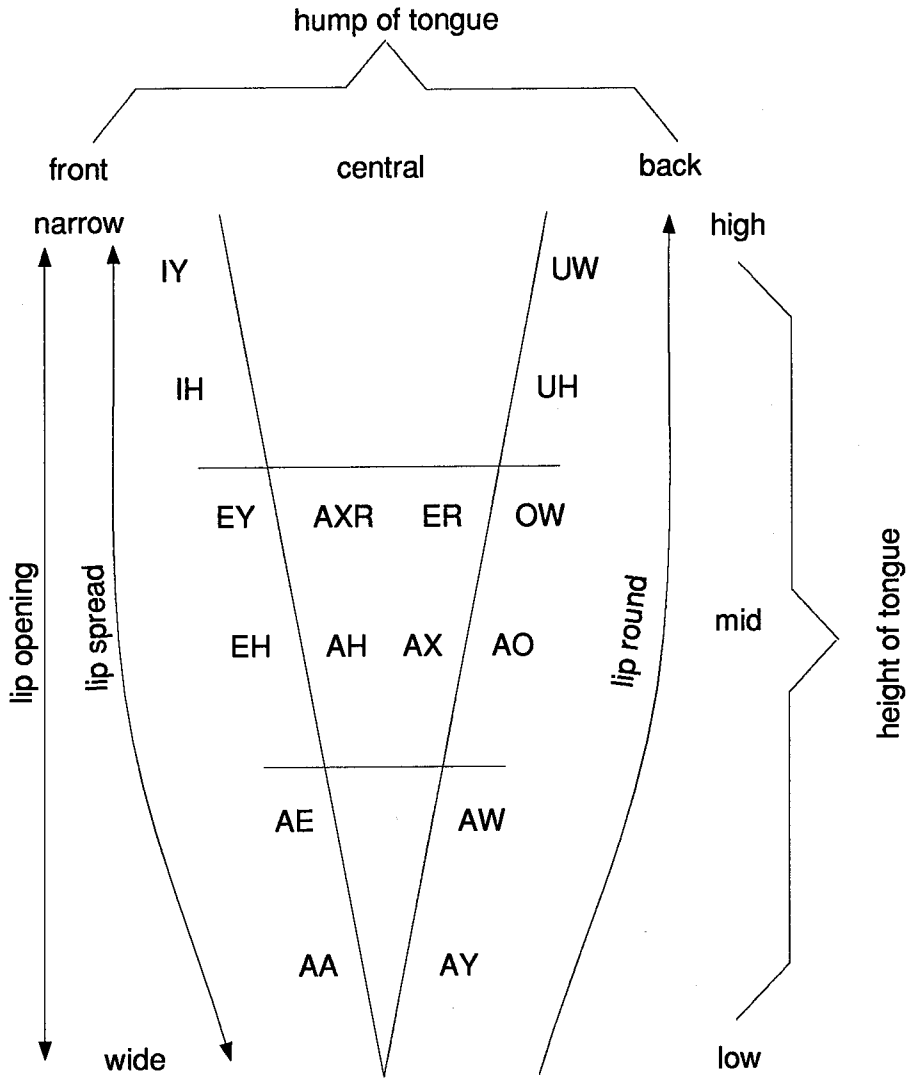


Figure 2.1 The Visual Vowel Triangle (3: 78).

Table 2.4 Percent of Visual Identifications and Confusions of Vowels (3: 82). Blank entries indicate less than 5%.

	IY	IH	EY	EH	AE	AA	AO	OW	UH	UW	AH	ER
IY	73	27										
IH	31	28	15	12							19	
EY		16	31	8	32						6	
EH		8	29	18	17	14					9	
AE			10	7	62	13						
AA					8	70	11				10	
AO						5	67	20				
OW								83		9		
UH									36	38		19
UW									11	74		5
AH	5	15		7		7	6		9		37	10
ER									12	16	8	58

points in Summerfield's work was the presentation of the different confusion trees for acoustic and visual channels which are recreated in figures 2.2 and 2.3. From these confusion trees it is easy to see the potential benefit of lip reading in automatic speech recognition. For example the M and N are acoustically hard to distinguish, but visually easy to separate. An automatic speech recognition system that includes this additional information should therefore improve accuracy in noise. Another researcher theorizes that the audio-visual fusion can best be modeled with fuzzy logic (16). Whatever the fusion process, clearly humans benefit from including the lip movement information in communication and therefore an automatic speech recognizer that utilizes this information should have improved performance.

2.3 Past Automatic Lip Readers

There have been several attempts to include lip reading in automatic speech recognition in the past. Table 2.5 is general overview of the past work. In general it is difficult to compare the results of the various methods tried because of the variety of test conditions. This section will discuss the methodology, benefits, and problems of each of the past approaches. A review

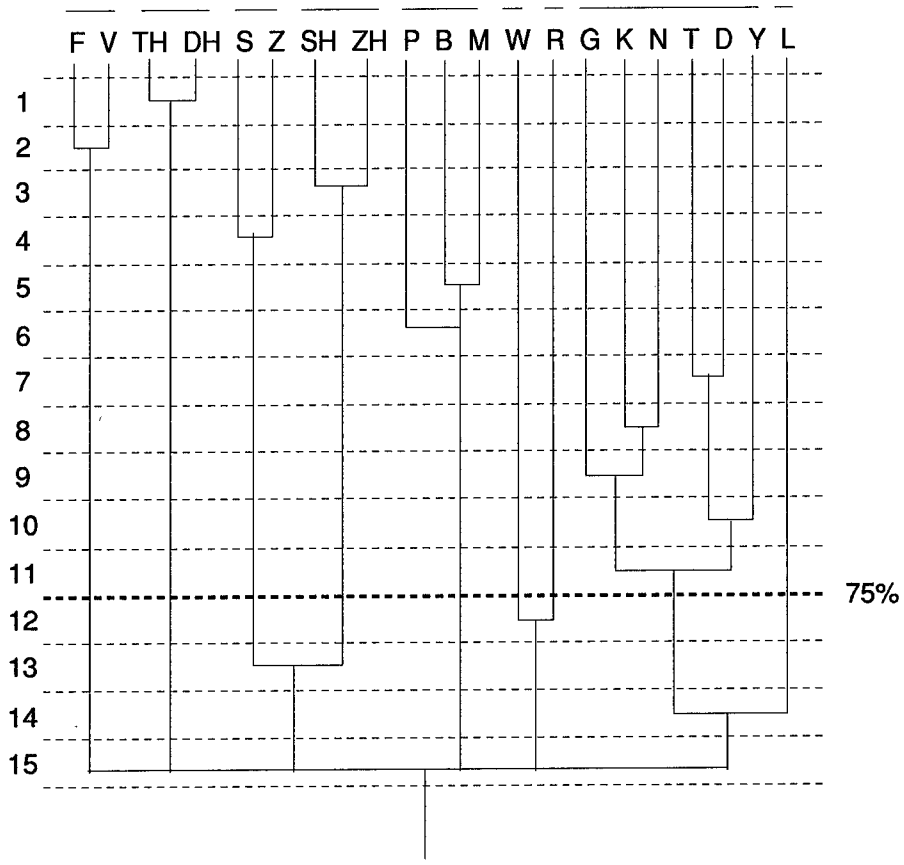


Figure 2.2 Visual confusions among consonants. The tree summarizes the results of a cluster analysis of the visual confusions made by trained hearing-impaired adult observers. The stimuli were consonants carefully articulated in CV syllables with the vowel "ah". Easily confused consonants cluster near the ends of the branches and dissimilar consonants cluster near the roots. The 9 groups of consonants defined after the formations of the 11th cluster can be considered to be distinct visemes and on 75% of the presentations these consonants were identified as belonging to their parent group (37: 13).

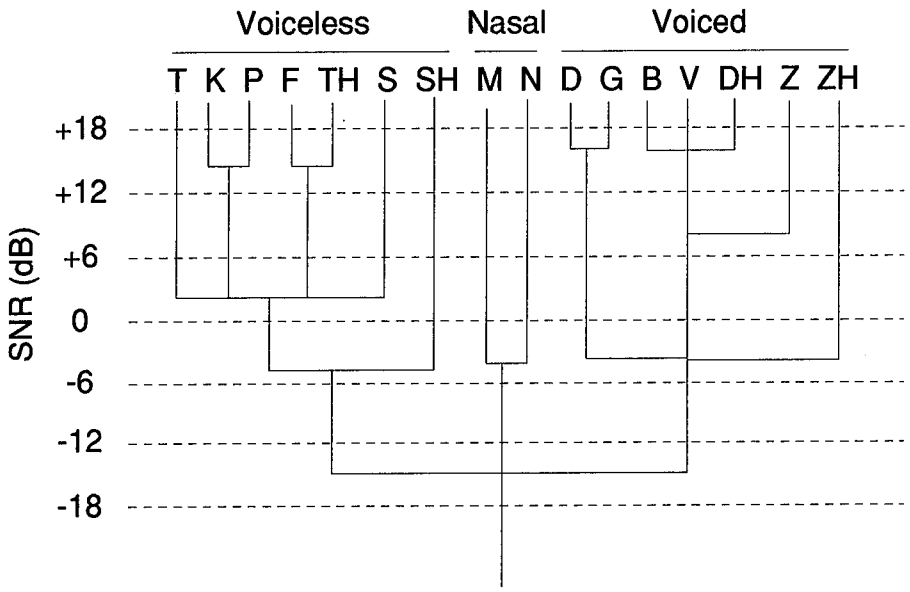


Figure 2.3 Acoustic confusion among consonants presented in noise. The tree summarizes the confusions that occurred when the consonants arrayed at the ends of the branches were spoken in CV syllables with the vowel "ah" and presented for identification in white noise at the signal-to-noise ratios indicated on the left-hand side of the figure (37: 15).

of these past efforts helps to understand the main problems that need to be overcome to make automatic lip reading a viable source of information in automatic speech recognition.

2.3.1 Traditional Pattern Recognition. In 1984, Eric Petajan (22) attempted to incorporate lip reading with acoustic information to enhance automatic speech recognition, using traditional image processing and pattern recognition techniques. He used a solid state camera to collect digitized images of the mouth. The system matched the input images against image templates for each utterance in a predefined vocabulary. Then the video processing section designated the closest matching image template as the visual recognition candidate. The system then compared the visual recognition candidate with the result of an acoustic recognition system. If the classification of the acoustic and visual classifiers matched, the additional visual information confirmed the result of the acoustic classifier. If the acoustic and visual classifiers differed he used a set of heuristic rules, based on the reliability of each classifier, to determine what the final classification should be. He tested his design with one

Table 2.5 Summary of past attempts to use lip reading in automatic speech recognition. Author: the primary researcher, Speaker: identifies the system as speaker dependent or speaker independent, Level: identifies the systems as phoneme, isolated word, or continuous speech recognizer, Raw Data: specifies the raw source used to collect the lip information, Features: specifies the features extracted from the raw data, Classifier: identifies the pattern recognition classifier used, Fusion: indicates the basis of the fusion technique, Vocab: specifies the vocabulary tested, ALR Results: indicate the automatic lip reader results achieved.

Year	84	88	89-90	91	92	92	93
Author	Petajan	Petajan	Sejnowski	Mase	Stork	Marshall	Silsbee
Speaker	Dep.	Dep.	Dep.	Indep.	Indep.	Dep.	Dep.
Level	Words	Words	Phon.	Cont.	Words	Words	Phon.
Raw Data	Video	Video	Video	Video	Video	LED refl.	Video
Features	Pixel	Pixel	Pixel	Opt. Flow	Face Markers	Photo Det.	Pixel
Classifier	Wt. Dist.	VQ/DTW	NN	Wt. Dist.	TDNN	DTW	VQ/HMM
Fusion	Rules	Rules	Wt. Sum STSAE	None	With NN arch.	Rules	Wt. HMM Scores
Vocab	Digits / Alpha.	Digits / Alpha.	Vowels	Digits	10 Alpha Utter.	Digits, 'Yes','No'	14 Vowels, Diphthongs
ALR Results	99% / 65%	93-100% / 80-94%	55.6%	73-100%	51%	12-62%	50%

speaker using two vocabularies, one consisting of the alphabet, the other of digits. The result of his work showed a clear improvement in the overall recognition rate, but at a cost. The use of images greatly increased the amount of computations and memory needed for classification and, therefore, made the system prohibitive for real time implementation. The system also restricted the user's movement because it required a boom mounted camera.

In 1988, Petajan made some improvements to his design (23). He reduced the computation cost by implementing vector quantization (27). Petajan matched each input image against 255 reference images. The system then assigned the number of the closest matching reference image to the input image, thereby reducing the time sequence of images to a vector of 8 bit numbers. It also vector quantized each reference utterance in the vocabulary and stored them as templates. The system then matched these templates against the input vector using Dynamic Time Warping (DTW) (21, 27). DTW allowed the closest match of two vector sequences, considering the variable speed of the utterances. This technique reduced the computations required for classification. Petajan used the same test vocabularies as before, and tested this new system with four speakers. Even though he used four speakers, he tested the device as speaker dependent. The results of these design changes enhanced the overall performance of the system achieving accuracy rates from 80-100% using the visual signal alone. The results of the combined visual and acoustic classifier, as before, showed improved performance over acoustic or visual alone. Overall this improved design did reduce the amount of computations, but it still needed a large memory and required the speaker to wear a cumbersome head mounted camera.

2.3.2 Neural Networks. The next effort to include visual information into an automatic speech recognition system uses neural networks. One approach using neural networks requires two different networks, one for the acoustic signal and one for the visual signal. The system makes its classification based on a weighting function. Sejnowski, Yuhas, et al. developed a neural network with this kind of structure, to predict the short time spectral amplitude envelope (STSAE) of the acoustic signal, using both the audio and visual inputs (31, 32, 40). The weighting function in their system was based on the acoustic signal to noise ratio. They

tested the system using a single speaker and a vocabulary of nine vowels. The accuracy of the system ranged from 70-80%.

2.3.3 Optical Flow. Another approach using video images is based on motion versus a sequence of static images. Optical flow is the apparent motion of brightness patterns due to the movement of a video camera relative to an object (2, 8). Pentland and Mase (15) used a video camera and four segmented regions of the mouth to collect the optical flow data from the mouth. The benefits of using optical flow are the data are easier to collect than lip shape extraction and can be used to identify word boundaries since there is a stop between each word. The system they designed was based on the physical nature of the muscles of the mouth. They tested the system as speaker independent with three speakers. The vocabulary consisted of continuous spoken utterances of three to five digits. The system achieved 73 to 100 percent accurate word recognition. The results clearly demonstrate that lip movement can be used for identification and may be particularly useful in a continuous speech recognition system and provide a better speaker independent method of collecting lip information.

2.3.4 Time Delayed Neural Networks. Another neural network approach utilizes a time delay neural network which uses temporal inputs, allowing it to take advantage of the predictable nature of speech. Stork *et. al* developed a time delay neural network to improve speech recognition with lip reading (35). They designed an acoustic and a video neural network and then expanded the network to combine the outputs. They also simplified the input by placing ten reflective markers on the face of the speaker which greatly reduced the video processing and computations required for classification. Unlike the other approaches, they tested their design as a speaker independent system. The test vocabulary consisted of ten alphabetic utterances. The consonants chosen were specifically chosen to illustrate the benefits of an automatic lip reader. They were "b,d,f,m,n,p,s,t,v,z". Some pairs were acoustically difficult to separate such as "b" and "d" and some visually difficult such as "d" and "t". This vocabulary is well suited for testing the benefits of adding automatic lip reading to automatic speech recognition. They tested the system as acoustic, visual, and acoustic-visual

combined. The results were 51% for visual, 64% for acoustic, and 91% for the combined acoustic-visual classifier.

2.3.5 Active Photo Sensors. All the systems considered thus far have used video images to capture the lip information. As noted above acquiring these video images is difficult and processing them often requires an extensive amount of computations and memory. Marshall (14) demonstrated that it is possible to track some motions of the mouth with a simple photo detector. He used a modulated LED to illuminate the mouth and then recorded the reflected signal. He attempted recognition using a lip motion detector alone and combined with an acoustic recognition system. He tested his design with five speakers, using both speaker independent and speaker dependent tests. His results varied drastically depending on the person and the test. Overall the results showed a slight improvement in recognition when incorporating the lip motion detector. Even though his results were not as good as the previous results using video images, his design greatly reduced the problem of acquiring and processing the lip information.

2.3.6 Hidden Markov Models. Another approach in speech recognition takes advantage of the predictable nature of speech. Certain sounds in speech are more likely to follow each other. Rabiner gives a complete description of Hidden Markov models which take advantage of this predictability (27). Silsbee and Bovik developed an automatic speech recognition that used HMMs for both the visual and acoustic signals (33). The authors used vector quantization techniques to encode the visual images similar to the method used by Petajan. The two classifiers gave results in visual and acoustic HMM scores. The design then combined the two HMM scores, which resulted in an overall classification. The authors tested the design using a set of 14 vowels with one speaker. They also tested the system with various levels of "additive white Gaussian noise." In low noise, the combined audio visual system achieved a 96% accuracy rate. In a high noise environment, the results showed that combined audio-visual recognition improved recognition from 69% to about 77% over acoustic recognition alone. Although this method may offer improved results over the

traditional approach, it increased the computations required, still required a large memory, and didn't address the problem of acquiring the mouth images.

2.3.7 Conclusion. The key problem with most of the past work is acquiring and processing the lip information. In all but one case video was the primary source of information. The problem with video is that it is computationally very expensive to process and extract the necessary lip reading features from video. Marshall's use of a modulated light signal to acquire the lip information was the only approach that did not use video, but his tests results were inconsistent. Even though an active light sensor did not provide very good results, a simple device still might be able to acquire the lip information. Another consistent problem in the past research is an effective way to integrate the automatic lip reader into an automatic speech recognition system. In most cases the fusion technique was based on heuristic rules.

2.4 Classification and Fusion Techniques

This section reviews classification and fusion techniques used in this thesis. First a review of dynamic time warping and linear predictive coding is given, emphasizing their purpose. Next a feature fusion technique is presented based on preprocessing the signals prior to dynamic time warping. Finally, a classifier fusion technique is presented based on a pseudo probability mass function derived from the dynamic time warping distances.

2.4.1 Linear Predictive Coding. Linear predictive coding (LPC) has been found to be a very successful technique for compactly encoding the essential information of speech for recognition. The coding is based on an all pole filter model of speech. The all pole model is effective because of the physical nature of the human vocal tract. The basic idea behind the LPC model is that for a given speech sample at time n the speech signal $s(n)$ can be approximated as a linear combination of the past p samples of the speech signal, such that

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p s(n-p) + e(n) \qquad (2.1)$$

where the coefficients $a_1, a_2, \ldots, a_p$ are assumed constant over the speech analysis frame and $e(n)$ is the error between the actual and predicted value of the signal. Efficient algorithms have been developed to extract the LPC coefficients from a speech utterance and high speed computers can now perform the calculations at real time. In most automatic recognition systems the LPC coefficients are typically transformed to cepstral coefficients. The cepstral coefficients, which are the coefficients of the Fourier transform representation of the log magnitude spectrum, have been found to be a more robust and reliable feature set for recognition than the LPC filter coefficients (1, 12, 27). The result of LPC analysis is that a speech utterance is transformed into a time sequence of LPC cepstral coefficient vectors. The length of the vectors is dependent on the parameter p chosen, typically p is chosen between 8 and 16. After LPC analysis, the signal can be input to a dynamic time warping algorithm for classification.

2.4.2 Dynamic Time Warping. One method of determining classification of an unknown utterance in automatic speech recognition is to try to match the unknown utterance with earlier utterances recorded as training templates. Unfortunately, due to variable speaking rate, the unknown utterance is generally either shorter or longer than the training templates. Therefore some kind of time warping method must be used. Past research has shown that a simple linear warping is inadequate because of the dynamics of speaking within an utterance (21, 27). A better method of compensating for time misalignment is to use a procedure known as dynamic time warping (DTW) or dynamic programming. DTW uses a nonlinear warping algorithm to find the distance between an unknown utterance and a reference template. There are a number of different constraints that can be used in DTW to warp the signals together. A study of various techniques has shown that unconstrained endpoints provides the best overall performance in DTW (25). Unconstrained endpoints allows the algorithm to compensate for imprecise endpoint detection. The output of the DTW algorithm is a distortion measure. The smaller the distortion the more similar the two signal are. DTW is used in classification by measuring the DTW distance between the unknown utterance and all the training templates and assigning the unknown utterance the same class as the closest template.

2.4.3 Feature Fusion. In feature fusion the goal is to combine the features of different analysis techniques together prior to submission to a classifier as illustrated in Figure 2.4. The goal is to combine the different kinds of features without suppressing or losing any information. The problem is that different features may have drastically different magnitudes and therefore certain features may dominate a DTW distortion metric. One way to overcome this problem is to normalize the magnitudes of the various features prior to fusion. Choices of normalization include range compression, Gaussian transform to identical means and variances, or energy normalization. Range compression is performed as

$$n_i = \frac{x_i - min(x_i)}{max(x_i) - min(x_i)} \tag{2.2}$$

where n_i is the normalized sequence and x_i is the original sequence. Range compression is susceptible to distortion due to noise spikes. Gaussian transformation to identical means and variances is done by letting

$$n_i = \frac{(x_i - \mu_x)\sigma}{\sigma_x} + \mu \tag{2.3}$$

where n_i is the normalized sequence and x_i is the original sequence, μ_x and σ_x are the mean and standard deviation of x, and μ and σ are the desired mean and standard deviation. The result is a sequence n_i with mean μ and standard deviation σ. Another normalization technique is energy normalization. In this technique

$$n_i = \frac{x_i}{\sqrt{\sum_{k=1}^{N} x_k^2}} \tag{2.4}$$

where n_i is the normalized sequence and x_i is the original sequence, N is the total length of the sequence. Each of these normalization techniques is effective depending on the problem.

2.4.4 Classifier Fusion. Classifier fusion involves the combination of multiple classifiers into one overall best classification as depicted in Figure 2.5. The goal is to improve accuracy by taking advantage of the different confusion matrices of different classifiers. A

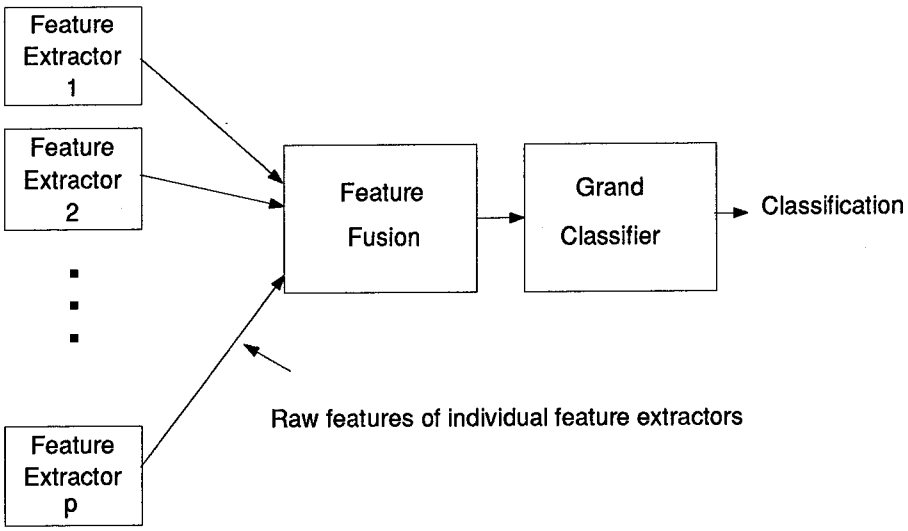


Figure 2.4 Generic feature fusion.

number of different classifier fusion techniques are considered in a paper by Xu (39). Their analysis considers classifier fusion based on three different levels of information output from the individual classifiers. The first level of information from the individual classifiers is simply the best class as determined by each individual classifier. The second level is a ranked list of closest classes. The third level is a distance metric from each class. All classifiers output at least the first level and most are capable of outputting information at level three. A DTW based classifier can certainly output the third level of information, the distortion distances from the test utterance and each training template. Xu also presents a method of converting the DTW distances to a pseudo probability mass function. The probability of each class i based on classifier k is

$$p_k(i) = \frac{\frac{1}{d_k(i)}}{\sum_{i=1}^{M} \frac{1}{d_k(i)}}, \tag{2.5}$$

where $d_k(i)$ is the distance from the test utterance to class i for classifier k, and M is the total number of classes. The final probability is based on an average of the pseudo probabilities of

each classifier. The final probability is

$$P(i) = \frac{1}{K}\sum_{k=1}^{K} p_k(i) \tag{2.6}$$

where K is the total number of classifiers. This fusion assumes that each classifier is equally reliable, however if they are not the final probability could also be derived as a linear combination of the individual pseudo probability mass functions. The final probability in this case would be

$$P(i) = \sum_{k=1}^{K} c_k p_k(i) \tag{2.7}$$

where c_k are the coefficients and

$$\sum_{k=1}^{K} c_k = 1. \tag{2.8}$$

The values of the c_k are based on the confidence of each of the individual classifiers.

2.5 *Ultrasound Basics*

This section will explain some of the basics of ultrasound. Ultrasonic waves are sound waves that are above the hearing threshold of humans, that is sound waves higher than 20,000 Hz. Ultrasound can be generated with a piezoelectric device driven by an electronic oscillator. The wavelength of the signal is related to the speed of sound and the frequency as

$$\lambda = \frac{v}{f}, \tag{2.9}$$

where v is the speed of sound and f is the frequency. The speed of sound in air at 20°C and standard atmospheric pressure is 343 m/s. Therefore, for an oscillator operating at 40Khz the wavelength would be 0.8575cm. To use ultrasound to detect lip motion we need to be able to propagate ultrasound through air and examine the reflected signal off of the speaker's mouth. The higher the frequency the better the range resolution, but the higher the frequency the

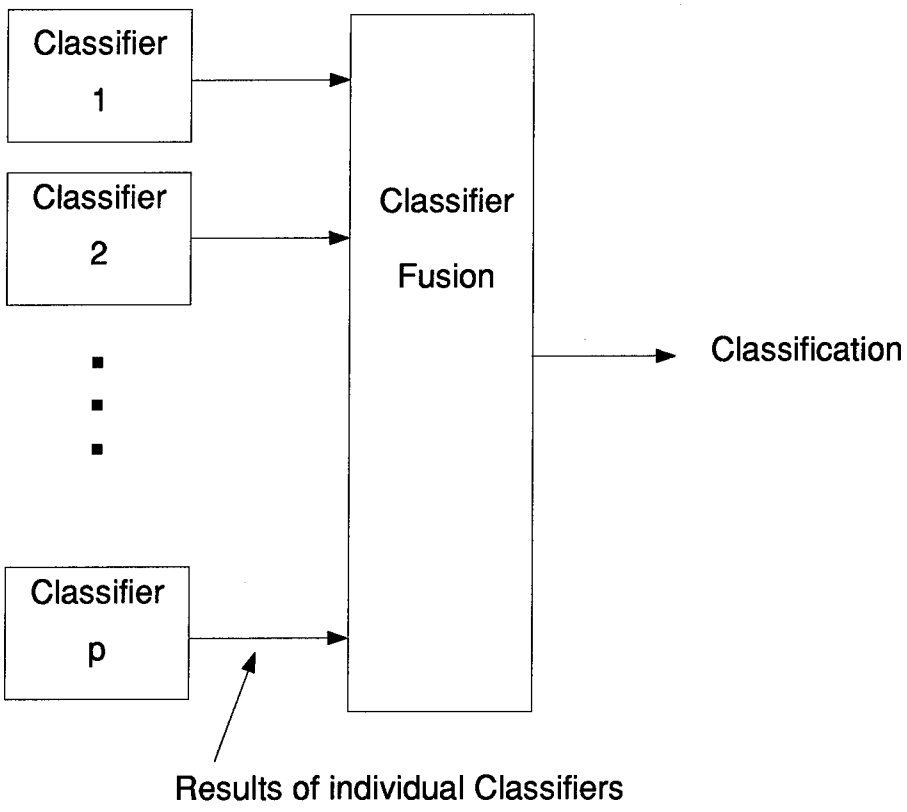


Figure 2.5 Generic classifier fusion.

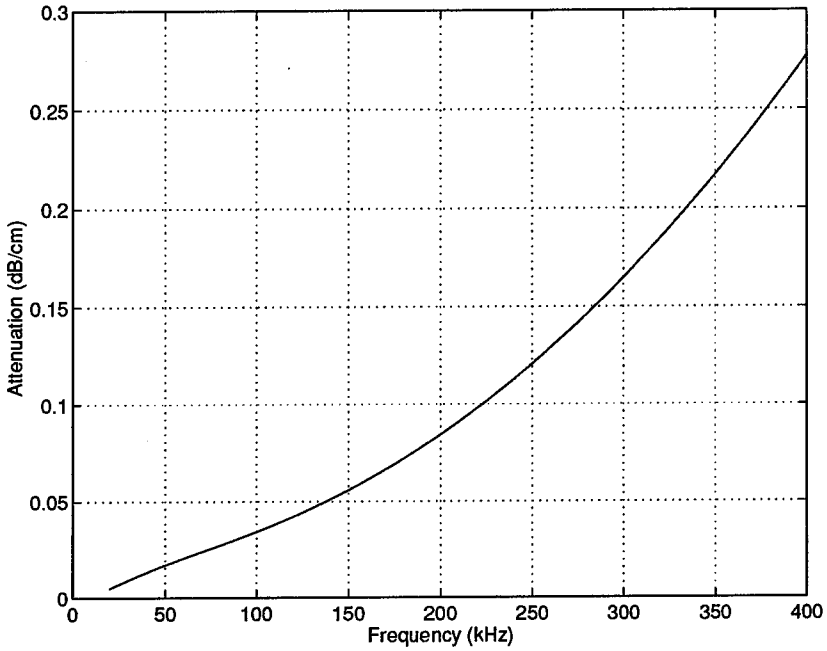


Figure 2.6 Attenuation of sound in air at 20°C, 50% humidity, and 1 atmosphere.

more severe the attenuation of sound in air (11). Figure 2.6 shows the relationship between frequency and attenuation. As noted the speed of sound is a function of frequency, temperature, humidity, and atmospheric pressure. For the normal operating ranges of a human frequency is the dominant factor in determining attenuation. At 40kHz the 3dB distance is approximately 1.5 meters, whereas at 200kHz the 3dB distance is only 0.4 meters. Since the devices used in this thesis are located approximately 2cm from the talker's face, attenuation is not a significant problem at the frequencies used. There are a number of different ways that ultrasound can be used. Both the devices considered in this thesis operate in the continuous mode. A transmitter outputs a continuous sinusoidal ultrasonic signal and information is derived based on the reflected signal.

2.6 Support Hardware and Software

The primary support hardware used in this thesis is a Sun Sparc workstation and an Aerial Proport (24). The Aerial Proport is a A/D and D/A converter with both mike and line

level inputs and outputs. The Proport allows two channels to be recorded simultaneously and allows up to 40dB of gain for each channel. The Proport supports a number of different sampling rate, but all the work in this thesis used a sampling rate of 16Khz. The Proport also provides antialiasing filtering corresponding to the sampling rate. As well as input the Proport also provides D/A for playback purposes. The Proport is connected to a Sun Sparc workstation and can be controlled with the primary software used in this project, Entropic Signal Processing Software (ESPS) (7). ESPS is a software package well suited for developing and testing automatic speech recognition. The program includes 85 user-level programs that can be called from a UNIX shell and includes a library of 200 functions. This software allows quicker design and testing of automatic speech recognition systems and also readily allows subsequent researchers to repeat or continue the work done in this project. The primary algorithms used in the ESPS package include a feature extraction routine "ACF" and a dynamic time warping program "DTW_REC". A complete summary of the input script and parameter files used with ESPS is given in Appendix A.

2.7 Conclusion

This chapter has reviewed the past efforts in automatic lip reading and the techniques used to extract and fuse the lip information into a traditional, acoustic based, automatic speech recognition system. The past efforts have demonstrated the potential benefits of including lip information in automatic speech recognition and also the difficulty of acquiring and processing this lip information. If an effective and efficient method of extracting and fusing the lip information into an automatic speech recognition system can be found, automatic speech recognition in adverse environments would be possible.

III. System Design

All of the classifiers in this research are designed as speaker dependent, isolated word recognizers. This chapter will give the details of the design of the systems to be tested. The first two sections describe the devices investigated in this thesis. Then the design details of the automatic speech recognizer and the automatic lip readers are given. Finally, two combined designs, one based on classifier fusion and one based on feature fusion, are presented.

3.1 Ultrasonic Mike

The "Ultrasonic Mike" is the first device that was investigated in this research. Figure 3.1 gives an overview of how the device works. The device using a 40Khz oscillator drives a piezoelectric material which creates an ultrasonic signal. The ultrasonic signal is directed at the talker's mouth, repeated reflections from the talker's mouth and the device establish a standing wave. Any movement of the mouth results in a change in the standing wave. An ultrasonic receiver also located in front of the talker's mouth converts the ultrasonic signal back to an electrical response. The electrical signal is passed through an envelope detector. The result is a device that outputs a signal that changes in response to the movement of the talker's mouth and is zero when ever the talker's mouth is still. Figure 3.2 is a recorded estimate of the step and impulse response of the device. The simulated step response on average lasted 0.4 seconds and the simulated impulse response lasted 0.1 seconds. The step response was simulated by abruptly moving a flat surface away from the device. The impulse response was simulated by quickly passing a thin object across the field of view of the device. Initial recording with the device demonstrated the device was very sensitive to position, resulting in drastically different waveforms from utterance to utterance. Therefore, the ultrasonic transmitter and receiver were mounted on a boom mike of a headset to maintain their relative position with the mouth. The position sensitity is later confirmed, resulting in cross session accuracy of approximately 13%, only slightly better than chance. A computer simulation of the device was designed and also demonstrated that the device was very sensitive to position and that even a millimeter

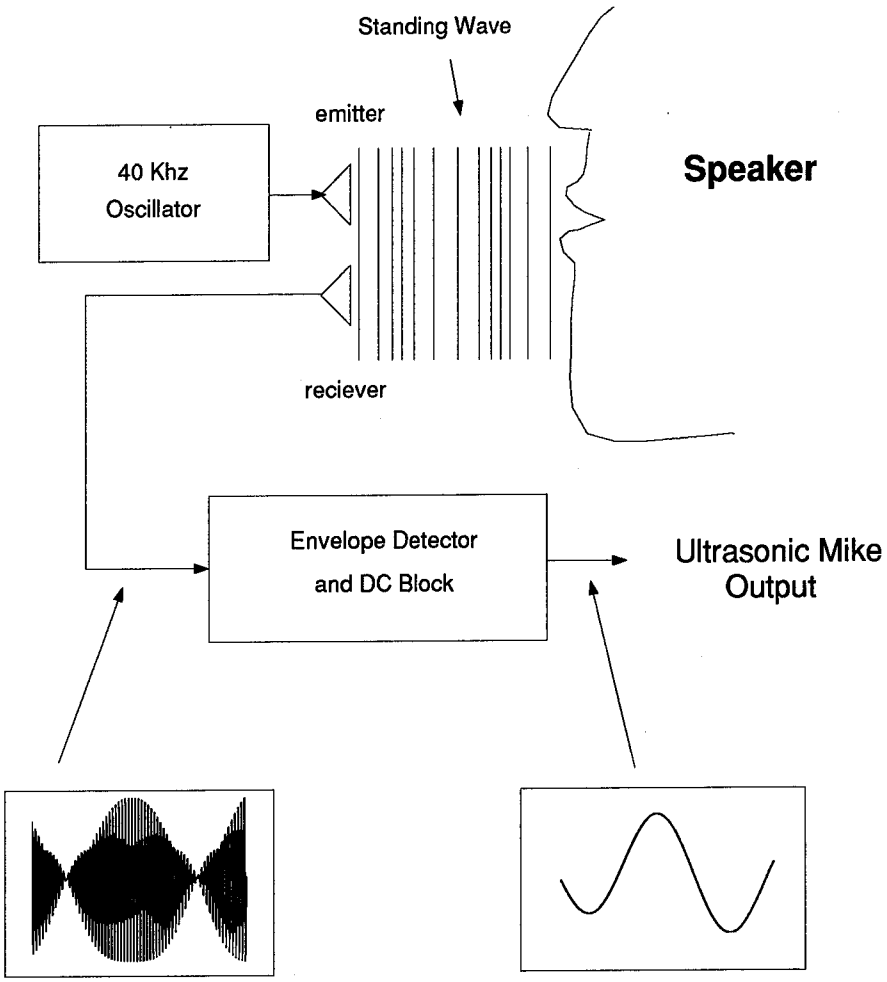


Figure 3.1 Overview of operation of the "Ultrasonic Mike."

change in position could significantly change the output of the device. The results of the simulation are presented in Appendix B.

3.2 Lip Lock Loop

The "Lip Lock Loop" is the second device considered in this research project. Figure 3.3 is a basic overview of how this device works. This device uses a phase lock loop with the lip position in the loop. A VCO with a center frequency at 23Khz is used to drive an ultrasonic transmitter. The signal is directed at the talker's mouth and the return signal from the mouth is collected with an ultrasonic receiver. The difference between the phase of the transmitted

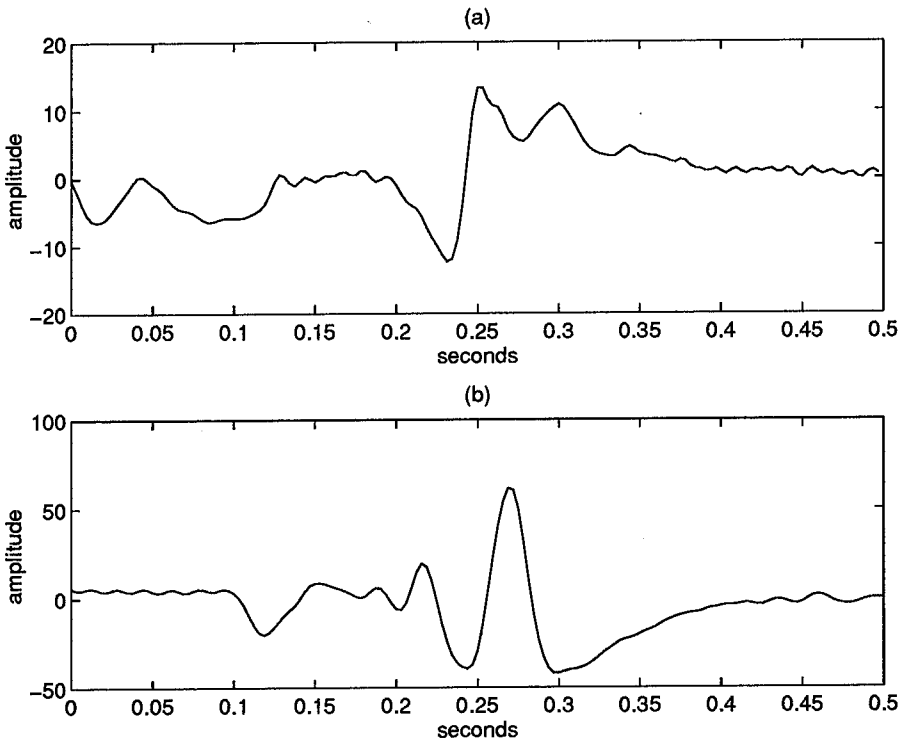


Figure 3.2 (a) Step response and (b) Impulse response of the "Ultrasonic Mike"

signal and the return signal causes the phase comparator to output a signal which changes the frequency of the VCO until the transmitted and received signal are in phase. The output of the device is derived from the control voltage of the VCO. The resulting signal has a DC level corresponding to the position of the lips relative to the device. The signal changes corresponding to mouth movement. Again the transmitter and receiver pair were mounted on a boom mike of a headset to maintain a position fixed relative to the movement of the talker's head. Figure 3.4 is the estimated impulse and step response of this device. The average step response was found to be approximately 0.25 seconds and the impulse response was around 0.1 seconds. The goal of this modified design was to reduce the position sensitivity of the device, but initial experiments again demonstrated the device to be quite sensitive to position. Cross session accuracies ranged from 6-15%.

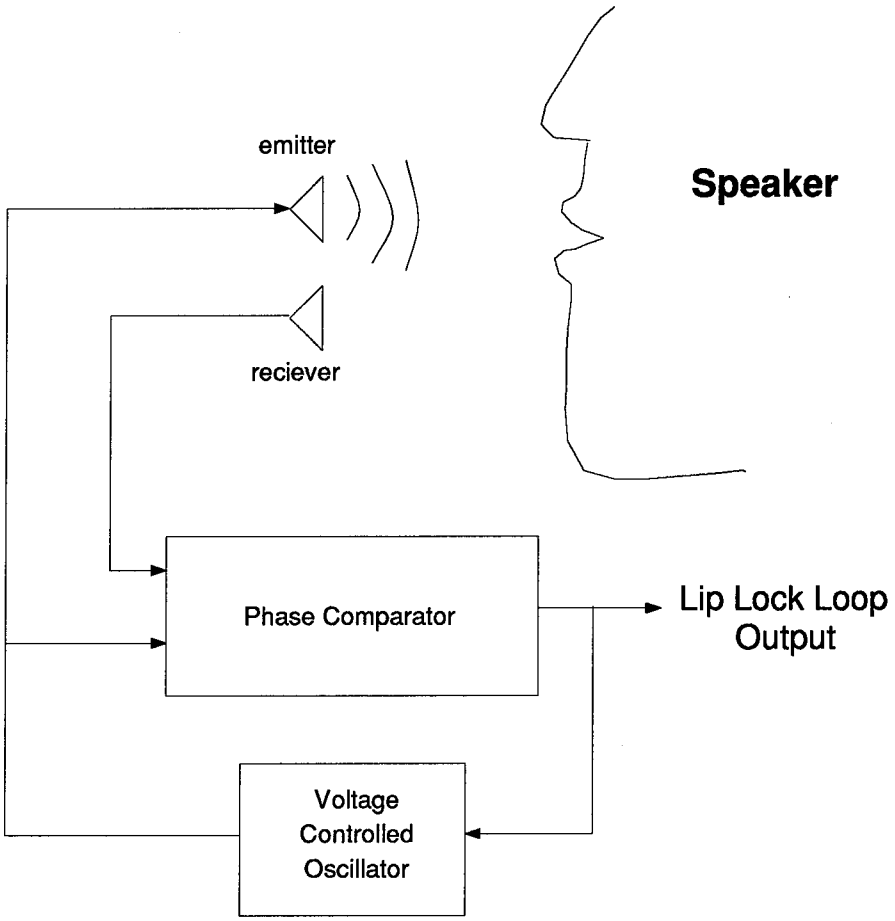


Figure 3.3 Overview of operation of the "Lip Lock Loop."

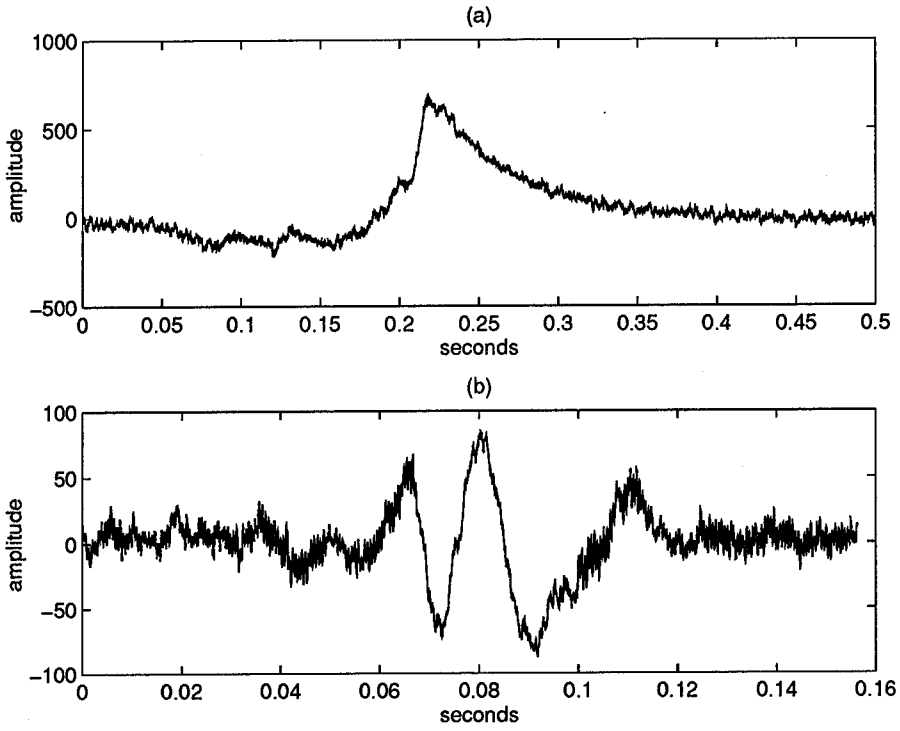


Figure 3.4 (a) Step response and (b) Impulse response of the "Lip Lock Loop"

3.3 Automatic Speech Recognizer

The automatic speech recognizer design is based on tried and tested methods of the past. The recognizer is not meant to be a state of the art design, but simply a basic recognizer to be used in fusion tests. Figure 3.5 gives a basic overview of the design used in this project. The sample rate was 16Khz. The sample signal was first passed through a preemphasis filter to improve the high frequency information. Then the signal was frame blocked with 10ms windows with a 50% overlap. This corresponds to 160 samples per window, moving 80 samples ahead for each frame. Then a Hamming window was applied to the window and ten linear predictive coding (LPC) cepstral coefficients were calculated. Each of the utterances submitted to the system were 1.5 seconds long, resulting in 299 frames per utterance. The acoustic classifier used template matching based on DTW distances. A test utterance's LPC cepstral vectors were matched against all of the training template's LPC cepstral vectors, and the class of the closest matching template was assigned to the test utterance. The acoustic

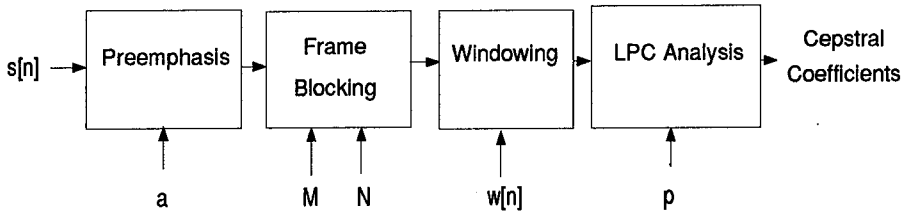


Figure 3.5 Block diagram of the implemented automatic speech recognizer used in this thesis. The numbers used in this thesis are a : preemphasis parameter (0.95), M : window size in samples (160), N : step size in samples (80), $w[n]$: is a Hamming window, p : LPC order (10).

classifier outputs a ranked list of the closest ten matches from the training templates and the corresponding distances.

3.4 Automatic Lip Reader

Both the "Ultrasonic Mike" and the "Lip Lock Loop" have similar outputs; therefore, the design of the automatic lip readers were identical for both devices. From the initial recordings of the devices, the output, as shown in Figures 3.6 and 3.7, appeared to be predominately low frequency. The Fourier transform of the ultrasonic signals revealed, as shown in Figures 3.8 and 3.9, no significant content above 150Hz, therefore the signal could be significantly down sampled without loss of information. To correspond to the acoustic recognizer the ultrasonic signal was down sampled by 80, resulting in 299 data points per 1.5 second utterance. Since the signal is now only 299 points long, the raw signal was used as input to the dynamic time warping algorithm. The lip reader also used template matching based on DTW distances. Test utterance's are matched against all of the training utterances, and the class of the closest matching template is assigned to the test utterance. The automatic lip reader also outputs a ranked list of the closest ten matches from the training templates and their corresponding distances. Figure 3.10 is a functional block diagram of the automatic lip reader.

3.5 Combined Systems

This research considers two designs for fusion. The first design is based on classifier fusion and the second design is based on feature fusion prior to dynamic time warping.

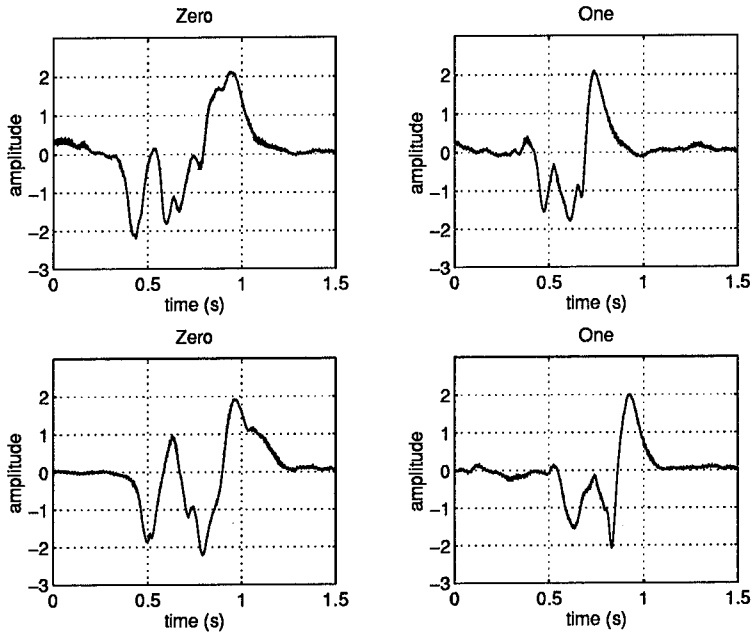


Figure 3.6 Typical signals from the "Ultrasonic Mike."

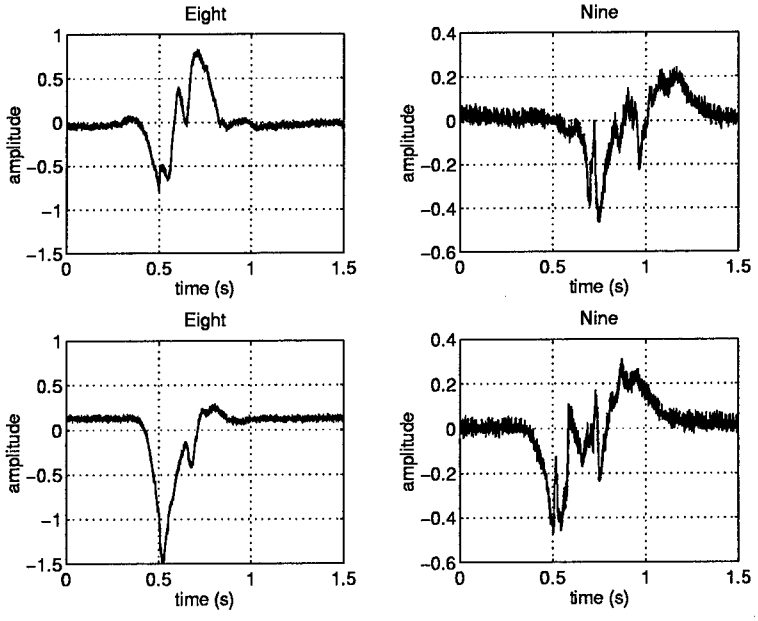


Figure 3.7 Typical signals from the "Lip Lock Loop."

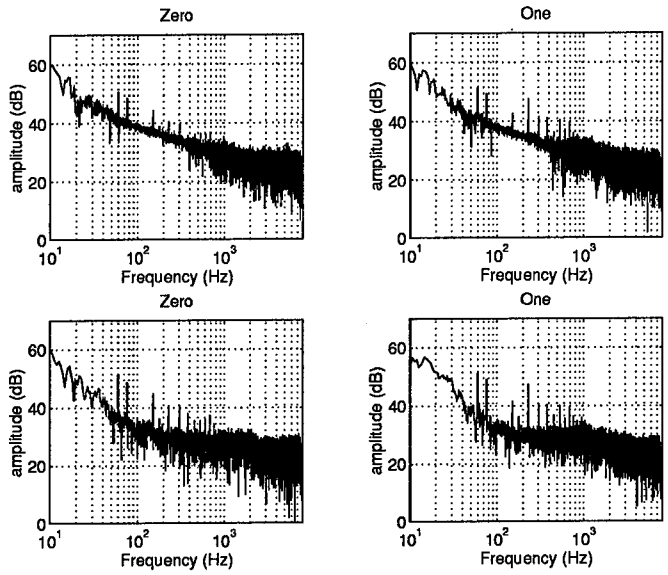


Figure 3.8 Typical Fourier transform of the "Ultrasonic Mike" signal.

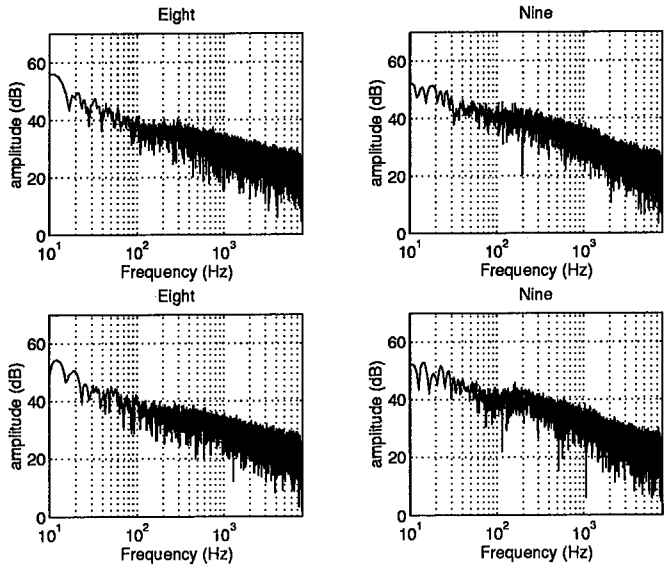


Figure 3.9 Typical Fourier transform of the "Lip Lock Loop" signal.

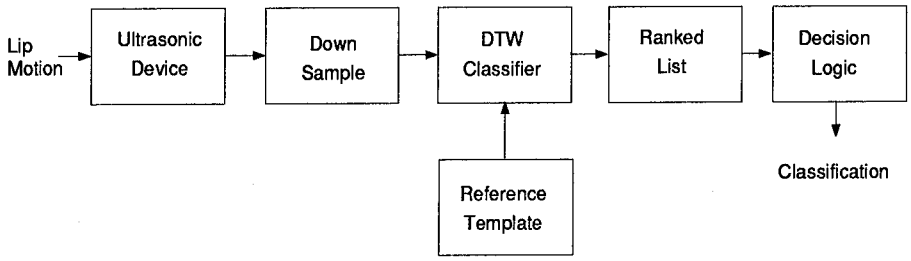


Figure 3.10 Block diagram of the automatic lip reader design used in this thesis.

3.5.1 Classifier Fusion. A block diagram of the classifier fusion design is given in Figure 3.11. In this design the automatic speech recognizer and the automatic lip reader function separately and input the ranked list of classifications and corresponding distances to the classifier fusion section. The number of ranks has to be at least two, but a larger number of ranks allows more flexibility in the fusion process. In this thesis ten ranks were used, corresponding to the number of classes. First the minimum distance from the test utterance and each class is determined. The fusion algorithm requires a distance measurement from each class; therefore, if a given class is not ranked, then the distance to that class is set equal to the largest distance. Then the pseudo probability mass (PMF) function for each classifier is calculated. The pseudo probability mass functions (PMF) is derived from the DTW distance as

$$p_k(i) = \frac{\frac{1}{d_k(i)}}{\sum_{i=1}^{M} \frac{1}{d_k(i)}}, \tag{3.1}$$

where k is the classifier, i is the class, M is the total number of classes, and $d_k(i)$ is the distance from the test utterance to class i for classifier k. The final overall classification is based on a linear combination of the automatic lip reader and the automatic speech recognizer probability mass functions. The overall probability mass function is

$$P(i) = \lambda p_{ac}(i) + (1 - \lambda) p_{lr}(i) \tag{3.2}$$

where p_{ac} is the probability mass function for the acoustic classifier, p_{lr} is the probability mass function for the lip reader, λ is a scale factor that is a function of the SNR and the accuracies of

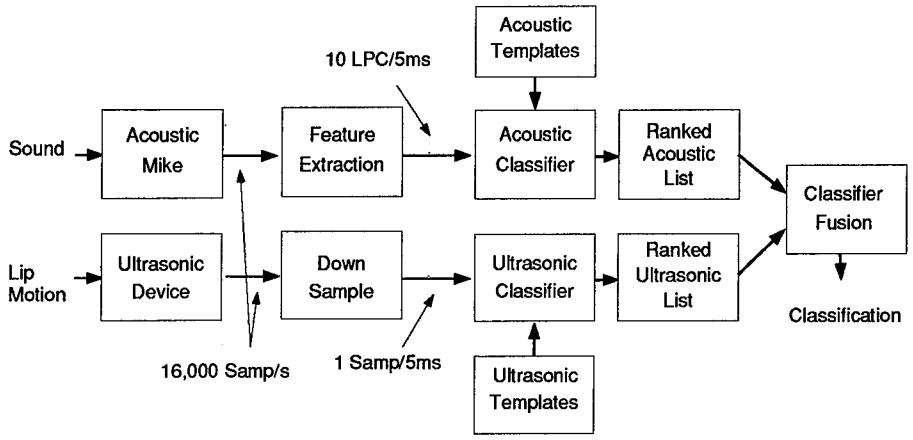


Figure 3.11 Combined design based on classifier fusion.

the recognizers, and i is the class. The final classification is based on the maximum $P(i)$ over all i. The classifier can also be set to reject the classification if the second largest $P(i)$ is too close to the maximum. The benefit of this scheme is that as noise causes the distances of the acoustic recognizer to become equal, the automatic lip reader results automatically dominate the overall decision. This procedure can also be considered as a fuzzy logic approach. In Appendix C, a fuzzy logic approach is considered that allows the scale factor λ to be fixed.

3.5.2 Feature Fusion. A block diagram of the feature fusion based classifier is given in Figure 3.12. The basic idea behind the feature fusion design is to combine the information before dynamic time warping, thereby warping the lip signal and the acoustic signal together. In this design the acoustic LPC features and a normalized version of the ultrasound signal are fused prior to dynamic time warping. Normalization was performed by subtracting the mean and dividing by the standard deviation, then multiplying by a constant to achieve a new standard deviation. The resulting normalized ultrasonic signal is

$$u_n[k] = \left(\frac{u[k] - \mu_u}{\sigma_u}\right)\sigma \tag{3.3}$$

where $u[k]$ is the original ultrasonic sequence, μ_u is the average of the sequence, σ_u is the standard deviation of the sequence, and σ is the scale factor. The resulting sequence will have zero mean and a standard deviation equal to σ. Again the best value of σ to use is a function

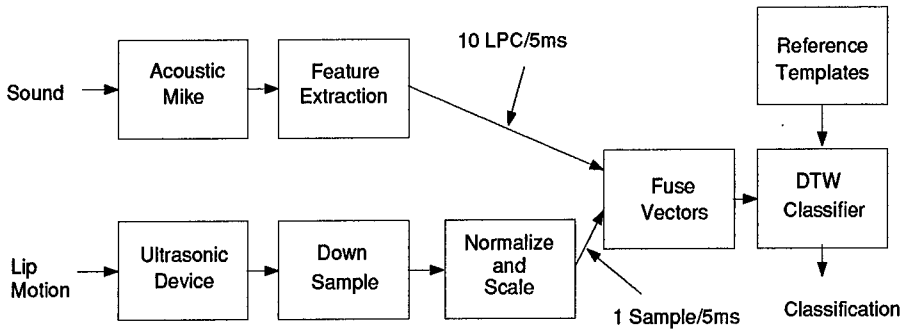


Figure 3.12 Combined design based on feature fusion.

of the SNR and the performance of the classifiers alone. The final classification is based on the closest matching template, and rejection can be implemented by using a threshold based on the difference between the closest and second closest match.

3.6 Conclusions

This chapter presents the designs of the various systems that are investigated in this project. The first design presented is a basic automatic speech recognition system based on methods found to be successful with isolated word recognition. Next an automatic lip reader design is presented based on the output of the ultrasonic devices investigated in this project. Finally, two different fusion designs are presented, one based on feature level fusion and one based on classifier level fusion. Now that the designs have been established the next chapter presents the results of tests on these designs.

IV. Experimental Results

This chapter will present the results of experimentation with the designs presented in chapter 3. The first section discusses the results of the acoustic recognizer alone. The acoustic recognizer is tested with various levels of white Gaussian noise added. Next, the automatic lip reader is tested as a stand alone recognizer. Finally tests are conducted on the combined systems. All of the following results are speaker dependent results on a isolated word recognition task. The vocabulary is the digits from zero to nine. All utterances are 1.5 seconds in duration and no endpoint detection was performed. Fourteen utterances of each digit are made. The sampling rate is fixed at 16Khz. The acoustic signal is recorded on one channel and the ultrasonic signal is recorded on the other. Then the two channels are demuxed and processed individually.

4.1 Automatic Speech Recognizer Alone

Figure 4.1 presents the typical results of the acoustic recognizer. The results are based on a system with 10ms windows with a 50% overlap. Preemphasis and a Hamming window were also used. The acoustic system used was based on 10 LPC cepstral coefficients. Other systems were tested and gave similar results, but this system design was used in all subsequent fusion tests. The classifier was based on DTW using up to four templates per utterance. The noise level was estimated by measuring the power during a non speech segment of the utterance and the signal plus noise level was estimated by measuring the power during a speech segment of the utterance. The estimated SNR was then derived as,

$$SNR = \frac{SN - N}{N} \tag{4.1}$$

where SN is the estimated signal plus the noise power and N is the estimated noise power. Various signal to noise ratios were tested by adding white Gaussian noise to the recorded utterance. As can be seen from the figure, the accuracy falls off drastically as the SNR falls below zero. By adding additional templates accuracy is improved at the expense of additional

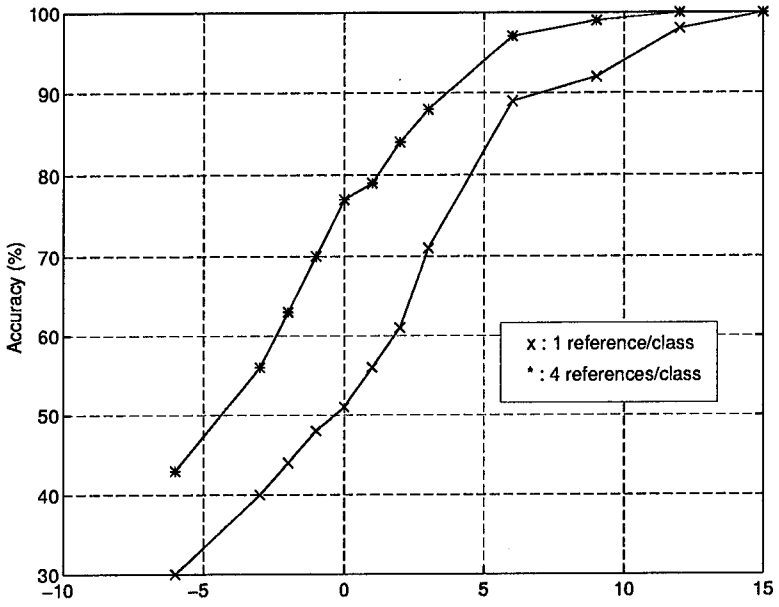


Figure 4.1 Typical results of the acoustic classifier alone. *: 4 reference templates per class,x: 1 reference Template per class.

computations and memory storage, but the accuracy is still far below acceptable levels for implementation. Table 4.1 represents the typical confusion matrix for the acoustic recognizer at 0dB. From the confusion table it can be seen that the accuracy of certain digits fall off more rapidly than others, such as the digit "2."

4.2 Automatic Lip Reader Alone

The average cross session results of the "Ultrasonic Mike" and the "Lip Lock Loop" are 12.6% and 10.6% respectively. These results are the average of 5 runs, using different sets of reference templates from one session and different sets of test utterances from a another recording session. As can be seen from these results, cross session results are only slightly better than chance. These results demonstrate that both devices are very position sensitive; therefore, great care had to be taken not to move the position of the device relative to the talker's mouth. Table 4.2 presents the average results of the automatic lip reader over five different sets of reference templates and test utterances. For these results the training templates

Table 4.1 Typical confusion matrix for acoustic recognizer at 0dB using 4 reference templates.

Actual Class	Called										Error Rate (%)
	0	1	2	3	4	5	6	7	8	9	
0	5				5						50
1	1	8			1						20
2	1		3	2			4				70
3	1		1	7			1				30
4					10						0
5					1	9					10
6							10				0
7	1							9			10
8	1							1	8		20
9										10	0

and test utterances were recorded during the same session. In each case 100 test utterances were used, 10 from each class. This allowed up to 4 reference templates per class. There are optimal ways to select templates from a give set of training templates (26), but for the results presented the templates were selected randomly. The "Ultrasonic Mike" and "Lip Lock Loop 1" results were recorded in sequence such as (0,0,...,1,1,...). Recording in this manner reduced the chances of moving the device between utterances. The "Lip Lock Loop 2" results were recorded in a random order, for example (3,2,4,1,...). In each case adding additional reference templates significantly increases the performance of the system, again at the cost of additional computations and memory. This improvement is related again to the position sensitivity of the devices tested. By using multiple templates, slight movements of the device are compensated for. Typical confusion matrices for the two ultrasonic devices are presented in tables 4.3 and 4.4. In this case certain digits have problems due mostly to the sensitivity of the devices. Note, the acoustic system results in different confusions from that of the automatic lip reader; therefore, a combined system should provide better accuracy than either individual system.

Table 4.2 Average automatic lip reader accuracies over 5 tests in percentages within one session. The Ultrasonic Mike and Lip Lock Loop 1 results were recorded sequentially and Lip Lock Loop 2 was recorded randomly.

Number of Templates	Ultrasonic Mike	Lip Lock Loop 1	Lip Lock Loop 2
1	75.2	77.2	49.4
2	80.2	85.2	51.4
3	86.2	87.8	62.0
4	89.2	89.0	65.8

Table 4.3 Typical confusion matrix for the "Ultrasonic Mike" using 4 reference templates.

Actual Class	Called										Error Rate (%)
	0	1	2	3	4	5	6	7	8	9	
0	8	1	1								20
1		9					1				10
2	2	1	4					2		1	60
3				10							0
4		3			7						30
5	1					8		1			20
6							10				0
7								10			0
8									10		0
9		1								9	10

Table 4.4 Typical confusion matrix for the "Lip Lock Loop" using 4 reference templates.

Actual Class	Called										Error Rate (%)
	0	1	2	3	4	5	6	7	8	9	
0	10										0
1		9					1				10
2		2	8								20
3		1		9							10
4					10						0
5		2				8					0
6							10				0
7				1			4	5			50
8					1	1	1		7		30
9									2	8	20

4.3 Combined Systems

4.3.1 Classifier Fusion. Figure 4.2 presents the typical results of classifier fusion with the "Ultrasonic Mike." The results presented are for a fixed λ value of 0.97. The value of 0.97 not only implies the acoustic signal was favored, but also compensates for the dramatic difference in the DTW scores between the automatic lip reader and the acoustic recognizer. For example in one case the average in class distance for the acoustic recognizer was approximately 100, whereas the automatic lip reader in class distance was around 45,000. These dramatically different numbers affect the pseudo probability mass function based classifier fusion. Therefore, λ is also a function of the different magnitudes of the DTW scores for the different classifiers. As can be seen from Table 4.5, some improvement in performance may be possible, by making λ a function of the SNR and best performance of the two classifiers, but overall a fixed λ provides good results. A promising feature of Figure 4.2 is that the combined results are always equal to or better than the best of the individual classifiers.

Figure 4.3 presents the results for the "Lip Lock Loop" using classifier fusion. The "Lip Lock Loop" results were almost identical to the "Ultrasonic Mike". However, in the "Lip Lock Loop" tests, there were some instances when the combined results were lower than the automatic lip reader alone. As can be seen from Table 4.6, a variable λ based on signal to noise ratio would overcome these inadequacies. Overall, including the lip information as derived from these ultrasonic devices is equivalent to approximately a 5dB increase in the signal to noise ratio.

4.3.2 Feature Fusion. Feature fusion is done by altering the mean and variance of the ultrasonic lip signal to levels proportional to the LPC features. Figure 4.4 presents the results of feature fusion with the "Ultrasonic Mike." These results were based on using a single reference template for each class and a fixed σ of 0.20. Results from various levels of σ are presented in Table 4.7. Unlike the classifier fusion the combined results were not always better than the best of the individual classifiers. Even if the best results for each signal to noise

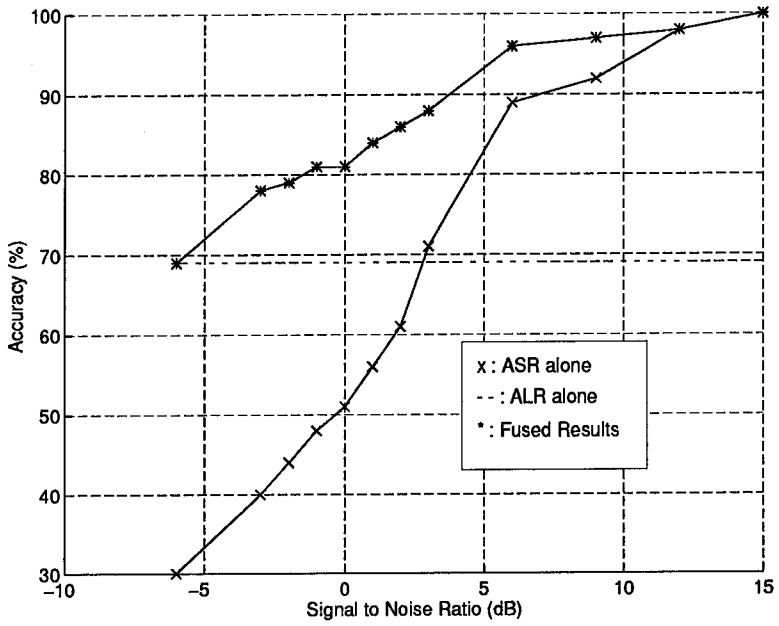


Figure 4.2 Classifier fusion with the "Ultrasonic Mike" for $\lambda = 0.97$ and 1 template. * : fused accuracy, $\times$: acoustic accuracy, and - - : automatic lip reader accuracy.

Table 4.5 Classifier fusion with the "Ultrasonic Mike" with 1 template, percentage accurate. Bold face indicates the two highest levels of accuracy for a given SNR.

	Values of λ									
SNR (dB)	1.0	0.99	0.98	0.97	0.96	0.95	0.94	0.93	0.92	0.91
15	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
12	**98**	**98**	**98**	**98**	**98**	**98**	**97**	**97**	**97**	96
9	92	**97**	**97**	**97**	**97**	**97**	**96**	**96**	**96**	94
6	89	**96**	**97**	**96**	**96**	94	93	93	92	91
3	71	83	**87**	**88**	**88**	**87**	86	**87**	**87**	85
2	61	79	83	**86**	**86**	**85**	**86**	**85**	84	83
1	56	71	82	**84**	**84**	**84**	**84**	83	81	81
0	51	69	79	**81**	**82**	**82**	**81**	80	80	79
-1	48	66	78	**81**	79	**81**	**80**	**80**	**80**	78
-2	44	61	72	**79**	**79**	**80**	**80**	78	78	76
-3	40	58	73	**78**	**77**	**77**	**77**	**78**	76	76
-6	30	42	61	69	72	**74**	**74**	**75**	**75**	**74**

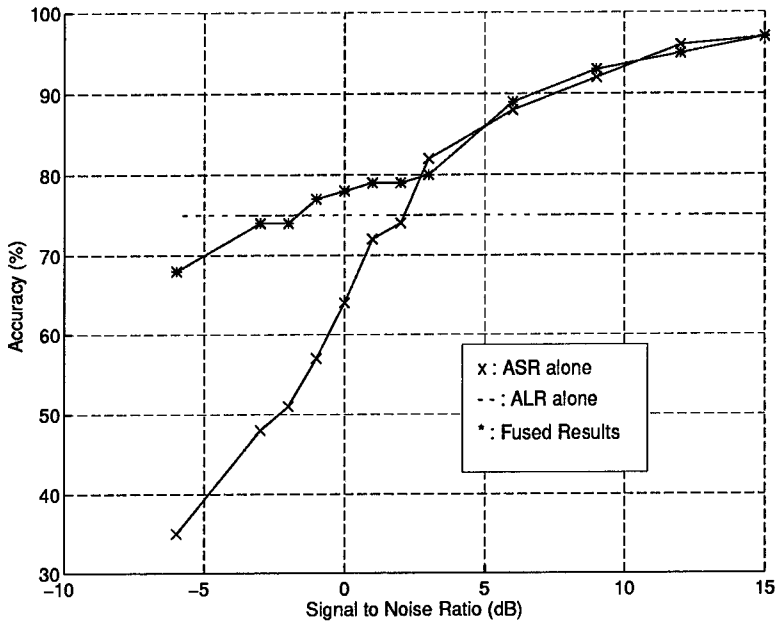


Figure 4.3 Classifier fusion with the "Lip Lock Loop" for $\lambda = 0.97$ and 1 template. * : fused accuracy, $\times$: acoustic accuracy, and - - : automatic lip reader accuracy.

Table 4.6 Classifier fusion with the "Lip Lock Loop" with 1 template, percentage accurate. Bold face indicates the two highest levels of accuracy for a given SNR.

	Values of λ									
SNR (dB)	1.0	0.99	0.98	0.97	0.96	0.95	0.90	0.85	0.80	0.70
15	**97**	**97**	**97**	**97**	**97**	**97**	**95**	91	89	88
12	**96**	**96**	**96**	**95**	**95**	**95**	92	89	88	87
9	**92**	**93**	**93**	**93**	**93**	**93**	87	86	86	85
6	88	**89**	**91**	**89**	**89**	86	83	82	81	80
3	82	**84**	81	80	80	81	**83**	82	81	80
2	74	**81**	80	79	79	79	**82**	**82**	**81**	79
1	72	74	78	79	79	79	**82**	**81**	**81**	79
0	64	72	77	78	79	79	**81**	**81**	**80**	79
-1	57	69	77	77	79	78	**80**	**81**	79	**80**
-2	51	63	73	74	76	77	76	**79**	78	**80**
-3	48	60	70	74	75	75	**78**	**79**	**78**	**79**
-6	35	47	60	68	70	73	76	**78**	**78**	**79**

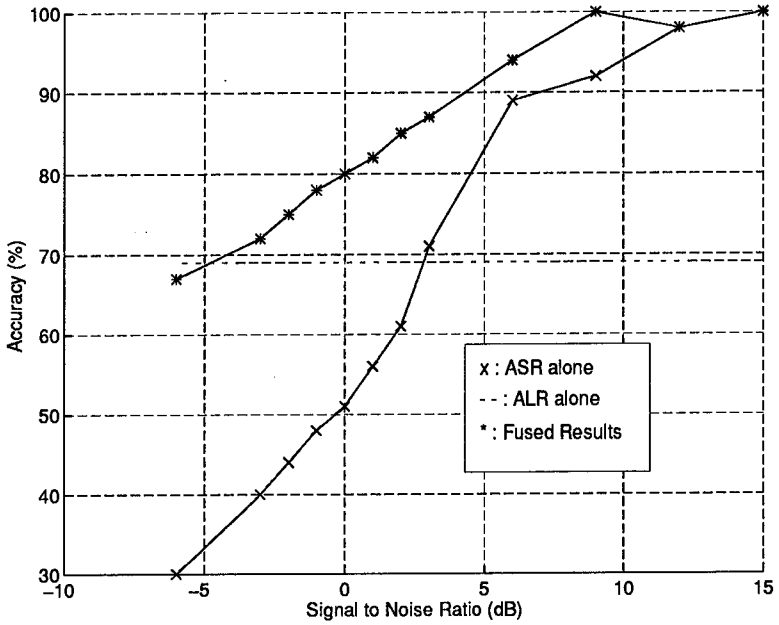


Figure 4.4 Feature fusion with the "Ultrasonic Mike" for $\sigma = 0.20$ and 1 template. * : fused accuracy, x : acoustic accuracy, and – : automatic lip reader accuracy.

ratio are chosen, there are some combined results that would be lower than the automatic lip reader alone.

Feature fusion results for the "Lip Lock Loop" are similar to the "Ultrasonic Mike." However, in this experiment using one template and a fixed σ of 0.20 resulted in numerous combined accuracies slightly worse, than the individual classifiers.

4.4 Conclusions

In this chapter the acoustic speech recognizer results are comparable to results found in the literature. The ALR results demonstrate the position sensitivity of the devices, resulting in cross session accuracies only slightly better than chance. The within session results for the automatic lip readers ranged from 75.2% to 89.2% depending on the number of reference templates used. Both fusion techniques provide improved performance in most cases, but classifier fusion provides more consistent improvement. In all but three instances, classifier fusion provided equal or superior accuracies than feature fusion. These results are probably

Table 4.7 Feature fusion with the "Ultrasonic Mike" with 1 template, percent accurate. Bold face indicates the highest two results for a given SNR.

	Values of σ								
SNR (dB)	0.00	0.05	0.10	0.15	0.20	0.25	0.50	0.75	1.00
15	**100**	**100**	**100**	**100**	**100**	**100**	**99**	98	95
12	**98**	**98**	**99**	**99**	**98**	**98**	92	90	87
9	92	96	95	**97**	**100**	96	89	87	83
6	89	93	**94**	**96**	**94**	93	85	80	77
3	71	80	**86**	85	**87**	**86**	80	77	75
2	61	76	82	**84**	**85**	**84**	78	76	75
1	56	71	78	**83**	**82**	**82**	77	74	71
0	51	60	75	**79**	**80**	**79**	77	73	70
-1	48	60	71	**78**	**78**	**75**	**75**	72	69
-2	44	56	65	**74**	**75**	**74**	**74**	70	68
-3	40	52	62	**73**	**72**	**73**	**73**	69	67
-6	30	37	57	64	**67**	66	**68**	66	64

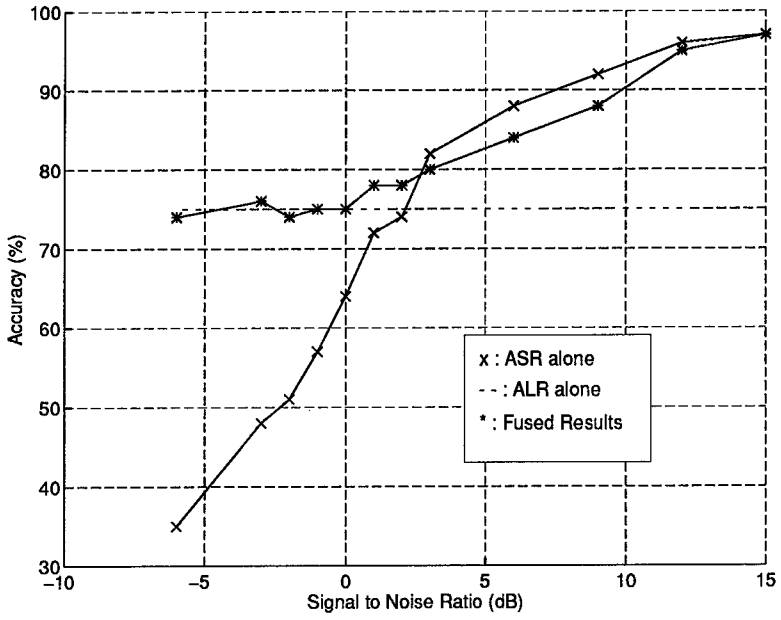


Figure 4.5 Feature fusion with the "Lip Lock Loop" for $\sigma = 0.20$ and 1 template. * : fused accuracy, x : acoustic accuracy, and – : automatic lip reader accuracy.

Table 4.8 Feature fusion with the "Lip Lock Loop" using 1 template, percent accurate. Bold faced numbers indicate two highest results for a given SNR.

	Values of σ								
SNR (dB)	0.00	0.05	0.10	0.15	0.20	0.25	0.50	0.75	1.00
15	**97**	**97**	**97**	**97**	**97**	**97**	90	85	84
12	**96**	**95**	**96**	**95**	**95**	93	87	83	80
9	**92**	**93**	**92**	**92**	88	88	84	81	80
6	**88**	**88**	**90**	85	84	85	81	80	80
3	**82**	**84**	**82**	80	80	80	80	79	78
2	74	**79**	**79**	**78**	**78**	**78**	**79**	**79**	76
1	72	75	**79**	77	**78**	76	**79**	77	76
0	64	73	75	74	75	**76**	**77**	**77**	75
-1	57	68	72	72	75	**76**	**77**	75	75
-2	51	65	71	71	74	**75**	**77**	**75**	**75**
-3	48	61	70	71	**76**	**75**	**75**	**75**	**75**
-6	35	51	61	**74**	**74**	**75**	**75**	**75**	**74**

due to the constraints imposed by feature fusion. In feature fusion the lip signal and the acoustic signal are time warped together, but classifier fusion allows separate warping of the two signals.

V. Conclusions and Recommendations

The goal of this thesis was to (1) demonstrate that a simple ultrasonic device can provide lip information that can be used for automatic lip reading and (2) to demonstrate an effective way to include the automatic lip reader results into a traditional automatic speech recognizer. Based on the results of tests of the two ultrasonic devices examined in this thesis a number of conclusions and recommendations for future work can be made. The first section of this chapter summarizes the successes and benefits of automatic lip reading with a simple ultrasonic device. The second section presents some of the problems that still need to be overcome to make automatic lip reading practical for implementation. The third section discusses possible future directions for this work.

5.1 Successes and Benefits

Clearly this research has demonstrated that a simple device can be used to improved automatic speech recognition in a noisy environment. The automatic lip reader alone under constrained conditions was fairly accurate, achieving accuracies as high as 89% . With further development a simple automatic lip reader could provide high accuracy isolated word recognition in situations where the acoustic channel is impossible to use. Making possible applications in high noise situations such as in the cockpit of a fighter aircraft or in situations where conflicting speech is a problem such as in a large crowded office. These devices also offer tremendous computational savings over video, the favorite source of other lip readers, reducing the data rate by as much as 12500 to 1. This research has also shown that a simple fusion algorithm can be used to combine the results of an automatic lip reader and an automatic speech recognizer, resulting in improved accuracy at all signal to noise ratios. The combined results were equivalent to approximately a 5dB increase in the signal to noise ratio at the 0dB signal to noise level. Both classifier and feature level fusion were shown to provide effective fusion of the lip information into the system. The classifier level fusion in most cases provided the best overall accuracy in the tests conducted in this research effort.

5.2 *Problems*

Even with the success of this research there are many obstacles still to overcome. The devices investigated were too sensitive to position, thereby making them impractical for implementation as is. The sensitivity of the devices is partly due to the fact that they react to any changes in the contour of the face and many of the contour changes of the mouth contain no relevant lip information. The idea features according to lip reading science are vertical lip separation and degree of pucker. Another problem is determining the best values of λ and σ, used in the fusion process. For implementation an effective way to automatically choose λ and σ, to optimally fuse the data, needs to be found.

5.3 *Future Work*

There are many areas that still need to be explored before effective automatic lip reading will become practical. In the future, new devices need to be explored. A device needs to found that can extract the best features of the mouth, such as vertical lip separation and lips puckered-extended (3, 9), without the computational expense of video. One possible device considered for future work might involve laser scanning (13). A combination of scanning and ranging could provide both the features listed above. Whatever device is explored, more concentration on phoneme level recognition needs to be made. Phoneme level recognition would allow fusion to many current systems and allow consideration of connected or continuous speech recognition. Another area that needs to be explored is the effect of environmental noise on an automatic lip reader. It is well known that people alter their speech when speaking in a noisy environment, this is known as the Lombard effect (5, 34). The Lombard effect may certainly have an effect on an automatic lip reader.

5.4 *Final Words*

Automatic lip reading has the potential to revolutionize automatic speech recognition and make possible many new applications. This research has demonstrated that a simple

device can provide the necessary lip information for automatic recognition. It is now only a matter of time before automatic lip reading will become a reality in commercial applications.

Appendix A. Isolated Word Recognition in ESPS

This appendix gives the details of how the Entropic Signal Processing Software (ESPS) is used in this research thesis. ESPS is a software package well suited for developing and testing automatic speech recognition. The program includes 85 user-level programs that can be called from a UNIX shell and includes a library of over 200 functions. This software allows quicker design and testing of automatic speech recognition systems and also readily allows subsequent researchers to repeat or continue the work done in this project. It is also easy to switch between ESPS, MATLAB, and ASCII format, allowing user written algorithms to be incorporated into the program. Although the ESPS program is extensive, the basic automatic recognition systems used in this thesis requires only a few of the algorithms available. The first section of this appendix presents the feature extraction methods. The second section presents the dynamic time warping program. The third section discusses the feature fusion methods used. The fourth section presents the methods used to determine the results and perform classifier level fusion. To assist in the explanation of the ESPS algorithms used, first a brief overview of the file management and file name conventions is necessary. The raw speech files have names such as d73, which is utterance number 7 of digit 3. The fourteen utterance numbers for each recording are (0,1,2,3,4,5,6,7,8,9,a,b,c,d). The orginal speech file contains both the speech and ultrasonic signal. Using the ESPS "demux" command the two signals are separated into two files with extensions ".a" and ".b", corresponding to the two channels recorded. Each recording session was maintained in a unique directory and each class of utterances was kept in a separate subdirectory. Typically in ESPS the file extension ".sd" implies a raw speech file and ".fea" implies a feature file.

A.1 Feature Extraction

ESPS has a feature extraction algorithm called "ACF." The program is specifically designed to extract acoustic features. The following Unix C-shell, entitled "xtract" is an example of how the ACF command is used to extract features. This C-shell interates through

a number of different levels corresponding to white gaussian noise levels that are added to the orginal signal. The ESPS program "TESTSD" was used to create the Gaussian noise signal. The parameter "count" specifies a seed for the noise generation and is incremented to result in different noise signals with the same noise level. The noise signal is added to the original signal prior to feature extraction. Inputs to the ACF command include a parameter file, an input speech file, and an output feature file. The parameter file specifies the feature extraction method to be used. A typical parameter file used in this thesis is also presented below. The program therefore transforms each raw speech signal file to a feature file. The C-shell "xtract" demonstrates a batch method to convert a large number of the raw speech files to feature files. Futher details of the ESPS commands can be found in the users manual which can be accessed online from the UNIX prompt with "eman (command)."

A.2 Dynamic Time Warping

The dynamic time warping algorithm in ESPS is called "DTW_REC". The algorithm is based on the results of Rabiner (25) and allows both constrained and unconstrained endpoints.

A typical command line for envoking this program is:

dtw_rec -P Params/dtw_params Lists/ref_list Lists/test_list Rslts/rslt_test1

The parameter file Params/dtw_params specifies which features are going to be warped, how many ranked candidates to include in the output list, and the maximum offset allowed between the two signals being tested. The Lists/ref_list file contains a list of the feature files to be used as reference templates. The Lists/test_list file contains a list of the feature files to be tested. The program takes each file listed in Lists/test_list and finds the DTW distances to each of the reference files listed in Lists/ref_list. The program then outputs the result to the file, Rslts/rslt_test1. A section of a typical output file is also given below. This program took the greatest amount of time to run. A typical run lasted 20-30 minutes on a Sparc 10, depending on the number of reference templates used. The results of the dynamic time warping program are used to determine classification performance.

A.3 Feature Fusion

If more than one kind of feature is to be used, the features must be fused prior to running the DTW program. The C-Shell entitled "warpdrv" is an example of how feature fusion was done for this research. The program interates through various levels of noise and values of σ corresponding to the normalization factor used. The program starts off by transforming the ultrasonic feature file to ASCII format, using the ESPS command "pplain". Then a C program, "normaliz", was written to adjust the mean and standard deviation of the ultrasonic feature to a new level. The parameter file for the "normaliz" program was used to specify the desired mean and standard deviation. After than the ESPS command "addfea" was used to create a new feature file consisting of the LPC features of speech and the normalized ultrasonic feature. Finally the features were fused into a common feature vector using the ESPS command "fea_deriv". Again the details of the ESPS commands can be found in the users manual. Now the DTW program can be used as before to determine classification performance. The C-Shell "warpdrv" is ran in the background and the results sent to a log file, allowing extensive runs to be performed automatically.

A.4 Evaluating the Results

As was pointed out above the output of the dtw_rec program is a file that contains all of the information necessary for classification and classifier level fusion. An example of a portion of a typical output file is presented below. Each line contains a test file, a reference file, and the DTW distortion calculated for that pair. To facilitate the evaluation of the results, a C program, "pmfclass", was written to extract the results and perform the necessary transformations for classifier fusion. The program also calculates the overall accuracy and confusion matricies.

Listing for C-Shell : xtract

```
#!/bin/csh
#
#
foreach level (142 466 839 942 1058 1188 1334 1886)
if ($level == 142) set dir = a
if ($level == 466) set dir = d
if ($level == 839) set dir = j
if ($level == 942) set dir = f
if ($level == 1058) set dir = k
if ($level == 1188) set dir = l
if ($level == 1334) set dir = g
if ($level == 1886) set dir = h
set count = 100
@ x = 0
while ($x < 10)
testsd -l $level -S $count -s 1.5 -T gauss -r 16000 noise.sd
@ count ++
addsd noise.sd Digits6/$x/d0$x.a rslt.sd
acf -P Params/ac2_params rslt.sd
Digits6/features/$dir/$dir\0$x.fea
testsd -l $level -S $count -s 1.5 -T gauss -r 16000 noise.sd
@ count ++
addsd noise.sd Digits6/$x/d1$x.a rslt.sd
acf -P Params/ac2_params rslt.sd
Digits6/features/$dir/$dir\1$x.fea
testsd -l $level -S $count -s 1.5 -T gauss -r 16000 noise.sd
@ count ++
```

```
addsd noise.sd Digits6/$x/d2$x.a rslt.sd
acf -P Params/ac2_params rslt.sd
Digits6/features/$dir/$dir\2$x.fea
testsd -l $level -S $count -s 1.5 -T gauss -r 16000 noise.sd
@ count ++
addsd noise.sd Digits6/$x/d3$x.a rslt.sd
acf -P Params/ac2_params rslt.sd
Digits6/features/$dir/$dir\3$x.fea
testsd -l $level -S $count -s 1.5 -T gauss -r 16000 noise.sd
@ count ++
addsd noise.sd Digits6/$x/d4$x.a rslt.sd
acf -P Params/ac2_params rslt.sd
Digits6/features/$dir/$dir\4$x.fea
testsd -l $level -S $count -s 1.5 -T gauss -r 16000 noise.sd
@ count ++
addsd noise.sd Digits6/$x/d5$x.a rslt.sd
acf -P Params/ac2_params rslt.sd
Digits6/features/$dir/$dir\5$x.fea
testsd -l $level -S $count -s 1.5 -T gauss -r 16000 noise.sd
@ count ++
@ x ++
end
end
```

Typical Parameter File for ACF acoustic classifier

string sd_field_name = ''samples'';

float preemphasis = 0.950000;

float frame_len = 160.000000;

float step = 80.000000;

string window_type = ''HAMMING'';

string units = ''samples'';

int ac_order = 10;

int lpccep_flag = 1;

string lpccep_fname = ''lpc_cepstrum'';

int lpccep_order = 10;

Listing for C-Shell : warpdrv

```
#!/bin/csh
#
#
set dir = Digits6/features
foreach sigma  (2 7 8)
echo start sigma = $sigma
date
foreach level  (A B C D E I J F K L G H)
echo start level = $level
date
foreach a (1 2 3 4 5 6 7 8 9)
foreach b (0 1 2 3 4 5 6 7 8 9 a b c d)
set ufile = $dir/U/u$b$a.fea
set afile = $dir/$level/$level$b$a.fea
set cfile = $dir/Z/c$b$a.fea
pplain $ufile > temp
normaliz normparm$sigma temp temp2
addfea -f samples -c all temp2 $afile temp.fea
fea_deriv field_file temp.fea temp2.fea
mv temp2.fea $cfile
end
end
echo start dtw for level sigma = $level $sigma
date
dtw_rec -P Params/dtw_parm Lists/ref_6a Lists/test_6
Rslts/comb_6
echo results for level sigma = $level $sigma 4 templates
```

```
Rslts/pmfclass Rslts/pmfparm2 Rslts/comb_6 Rslts/comb_6
dtw_rec -P Params/dtw_parm Lists/ref_6b Lists/test_6
Rslts/comb_6
echo results for 1 template
Rslts/pmfclass Rslts/pmfparm2 Rslts/comb_6 Rslts/comb_6
date
end
end
```

Portion of a result file from dtw_rec"

dtw_rec: Digits6/features/A/A40.fea Digits6/features/A/A30.fea 1.092519e+02

dtw_rec: Digits6/features/A/A40.fea Digits6/features/A/A34.fea 1.193438e+02

dtw_rec: Digits6/features/A/A40.fea Digits6/features/A/A31.fea 1.196075e+02

dtw_rec: Digits6/features/A/A40.fea Digits6/features/A/A32.fea 1.245525e+02

dtw_rec: Digits6/features/A/A40.fea Digits6/features/A/A37.fea 1.250912e+02

dtw_rec: Digits6/features/A/A40.fea Digits6/features/A/A33.fea 1.260365e+02

dtw_rec: Digits6/features/A/A40.fea Digits6/features/A/A35.fea 1.281371e+02

dtw_rec: Digits6/features/A/A40.fea Digits6/features/A/A36.fea 1.376709e+02

dtw_rec: Digits6/features/A/A40.fea Digits6/features/A/A39.fea 1.395755e+02

dtw_rec: Digits6/features/A/A40.fea Digits6/features/A/A38.fea 1.640215e+02

dtw_rec: Digits6/features/A/A41.fea Digits6/features/A/A31.fea 1.043715e+02

dtw_rec: Digits6/features/A/A41.fea Digits6/features/A/A34.fea 1.180687e+02

dtw_rec: Digits6/features/A/A41.fea Digits6/features/A/A30.fea 1.218792e+02

dtw_rec: Digits6/features/A/A41.fea Digits6/features/A/A37.fea 1.225160e+02

dtw_rec: Digits6/features/A/A41.fea Digits6/features/A/A35.fea 1.236908e+02

dtw_rec: Digits6/features/A/A41.fea Digits6/features/A/A32.fea 1.236928e+02

dtw_rec: Digits6/features/A/A41.fea Digits6/features/A/A33.fea 1.240421e+02

dtw_rec: Digits6/features/A/A41.fea Digits6/features/A/A39.fea 1.275375e+02

dtw_rec: Digits6/features/A/A41.fea Digits6/features/A/A36.fea 1.284683e+02

dtw_rec: Digits6/features/A/A41.fea Digits6/features/A/A38.fea 1.509813e+02

Appendix B. Simulation of the "Ultrasonic Mike"

This appendix describes the results of a simulation of the "Ultrasonic Mike." The simulation is designed in Matlab, a numeric computation and visualization software package. This appendix describes the design of the simulation and the resulting ramp, step, and impulse responses. The simulated results are also compared to the recorded results of the device when possible.

B.1 Design

Figure B.1 gives an overview of how the device works. The device using a 40Khz oscillator drives a piezoelectric material which creates an ultrasonic signal. The ultrasonic signal is directed at the talker's mouth, repeated reflections from the talker's mouth and the device establish a standing wave. Any movement of the mouth results in a change in the standing wave. An ultrasonic receiver also located in front of the talker's mouth converts the ultrasonic signal back to an electrical response. The electrical signal is passed through an envelope detector. The result is a device that outputs a signal that changes in response to the movement of the talker's mouth and is zero when ever the talker's mouth is still.

It would be extremely difficult to design a simulation to accurately depict the changing contour of a talker's mouth; therefore, the simulation is designed to model the reflection off a flat surface. The goal of this simulation is to demonstrate and evaluate the sensitivity of the device with respect to position. In the following sections a ramp, step, and impulse type input are considered. In each case the same input is presented starting at slightly different distance from the device. The results are cyclic with a period equal to the wavelength of the device. Operating at 40Khz, the wavelength is approximately 0.85cm. In the following examples, the starting position is moved by increments, dx, of $\frac{1}{3}$ of a wavelength and the corresponding outputs are calculated.

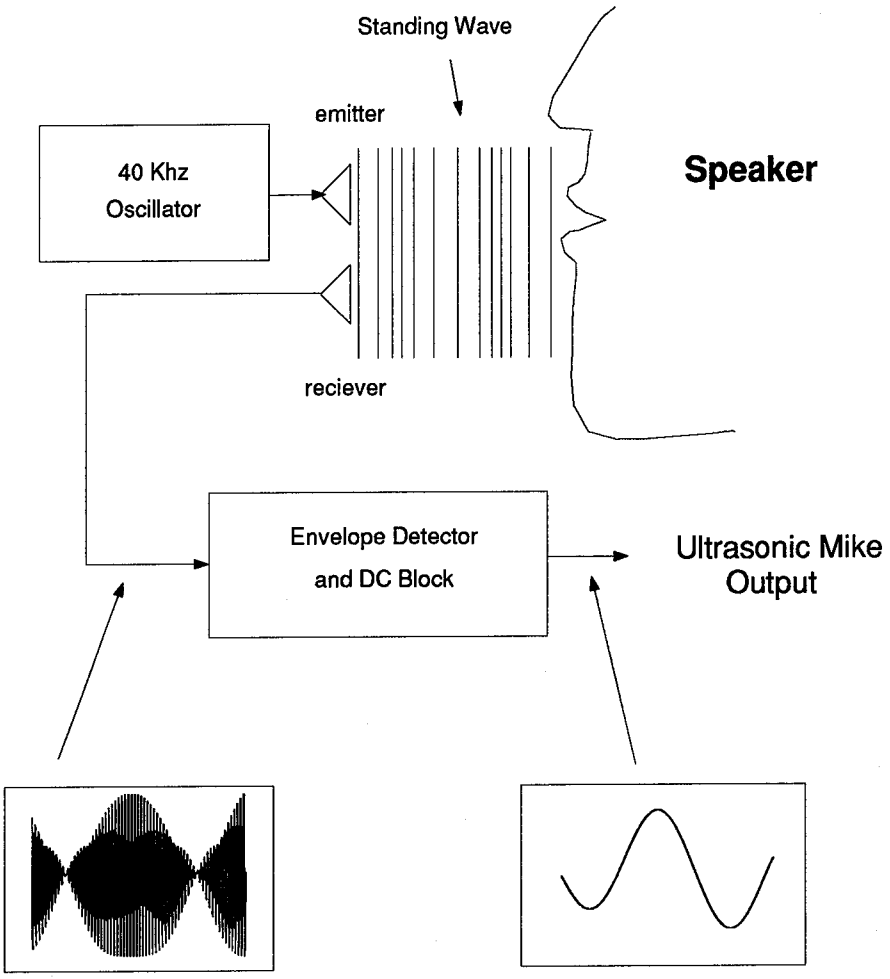


Figure B.1 Overview of operation of the "Ultrasonic Mike."

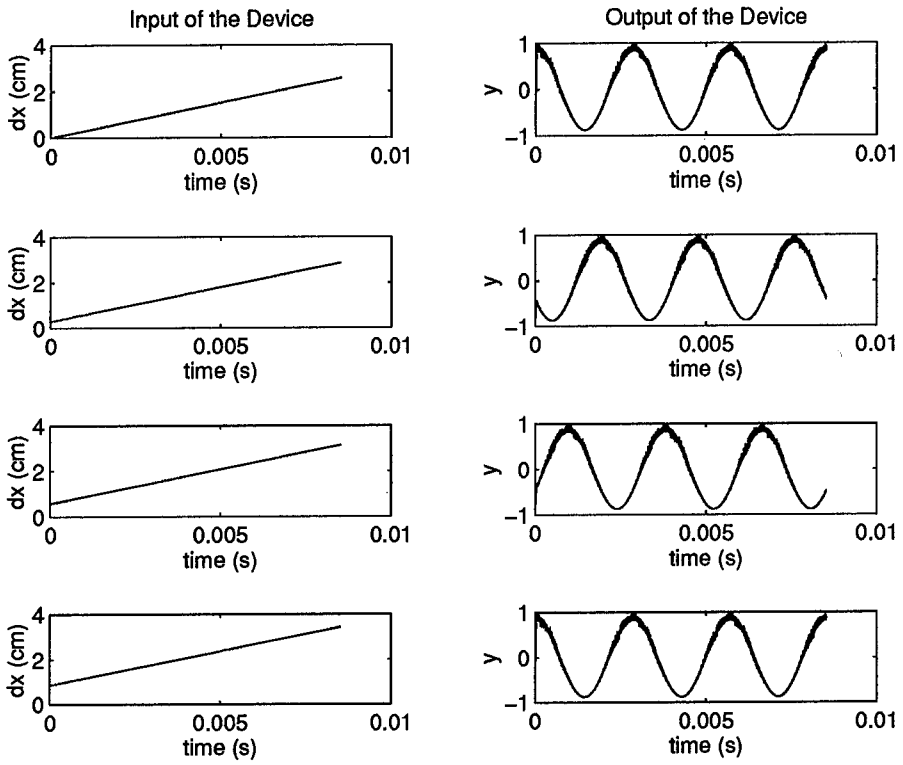


Figure B.2 Simulated response of a ramp imput.

B.2 Ramp Input

The ramp input seen in Figure B.2 ranges from 0 to 2cm. The starting distance is at one wavelength, but the same results are achieved for any integer multiple of wavelengths. The only apprarent difference in the outputs is a phase shift. A similar response is observed by hooking the output of the actual device to an osciloscope and moving a flat object away from or toward the device.

B.3 Step Input

The input for the step response, as seen in Figure B.3, is a step of 0.2cm toward the device. As before, four different starting positions are shown corresponding to changes, dx, of $\frac{1}{3}$ of a wavelength. As can be seen from the figure the output changes in magnitude and sign depending on the starting position of the input. To compare the actual device with the simulation a response was recorded by placing a flat surface in from of the device and during

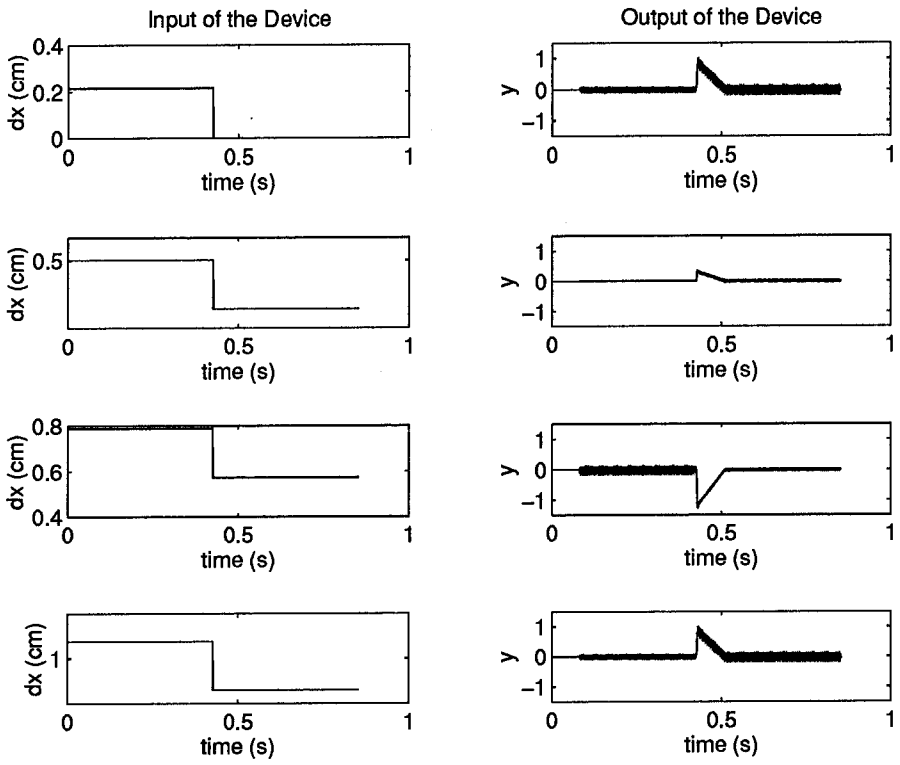


Figure B.3 Simulated response of a step input.

the recording quickly reducing the distance to the surface by a fixed amount. As can be seen in Figure B.4 the actual recorded step resonse is much more complicated than the simulation. The output of the actual device appears to ring up and down for approximately three tenths of a second before settling.

B.4 Impulse Response

The impulse was simulated as a brief change in the distance of the flat surface from the device. As is shown in Figure B.5 the magnitude of the impulse like input was chosen to be 0.2cm. As with the step response, the output was significantly different depending of the starting position. In this case the simulated and recorded responses are similar. A impulse like recording was made by quickly passing a pencil between the device and a flat surface.

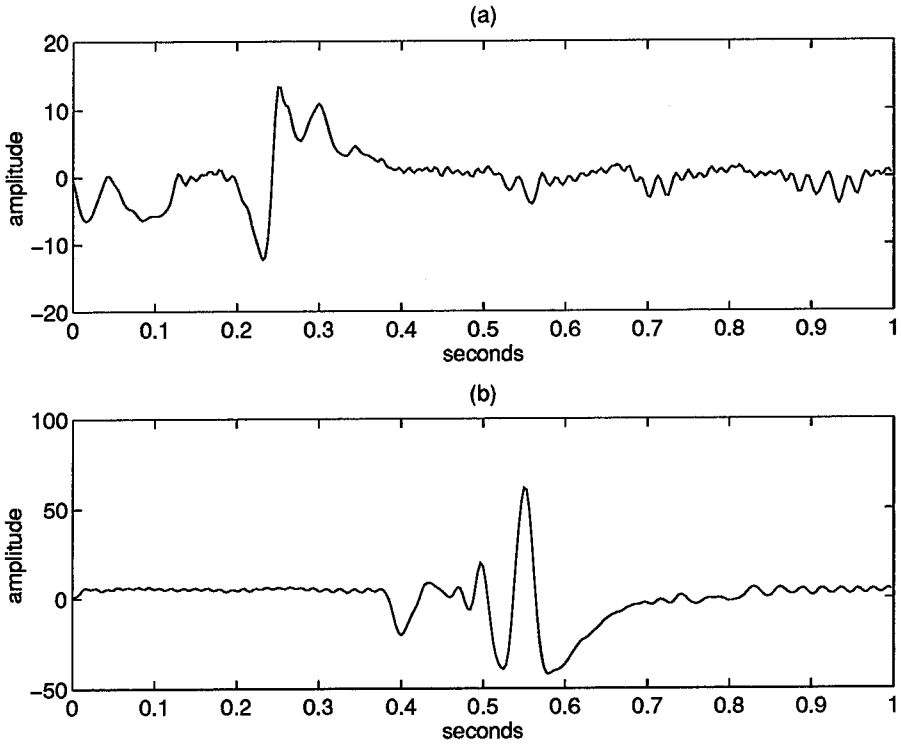


Figure B.4 (a) Step response and (b) Impulse response of the "Ultrasonic Mike"

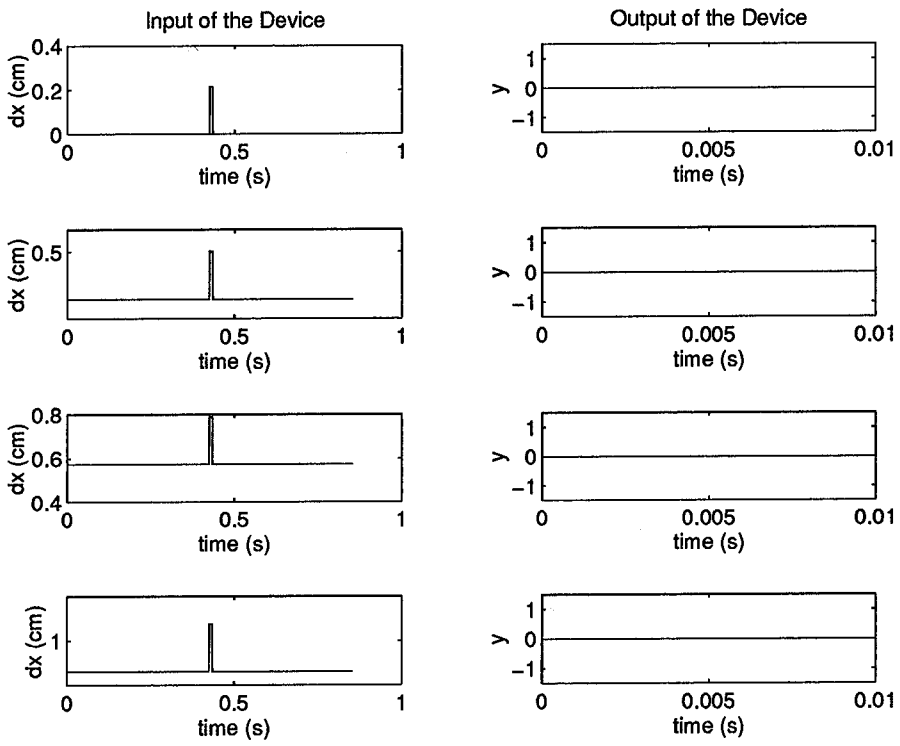


Figure B.5 Simulated response of a step imput.

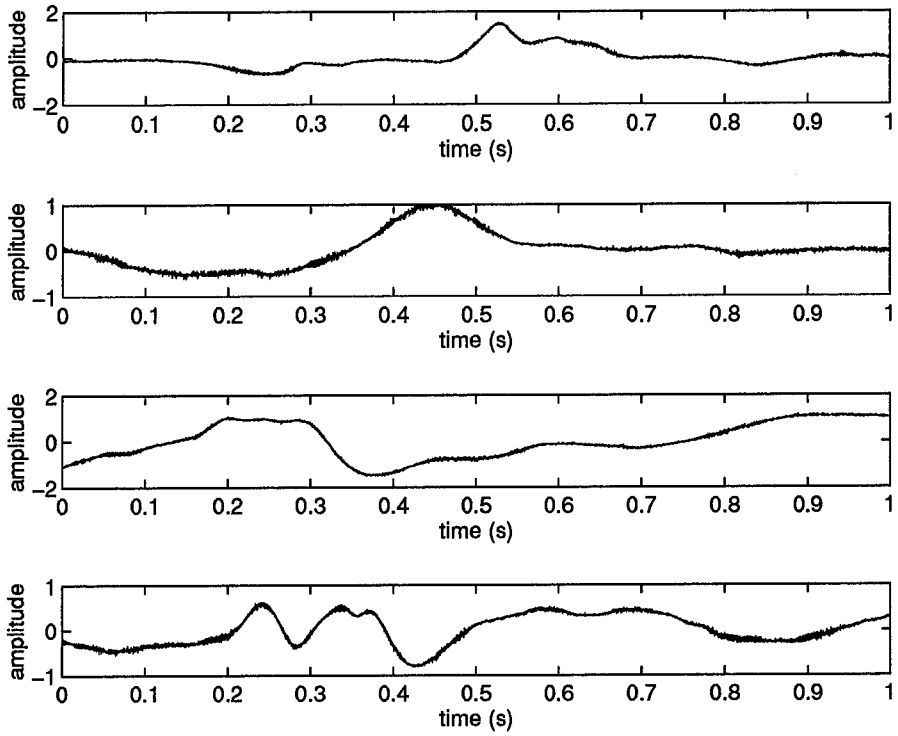


Figure B.6 Output of the "Ultrasonic Mike" from different device positions relative to the talker's mouth.

B.5 Conclusion

Although this is a very basic simulation it does clearly demonstrate the position sensitivity of the device. Therefore, unless the device is maintained at a fixed position with respect to the face, the output will be inconsistent and useless for recognition. Figure B.6 gives the results of multiple recordings of the digit zero from different positions. Obviously changing the position of the device relative to the talker's mouth drastically changes the output.

Appendix C. Classifier Fusion using Fuzzy Logic

This appendix demonstrates a classifier fusion method based on fuzzy sets derived from dynamic time warping distances. Two procedures are given for converting the similarity measurements of dynamic time warping to fuzzy sets. Fusion results from the combination of an automatic lip reader and automatic speech recognizer demonstrate that fuzzy sets derived from the relative dynamic time warping distances are superior to ones derived from the raw dynamic time warping distances.

C.1 Introduction

Classifier fusion is an important problem in many current pattern recognition applications. The basic idea behind classifier fusion is to combine a number of distinct classifiers in order to increase accuracy of the overall classification. Figure C.1 is an overview of the basic classifier fusion problem. Based on the type of individual classifiers used, different levels of information may be available for the fusion. Xu explains three different levels as abstract, rank, and measurement (39). At the abstract level the individual classifier only outputs the most likely classification. At the rank level the classifier outputs a subset of possible classes in order from most likely to least. At the measurement level the classifier outputs a subset of possible classes and a measurement indicating the likelihood of each rank. All classifiers output information at least at the abstract level and most are capable of outputting information at the measurement level. Xu *et. al* primarily investigate classifiers that output information at the abstract level. This appendix investigates, more thoroughly, fusion of classifiers that output information at the third or measurement level. In particular, this appendix will investigate methods of converting dynamic time warping distances to fuzzy sets. For an explanation of fuzzy logic see (4) and for an explanation of dynamic time warping see (21, 27).

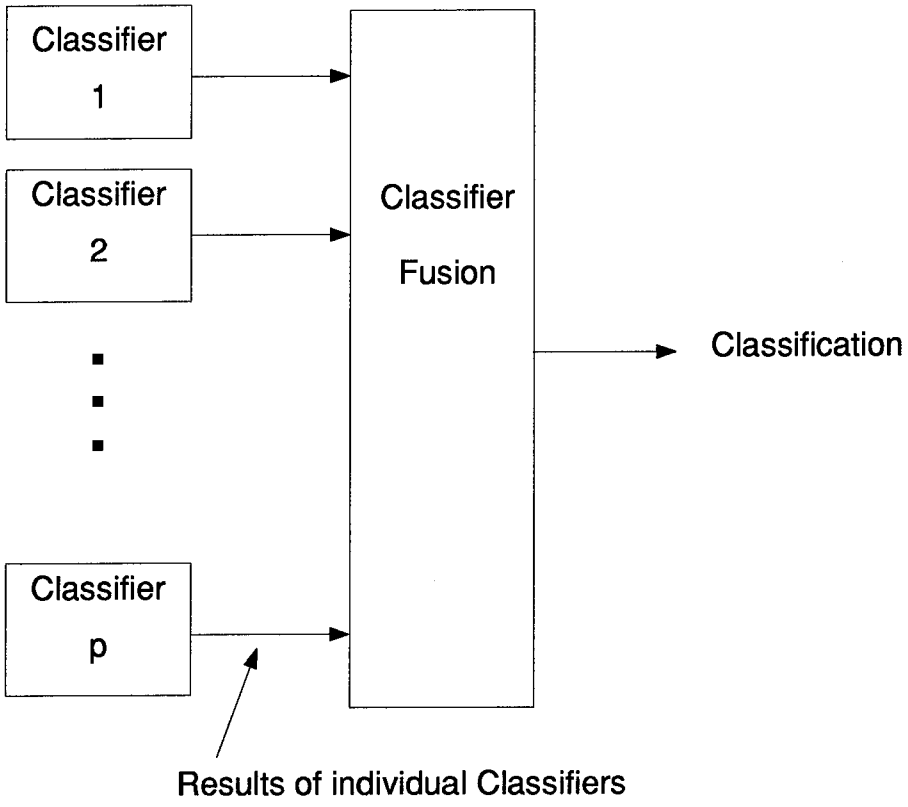


Figure C.1 Generic Classifier Fusion.

C.2 Theory

In this section two different methods of converting the dynamic time warping (DTW) distances to fuzzy set are presented. The first method is derived from the raw distance measurements, and the second is based on the relative differences between the distances. Basically, the output of a dynamic time warping procedure is a measurement of the similarity between two signals. The more similar the two signals are the lower the DTW distance between them. In a classifier based on DTW, a number of reference templates is compared to the test utterance and the class of the closest matching reference template is assigned to the test utterance. At the abstract level, the classifier outputs the class of the closest matching reference template. However, the classifier is capable of outputting information on the measurement level. The classifier could output the closest DTW distance to each class. These similarity measurements can then be converted to fuzzy set memberships for each class.

C.2.1 Memberships derived from the raw DTW distances. Xu's paper illustrates one way to convert these numbers to a pseudo probability mass function, but this could also be viewed as a way to convert the numbers to a fuzzy set membership. The membership numbers for class i and classifier a, $p_a(i)$, is computed as follows,

$$p_a(i) = \frac{\frac{1}{d_a(i)}}{\sum_{i=1}^{M} \frac{1}{d_a(i)}}, \tag{C.1}$$

where $d_a(i)$ is the distance from the test utterance to class i for classifier a, and M is the total number of classes. The problem with using raw distances is that depending on the magnitude of the signal, the raw distances from classifier to classifier can be drastically different. These differences in the raw distance magnitudes can lead to incorrect weighting of the fuzzy membership. This problem will be demonstrated in the experimentation section.

C.2.2 Memberships derived from the differences in the DTW distances. Another method to convert the DTW distances to fuzzy set membership is to use the relative magnitude of the DTW distances rather than the raw distances. This method requires the calculation of

the average and estimated variance of the difference between the "in class" and closest "out of class" DTW distances. This information can be derived from the training or reference templates. The basic idea behind this procedure is to capture the relative significance of the differences between the DTW distances. The procedure is as follows:

1. Use the reference templates to find the average difference, μ_x, and the estimated variance of the difference, σ_x, between the "in class" and closest "out of class" distances.
2. Start off by letting the membership number of class of the closest reference template, (the one with the lowest DTW distance), $p_a(0) = 1$.
3. Calculate a vector of differences from the lowest distance to the highest distance, $x_a(i)$, as

$$x_a(i) \;=\; d_a(i+1) - d_a(i) \quad i = 1, \ldots, M-1.$$

   Where $d_a(i)$ is the dynamic time warping distance from the test utterance for rank i for classifier a and M is the total number of classes. A separate vector is used to key track of what class each rank corresponds to.
4. In this step the membership numbers are adjusted iteratively as each of the differences in the DTW distances are considered. At this point an equation is needed to determine what share of the membership each subsequent class is to receive. In this appendix the function used was,

$$p_a(j) \;=\; \tfrac{j}{j+1} exp(\tfrac{-x_j}{\sigma_x}) p_a(j-1) \quad j = 1, \ldots, M-1.$$

   After each iteration the previous memberships are reduced as follows,

$$p_a(k)^+ \;=\; p_a(k)^- - \tfrac{1}{j} p_a(j) \quad k = 0, \ldots, j-1.$$

Both of these methods result in equal memberships in all classes if the DTW distance are all equal. In essence, each classifier is given one vote, instead of giving the entire vote to one

class, as is required in the abstract level, the vote is split according to the DTW distances from each class.

C.3 Experimentation

The thrust of this investigation was to determine the benefits and problems with these two methods of converting the DTW distances to fuzzy sets. In the following experiment two classifiers are considered for fusion, an automatic lip reader and an automatic speech recognizer. The automatic speech recognizer is tested under various signal to noise ratios while the automatic lip reader was held to a constant signal to noise ratio. Both classifiers are tested as isolated digit recognizers and output a ranked list of all classes and the corresponding DTW distances.

C.3.1 Memberships derived from the raw DTW distances. Figure C.2 represents a typical results of the fuzzy membership based on the raw DTW distances. For these results the signal to noise for the automatic speech recognizer (ASR) was approximately 15dB and achieved nearly 100% accuracy, but the membership was nearly equal for all the classes. The combined membership is calculated as the average of the ALR and ASR memberships. In this situation the automatic lip reader (ALR) is receiving two much emphasis and dominates the final decision. In this example the raw distances for the ASR system are around 100, whereas the raw distances for the ALR are approximately 45000. These different magnitudes have a detrimental affect on the calculation of the fuzzy memberships. Figure C.3 is the same, except the signal to noise for the ASR is now approximately -6dB. In this situation the membership for the ASR should be, and is, nearly flat allowing the ALR to dominate. The combined membership could be calculated as linear combination of the ALR and ASR memberships, with the weights determined by the signal to noise of the two individual classifiers.

C.3.2 Memberships derived from the differences in the DTW distances. Figures C.4 and C.5 represent the typical results using the fuzzy sets derived from the differences in the DTW distances. The variance of the difference is calculated from the 15dB signal to noise

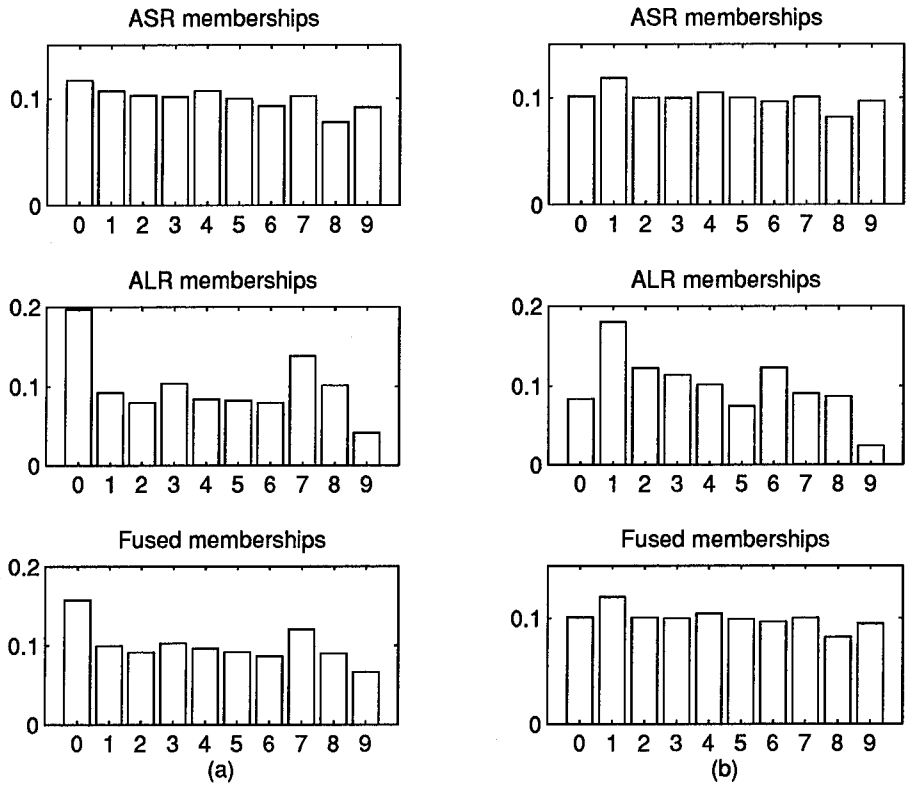


Figure C.2 Memberships calculated from raw distances (a) actual class is 0 (b) actual class is 1. Signal to noise 15dB.

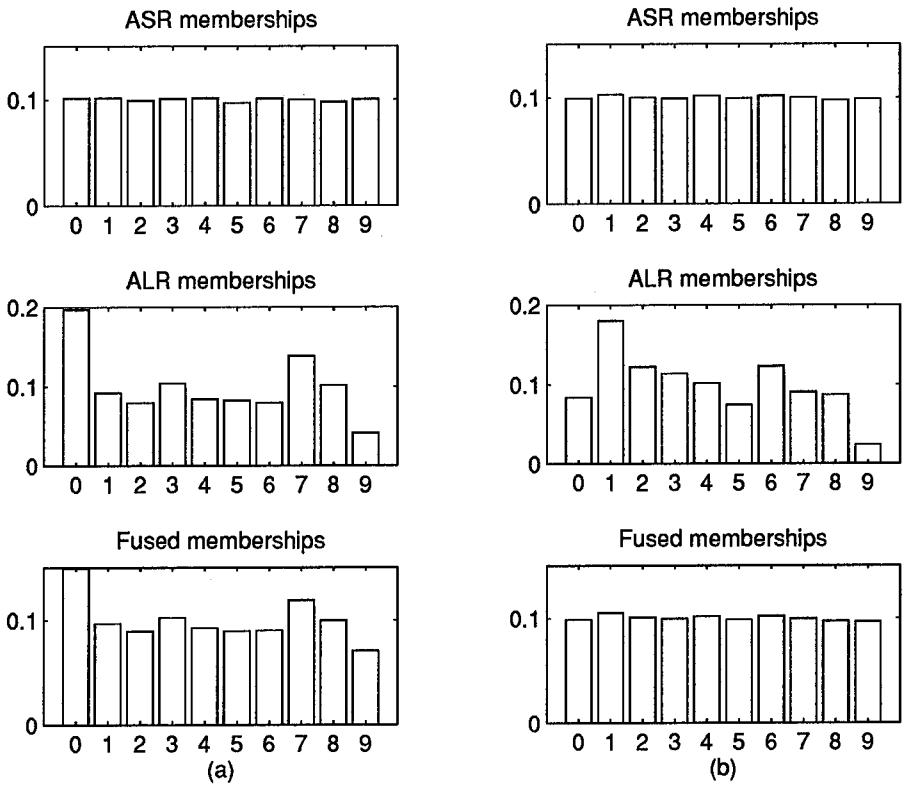


Figure C.3 Memberships calculated from raw distances (a) actual class is 0 (b) actual class is 1. Signal to noise -6dB.

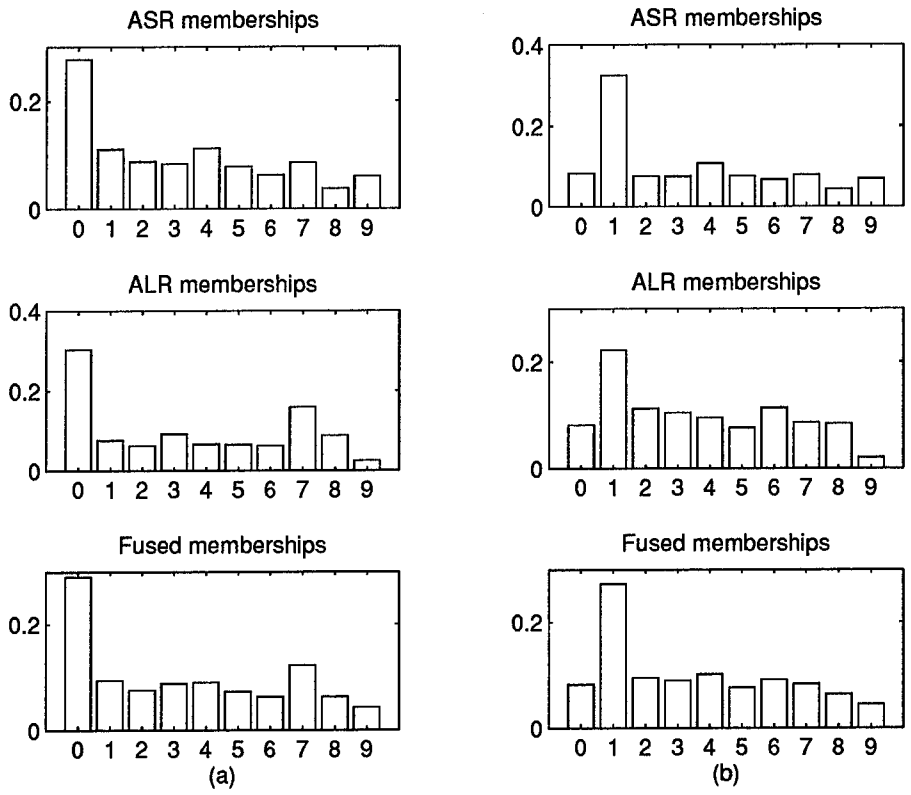


Figure C.4 Memberships calculated from differences in distances (a) actual class is 0 (b) actual class is 1. Signal to noise 15dB.

reference templates. The same two digits are shown as before. At the 15dB signal to noise ratio, the ASR is now favored and at the -6dB signal to noise ratio the ASR is flat again allowing the ALR to dominate. The combined fuzzy membership was determined as the average of the ALR and ASR memberships.

C.3.3 Comparison of two methods. Table C.1 summarizes the results of the two methods for various signal to noise ratios. In each case 100 test utterances were used. For the fuzzy memberships derived from raw data the combined membership was derived as,

$$P(i) = \lambda p_{asr}(i) + (1-\lambda)p_{alr}(i) \quad i = 1, \ldots, M$$

where p_{asr} and p_{alr} are the fuzzy memberships for the ASR and ALR systems respectively and M is the total number of classes. The Method 1 results presented in the table are the best

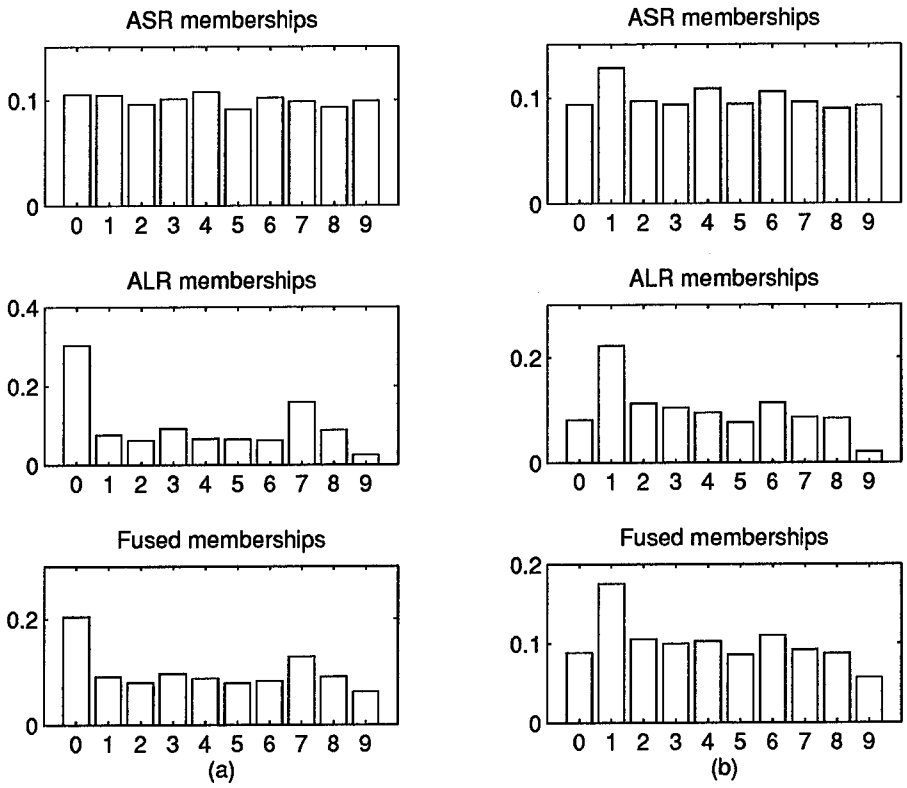


Figure C.5 Membership calculated from differences in distances (a) actual class is 0 (b) actual class is 1. Signal to noise -6dB.

Table C.1 Accuracies of ASR for various levels of signal to noise and fusion bases on Method 1 (membership based on raw DTW scores) and Method 2 (membership based on differences in DTW scores). The results of Method 1 were the best over all λ.

SNR	1 template ALR alone 82%			4 templates ALR alone 84%		
	ASR	METHOD 1	METHOD 2	ASR	METHOD 1	METHOD 2
15	97	97	97	96	95	96
12	96	96	95	96	95	95
9	92	93	93	95	93	93
6	88	89	88	93	93	93
3	82	84	87	91	91	90
2	74	82	87	88	89	88
1	72	82	86	82	88	88
0	64	82	84	79	87	88
-1	57	81	84	72	87	87
-2	51	80	82	66	87	87
-3	48	80	81	60	87	86
-6	35	79	79	42	83	84

results over all λ. The systems were also tested using both one template and four templates as references. Using more reference templates typically increases the accuracy of the system. As can be seen from Table C.1 and Figure C.6, the fuzzy memberships derived from the differences provided nearly equal or superior results over that of the fuzzy memberships derived from the raw data, without the problem of determining the best λ.

C.4 Conclusions

This appendix has demonstrated two methods of fusing classifiers based on dynamic time warping distances. These results could easily be extended to other classifiers that output a distortion metric of similarity measure. The benefits of effective classifier fusion are clear in the problem investigated in this appendix. The two different classifiers with different approaches were combined effectively, often increasing the overall performance of the system. The results of this appendix demonstrate the superiority of the fuzzy memberships based on the difference between DTW distances to that of the fuzzy memberships based on the raw DTW distances.

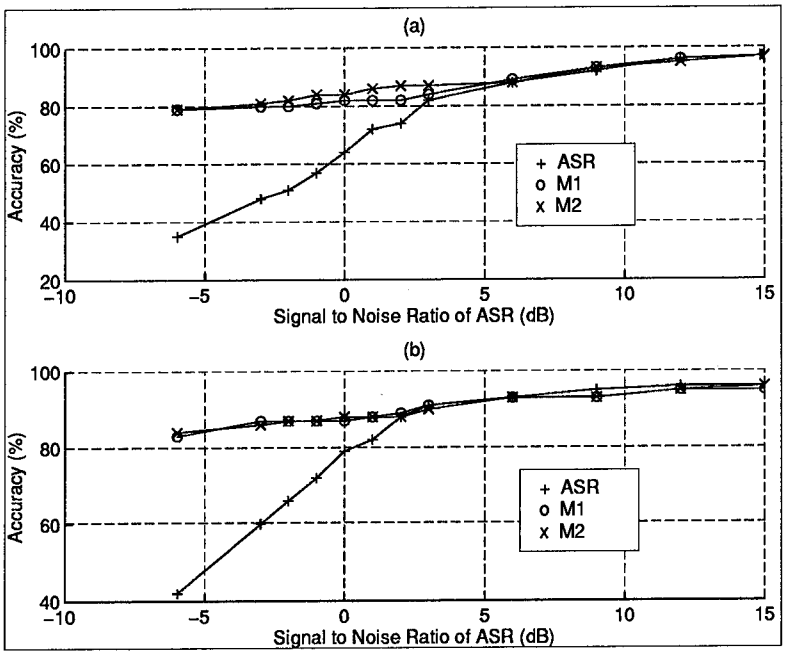


Figure C.6 Accuracy of fused systems vs signal to noise for the ASR. (a) 1 Template, ALR alone 82%. (b) 4 Templates, ALR alone 84%.

Appendix D. Combined System Results using Four Reference Templates

The results presented in this appendix are the combined results for tests on the combined system using four reference templates. The results are similar to those achieved using a single reference template. Figures D.1 and D.2 present the results of classifier fusion and a fixed λ of 0.97. The only significant difference with using four templates is that accuracies are improved for both acoustic automatic speech recognition and automatic lip reading. Tables D.1 and D.2 present the results for various values of λ. As was noted before, a variable $lambda$ may improve results. The feature fusion results are presented in Figures D.3 and refD.4. These results are also similar to the results using one template.

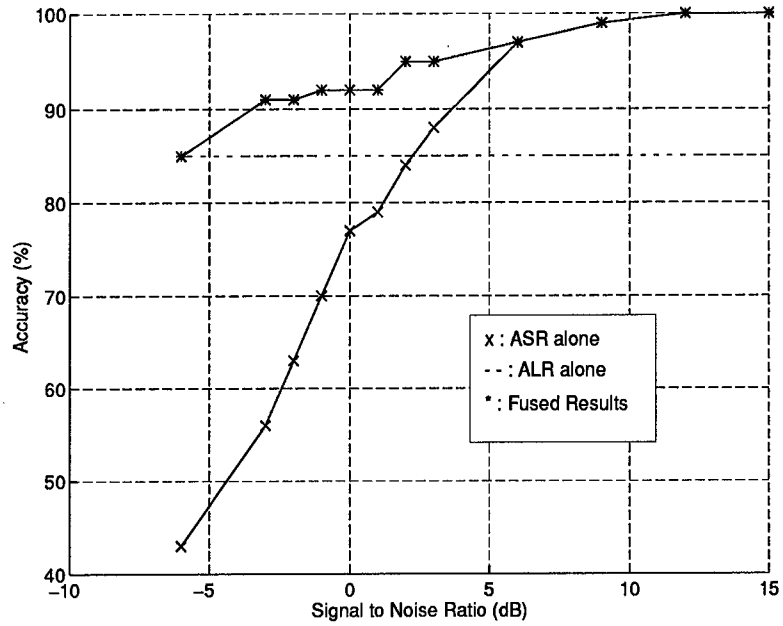


Figure D.1 Classifier fusion with the "Ultrasonic Mike" for $\lambda = 0.97$ and 4 templates. * : fused accuracy, $\times$: acoustic accuracy, and - - : automatic lip reader accuracy.

Table D.1 Classifier fusion with the "Ultrasonic Mike" with 4 templates, percentage accurate. Bold faced numbers indicate two highest results for a given SNR.

	Values of λ									
SNR (dB)	1.0	0.99	0.98	0.97	0.96	0.95	0.94	0.93	0.92	0.91
15	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
12	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**99**	**99**	98
9	**99**	**99**	**99**	**99**	**99**	**99**	**99**	97	95	94
6	**97**	**97**	**98**	**97**	96	95	94	92	91	91
3	88	93	**94**	**95**	93	92	91	91	91	91
2	84	92	**94**	**95**	93	92	91	91	91	91
1	79	91	**93**	**92**	**93**	**92**	91	91	91	91
0	77	90	**93**	**92**	**93**	**92**	91	91	91	91
-1	70	89	**91**	**92**	**91**	**91**	**91**	**91**	**91**	**91**
-2	63	87	**90**	**91**	**90**	**90**	**90**	**90**	**90**	88
-3	56	82	**90**	**91**	**89**	**90**	**90**	**90**	88	88
-6	43	70	83	85	**88**	**88**	**87**	**87**	**87**	**87**

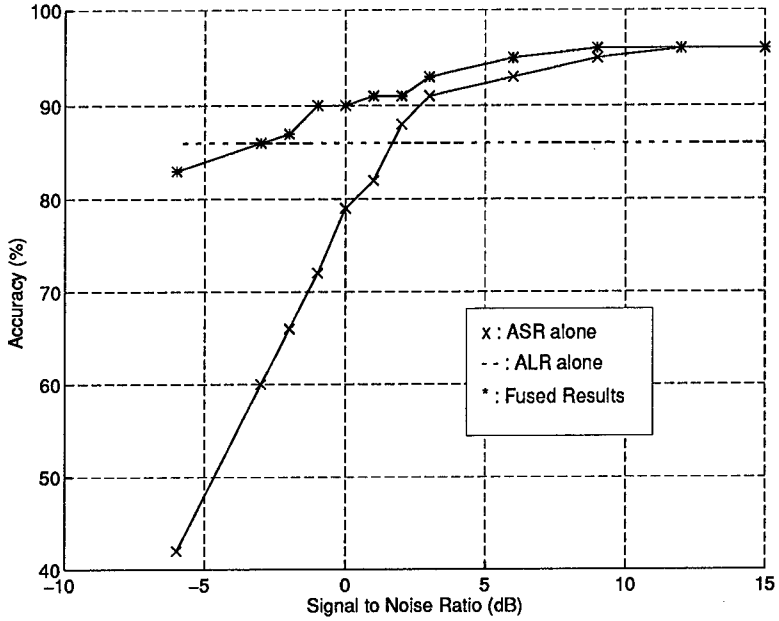


Figure D.2 Classifier fusion with the "Lip Lock Loop" for $\lambda = 0.97$ and 4 templates. * : fused accuracy, $\times$: acoustic accuracy, and - - : automatic lip reader accuracy.

Table D.2 Classifier fusion with the "Lip Lock Loop" with 4 template, percentage accurate. Bold faced numbers indicate two highest results for a given SNR.

	Values of λ									
SNR (dB)	1.0	0.99	0.98	0.97	0.96	0.95	0.90	0.85	0.80	0.70
15	**96**	**96**	**96**	**96**	**96**	**96**	**96**	**96**	**97**	94
12	**96**	**96**	**96**	**96**	**96**	**96**	**96**	**96**	**95**	94
9	**95**	**95**	**96**	**96**	**96**	**96**	**96**	**96**	**95**	91
6	93	**94**	**95**	**95**	**95**	**95**	**95**	**95**	91	88
3	91	**92**	**93**	**93**	**93**	**93**	**92**	90	91	87
2	88	89	**91**	**91**	**91**	**91**	**90**	**90**	88	87
1	82	87	**91**	**91**	**91**	**91**	88	**89**	87	86
0	79	87	89	**90**	**91**	**91**	88	88	86	86
-1	72	83	**89**	**90**	**90**	**90**	88	87	87	86
-2	66	81	87	87	**88**	**89**	**88**	87	86	86
-3	60	78	83	**86**	**87**	**87**	**87**	**87**	**86**	**86**
-6	42	62	79	83	**84**	**86**	83	**86**	**86**	**86**

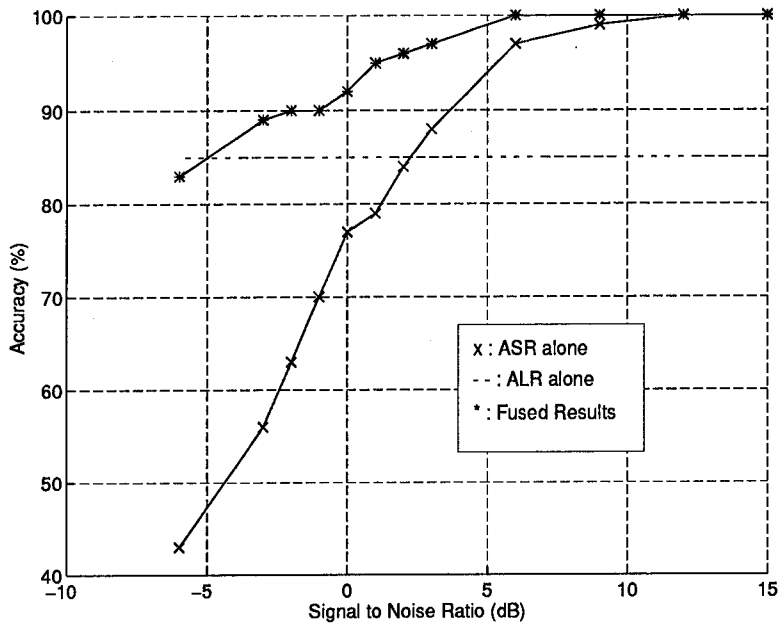


Figure D.3 Feature fusion with the "Ultrasonic Mike" for $\sigma = 0.20$ and 4 templates. * : fused accuracy, $\times$: acoustic accuracy, and - - : automatic lip reader accuracy.

Table D.3 Feature fusion with the "Ultrasonic Mike" with 4 templates. Bold faced numbers indicate two highest results for a given SNR.

	Values of σ								
SNR (dB)	0.00	0.05	0.10	0.15	0.20	0.25	0.50	0.75	1.00
15	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**99**
12	**100**	**100**	**100**	**100**	**100**	**100**	**99**	97	83
9	**99**	**100**	**99**	**100**	**100**	**99**	95	90	90
6	97	97	**100**	**100**	**100**	**99**	92	89	88
3	88	93	**97**	**98**	**97**	93	89	86	85
2	84	89	**96**	**98**	**96**	92	87	86	85
1	79	86	**95**	**97**	**95**	91	86	85	84
0	77	82	91	**96**	**92**	91	84	84	83
-1	70	82	**90**	**93**	**90**	**90**	82	83	83
-2	63	75	**90**	**92**	**90**	**90**	82	82	80
-3	56	71	87	**88**	**89**	87	82	82	83
-6	43	52	73	**84**	**83**	82	82	81	81

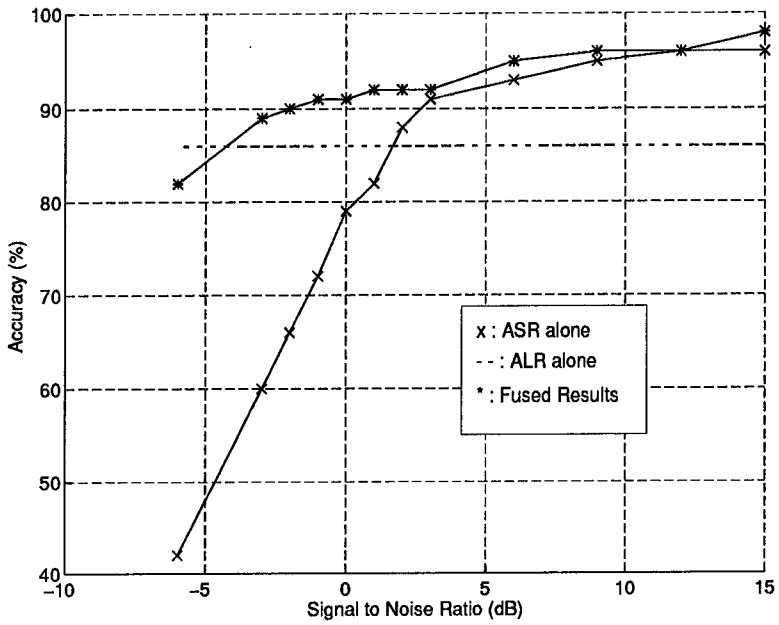


Figure D.4 Feature fusion with the "Ultrasonic Mike" for $\sigma = 0.20$ and 4 templates. * : fused accuracy, $\times$: acoustic accuracy, and - - : automatic lip reader accuracy.

Table D.4 Feature fusion with the "Lip Lock Loop" using 4 templates, percent accurate. Bold faced numbers indicate two highest results for a given SNR.

	Values of σ								
SNR (dB)	0.00	0.05	0.10	0.15	0.20	0.25	0.50	0.75	1.00
15	96	**97**	**97**	**97**	**98**	**98**	**98**	96	95
12	**96**	**96**	**96**	**96**	**96**	**98**	**98**	**96**	92
9	95	96	96	96	96	**97**	**98**	96	92
6	93	94	94	94	**95**	**96**	93	90	88
3	91	91	**93**	**93**	**92**	**93**	91	87	83
2	88	**92**	**92**	**92**	**92**	**93**	90	86	82
1	82	88	89	90	**92**	**93**	89	86	82
0	79	83	87	**89**	**91**	**89**	88	85	81
-1	72	83	87	**89**	**91**	**89**	87	83	81
-2	66	78	85	**89**	**90**	**89**	87	82	80
-3	60	75	83	86	**89**	**88**	86	81	79
-6	42	66	75	**81**	**82**	**82**	**82**	76	76

Bibliography

1. Atal, B. S. and Suzanne L. Hanauer. "Speech Analysis and Sythesis by Linear Prediction of the Speech Wave," *The Journal of the Acoustical Society of America, 50*:637–655 (April 1971).
2. Ballard, Dana H. and Christopher M. Brown. *Computer Vision.* Prentice-Hall, Inc., 1982.
3. Berger, Kenneth W. *Speechreading: Principles and Methods* (Second Edition). Herald Publishing House, 1978.
4. Bezdek, James C. and Sankar K. Pal. *Fuzzy Models for Pattern Recognition.* 345 East 47th Street, NY 10017-2394: IEEE Press, 1992.
5. claude Junqua, Jean and Yolande Anglade. "Acoustic and Perceptual Studies of Lombard Speech: Applications to Isolated-Words Automatic Speech Recognition," *IEEE, ICCASP, 2*:841–844 (1990).
6. Dodd, Barbara. "The Acquisition of Lipreading Skills by Normally Hearing Children." *Hearing by Eye: The Psychology of Lip-reading* edited by Barbara Dodd and Ruth Campbell, 163–175, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1987.
7. ESPS. Entropic Signal Processing System, Entropic Research Laboratory, Inc. 600 Pennsylvania Avenue, SE, Suite 202, Washington D.C. 20003.
8. Horn, Berthold Klaus Paul. *Robot vision.* MIT electrical engineering and computer science series, McGraw-Hill, 1986.
9. Jackson, Pamela L., et al. "Perceptual Dimemsions Underlying Vowel Lipreading Performance," *Journal of Speech and Hearing Research, 19*:796–811 (1976).
10. Kaplan, Harriet, et al. *Speechreading: A Way to Improve Understanding* (Second Edition). Gaullaudet University Press, 1985.
11. Kleppe, J. A. *Engineering Applications of Acoustics.* Artech House, Inc., 1989.
12. Makhoul, John. "Linear Prediction: A Tutorial Review," *Proceedings of the IEEE, 53*:561–580 (April 1975).
13. Marshall, Gerald F. *Laser Beam Scanning.* 270 Madison Avenue, New York, New York 10016: Marcel Dekker, Inc, 1985.
14. Marshall, Patrick. *Speech Recognition Using Visible and Infrared Detectors.* Master's thesis, Air Force Institute of Technology, Engineering Department, September 1992.
15. Mase, Kenji and Alex Pentland. *Automatic Lipreading by Optical-Flow Analysis.* Technical Report 117, Cambridge, MA: MIT Media Lab, Perceptual Computing Group, 1991.
16. Massaro, Dominic W. "Speech Perception by Ear and Eye." *Hearing by Eye: The Psychology of Lip-reading* edited by Barbara Dodd and Ruth Campbell, 53–83, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1987.

17. McGurk, Harray. "Hearing Lips and Seeing Voices," *Nature*, *264*:746–748 (December 1976).

18. Morton, Col. Paul and John Schnurer. Personal interviews. Dayton, OH, January - October 1994.

19. Oppenheim, Alan V. and Ronald W Schafer. *Discrete-Time Signal Processing*. Prentice Hall, Inc., 1989.

20. Papoulis, Athanasios. *Probability, Random Variable, and Stochastic Processes* (Third Edition). McGraw-Hill, Inc., 1991.

21. Parsons, Thomas. *Voice and Speech Processing*. Electrical Engineering, McGraw-Hill, 1987.

22. Petajan, Eric D. "Automatic Lipreading to Enhance Speech Recognition," *Proceedings IEEE Global Telecommunications Conference*, 265–272 (1984).

23. Petajan, Eric D. "An Improved Automatic Lipreading System to Enhance Speech Recognition," *Association for Computing Machinery Special Interest Group on Computer and Human Interaction*, 19–25 (1988).

24. Proport, Aerial. Aerial Corporation, Highland Park, N.J.

25. Rabiner, Lawrence. "Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition," *IEEE Transacations on Acoustics, Speech, and Signal Processing*, *26*:575–582 (December 1978).

26. Rabiner, Lawrence. "On Creating Reference Templates for Speaker Independent Recognition of Isolated Words," *IEEE Transacations on Acoustics, Speech, and Signal Processing*, *26*:34–42 (February 1978).

27. Rabiner, Lawrence and Biing-Hwang Juang. *Fundamentals of Speech Recognition.* Signal Processing Series, Englewood Cliffs, New Jersey: Prentice Hall, 1993.

28. Roe, David B. and Jay G. Wilpon. "Whither Speech Recognition: The Next 25 Years," *IEEE Communications Magazine*, 54–62 (November 1986).

29. Sanders, Mark S. and Ernest J. McCormick. *Human Factors in Engineering and Design* (Seventh Edition). McGraw-Hill, Inc., 1993.

30. Schalkoff, Robert. *Pattern Recognition.* John Wiley and Sons, Inc., 1992.

31. Sejnowski, Terrence J. and Moise Goldstein. *Massively Parallel Network Architectures for Automatic Recognition of Visual Speech Signals.* Final Technical Report DTIC AD-A226 968, Baltimore, MD: The Johns Hopkins University, 1990.

32. Sejnowski, T.J., et al. "Combining Visual and Acoustic Speech Signals with a Neural Network Improves Intelligibility." *Advances in Neural Information Processing Systems, 2* edited by D. Touretzky, 232–239, San Mateo, California: Morgan Kaufmann, 1990.

33. Silsbee, Peter L. and Alan C. Bovik. "Audio-Visual Speech Recognition for a vowel discrimination task," *International Society for Optical Engineering, SPIE, 2094*:84–95 (1993).

34. Stanton, Bill J., et al. "Robust Recognition of Loud and Lombard Speech in the Fighter Cockpit Environment," *IEEE, ICASSP*, *1*:675–678 (1989).

35. Stork, David G., et al. "Neural Network Lipreading System for Improved Speech Recognition," *IEEE International Joint Conference on Neural Networks*, 2:285–295 (1992).

36. Sumby, W. H. and I. Pollack. "Visual contribution to speech intelligibility in noise," *Journal of the Acoustical Society of America*, *26*:212–215 (1954).

37. Summerfield, Quentin. "Some Preliminaries to a Comprehensive Account of Audio-visual Speech." *Hearing by Eye: The Psychology of Lip-reading* edited by Barbara Dodd and Ruth Campbell, 3–51, Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1987.

38. Walden, Brian E., et al. "Benefit From Visual Cues in Auditory-Visual Speech Recognition by Middle-Aged and Elderly Persons," *Journal of Speech and Hearing Research*, 431–436 (April 1993).

39. Xu, Lei, et al. "Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition," *IEEE Transactions on Systems, Man, and Cybernetics*, *22*:418–435 (May 1992).

40. Yuhas, Ben P., et al. "Integration of Acoustic and Visual Speech Signals Using Neural Networks," *IEEE Communications Magazine*, 65–71 (1989).

Vita

Captain David L. Jennings was born on 6 June 1962 in DuQuoin, Illinois. He graduated from Murphysboro High School in June of 1980. In August of 1980 he entered the Air Force and attended Basic Military Training School at Lackland AFB, Texas. In the following months he attended technical training schools to become a Russian linguist. After completing this extensive training program he was assigned as a cryptologic linguist at RAF Chicksands, England. After a 3 year tour in England he returned to an assignment at the National Security Agency where he continued his service as a cryptologic linguist. In 1988 he was selected for the airman's educations commissioning program and began attending the University of Maryland at College Park, Maryland. In 1990 he graduated Magna Cum Laude with a Bachelor of Science degree in electrical engineering. After graduation David attended Officer Training School at Lackland AFB, Texas and was commissioned on 1 October 1990. His first assignment as an officer was at Falcon, AFB, Colorado Springs as a Satellite Engineering Officer. In May of 1993 he entered the School of Engineering, Air Force Institute of Technology at Wright-Patterson Air Force Base, Ohio, to pursue a Master of Science degree in Electrical Engineering. His primary emphasis lies in the fields of pattern recognition and communications. David is married to Colleen Jennings of Maple, Wisconsin and has six children Breanna, Larissa, Chantal, Bethany, Aaron, and Joshua.

Permanent address: 91 MEEHAN DRIVE
DAYTON,OH 45431

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	December 1994	Master's Thesis

4. TITLE AND SUBTITLE
MULTICLASSIFIER FUSION
OF AN ULTRASONIC LIP READER
IN AUTOMATIC SPEECH RECOGNITION

5. FUNDING NUMBERS

6. AUTHOR(S)
David L. Jennings

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology, WPAFB OH 45433-6583

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/GE/ENG/94D-16

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Lt Col Paul Morton
AL/CF
2610 7th St.
Wright Patterson AFB, OH 45433-7901

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)
This thesis investigates the use of two active ultrasonic devices in collecting lip information for performing and enhancing automatic speech recognition. The two devices explored are called the "Ultrasonic Mike" and the "Lip Lock Loop." The devices are tested in a speaker dependent isolated word recognition task with a vocabulary consisting of the spoken digits from zero to nine. Two automatic lip readers are designed and tested based on the output of the ultrasonic devices. The automatic lip readers use template matching and dynamic time warping to determine the best candidate for a given test utterance. The automatic lip readers alone achieve accuracies of 65-89%, depending on the number of reference templates used. Next the automatic lip reader is combined with a conventional automatic speech recognizer. Both classifier level fusion and feature level fusion are investigated. Feature fusion is based on combining the feature vectors prior to dynamic time warping. Classifier fusion is based on a pseudo probability mass function derived from the dynamic time warping distances. The combined systems are tested with various levels of acoustic noise added. In one typical test, at a signal to noise ratio of 0dB, the acoustic recognizer's accuracy alone was 78%, the automatic lip reader's accuracy was 69%, but the combined accuracy was 93%. This experiment demonstrates that a simple ultrasonic lip motion detector, that has an output data rate 12,500 times less than a typical video camera, can significantly improve the accuracy of automatic speech recognition in noise.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Classifier Fusion,Ultrasound,Lip Reading,Speech Recognition	98
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL