Augmented Speech Recognition

By

Hua Jian Guo

A thesis submitted to the
Faculty of Graduate Studies and Research Office
in partial fulfillment of the requirements for the degree of

Master of Applied Science in Electrical Engineering
Ottawa-Carleton Institute for Electrical and Computer Engineering
Department of Systems and Computer Engineering
Carleton University
Ottawa, Ontario, Canada

Your file Votre référence
ISBN: 978-0-494-26989-3
Our file Notre référence
ISBN: 978-0-494-26989-3

The undersigned is recommended to

the Faculty of Graduate Studies and Research

for acceptance of the thesis

Augmented Speech Recognition

Submitted by

Hua Jian Guo, B.Eng

in partial fulfillment of the requirements for the degree of

Master of Applied Science in Electrical Engineering

Adrian D.C. Chan, Thesis Supervisor

V. Aitken, Chair, Department of systems and Computer Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
Department of Systems and Computer Engineering
Carleton University

April, 2007

Author's Declaration

I hereby declare that I am the sole author of this thesis.

I authorize Carleton University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

Signature_____

I further authorize Carleton University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Signature_____

ABSTRACT

Two types of non-acoustic information sources, namely myoelectric signals (MES) and the general electromagnetic motion sensor (GEMS) signal, are investigated to overcome the limitations of conventional automatic speech recognition (ASR) systems; in particular, these limitations include degradation in noisy environments and reliance on the single acoustic signal modality.

A new training algorithm called the approximated maximum mutual information (AMMI) is demonstrated to improve the accuracy of MES ASR using hidden Markov models. Results show that AMMI training consistently reduces the error rates compared to conventional *maximum likelihood* training. Increases in accuracy of approximately 7% are observed at the empirically optimal operating point. A new ASR methodology using the GEMS signal is also presented. Classification accuracy of 68.9% is obtained for a ten-word vocabulary, confirming the presence of speech information in the GEMS signal.

Two types of multimodal ASR systems are developed combining MES, the GEMS signal, and the acoustic signal. Type I combined the output of multiple classifiers, each operating on a single signal modality. Type II combined the three modalities in a single classifier. Evaluation of the multimodal system under acoustic noisy conditions shows that performance of the multimodal systems was superior to the unimodal acoustic ASR system in noisy environment. An acoustic ASR system was demonstrated to have a classification error as high as 55.8% at an SNR of 15 dB, whereas the optimal multimodal ASR system classification error remained below 5.1% for the same range of noise.

ACKNOWLEDGMENTS

I would like to thank my thesis supervisor, Dr. Adrian D.C. Chan, for his technical insight, continual guidance, and encouragement, who had made this exploration most enjoyable.

I have been very fortunate having the opportunity to work with the graduate students Nabil Yazdani, Vesal Badee, Geoffrey Green, Fang Dai, Beverley Bradley and Mehran Talebinejad in the Biomedical Sensors and Signals Laboratory.

I would like to thank the volunteer subjects who participated in my experimental studies.

I want to thank the office administration staff and technical staff in the Department of Systems and Computer engineering for kindly providing me with support at various stages in my research.

I also want to thank the Natural Sciences and Engineering Research Council (NSERC) of Canada for the financial support for my research.

I also owe thanks to my family, my parents, who always show me with love, kindness and care.

Table of Contents

List of Figures

List of Tables

List of Acronyms

ABD	*anterior belly of the digastricus*
CER	classification error rate
ADC	analogue to digital converter
AMMI	approximated maximum mutual information
AR	autoregressive
ASR	automatic speech recognition
AWGN	additive white Gaussian noise
BPNN	back propagation neural network
CER	classification error rate
CPU	central processing unit
CTF	compact trace format
DAO	*depressor anguli oris*
dHMM	discrete hidden Markov model
DSP	digital signal processing
DTW	dynamic time warping
EM	expectation-maximization / electromagnetic
EMF	eclipse modeling framework
EMG	electromyography
GEMS	general electromagnetic motion sensor
HCM	hyper column model
HMM	hidden Markov model
HPF	high pass filter
IAV	integrated absolute value
LDA	linear discriminant analysis
LPF	low pass filter
LTW	linear time warping
MAP	maximum *a posteriori*
MCE	minimum classification error rate
MES	myoelectric signal
ML	maximum likelihood
MMI	maximum mutual information
MST	*masseter*
MU	motor unit
MUAP	motor unit action potential
PCA	principal component analysis
PLT	platysma
RF	radio-frequency
RMS	root mean square
SFAP	single fiber action potential
SNR	signal to noise ratio
SSC	slope sign change
STFT	short-time Fourier transformation

SVG	scalable vector graphics
SWT	standard widget toolkit
TDNN	time delayed neural networks
VAD	voice activity detection
WPT	wavelet packet transform
WT	wavelet transform
ZC	zero crossing
ZYG	*zygomaticus* major

Chapter 1 Introduction

1.1 Introduction

Conventional acoustic automatic speech recognition (ASR) systems can provide an acceptable performance in "clean" environments (i.e. low noise, low reverberation). However, the performance of such ASR systems, which rely solely on the acoustic signal, degrades dramatically when operating in noisy environments. For example, when an acoustic ASR system is trained and used under quiet ambient conditions, the error rate achieved can be less than 1%. The error rate, however, can increase to more than 50% when the same system is used in a cafeteria environment [1]. Such performance degradation is prohibitive for the general use of ASR.

Mismatches between the training and operating environments are the main reason which accounts for the deterioration in performance. As a result, many methods have been proposed to combat the drawbacks by reducing the mismatches between training and operating environment. These methods included filtering the noisy acoustic signal [2]-[3], giving more weigh to high SNR partitions of the acoustic signal in decision making [4]-[5], and exploiting *a priori* knowledge of both speech and noise signals [6]-[8]. While these methods have achieved success in certain applications of ASR, they also have shortcomings. For example, some signal filtering techniques such as Kalman filtering have been applied successfully to improve the SNR but not necessarily the intelligibility of the recognition accuracy. While removing the SNR, filtering also altered the spectrum of the speech signal [9]. In addition, despite improvements in acoustic ASR accuracy, methods

which operate on this unimodal approach are expected to saturate [10], and arguably this has already begun.

An alternative approach to improving the accuracy and robustness of the ASR system against noise is the use of complementary non-acoustic signals that contain speech information but are immune to noise, or at least react in a dissimilar manner to noise.

The presence of speech information in the myoelectric signal (MES) has been confirmed in many studies [10]-[21]. The MES pattern associated with speech can be recognized by a trained MES classifier, enabling MES to be used in a variety of ASR applications. Since the MES is unaffected by the acoustic noise [10], the output of the MES classifier could be used to compensated the shortcomings of the acoustic ASR system.

The general electromagnetic motion sensor (GEMS) signal is a signal that measures the vibration of human organs. The GEMS signal at the trachea area conveys the motion information of various tracheal tissues, including the tracheal wall, the vocal fold. When collected during speech, the GEMS signal is expected to include speech information. A study will be conducted to confirm this argument in this research. If the hypothesis is true, similar to the MES, the GEMS could be another useful alternative for ASR.

If the speech information within the MES, the GEMS, and the acoustic signal are not redundant, a multimodal ASR system combining the three information sources can be used to improve the classification accuracies and robustness under noise environment. The multimodal ASR system can be implemented by fusing the classification of individual classifier or by employing a classifier to classify all the information sources.

1.2 Objective

The objective of this thesis is to investigate non-acoustic speech signals that can be used to augment conventional acoustic ASR systems through a multimodal approach; in particular, the MES and GEMS signal will be examined. Specific objectives include:

1) Improved the classification accuracy of the MES for ASR. Hidden Markov models (HMM) have been demonstrated to be an effective method of classifying MES for ASR [14]. However, the *maximum likelihood* (ML) objective function used to train the HMM does not performing training to optimize classification accuracy. We proposed the use of the maximum mutual information (MMI) criterion for training HMM for MES ASR. This is demonstrated to have a consistent improvement in the MES ASR accuracy. Improvements in the MES ASR accuracy will in turn should enable an improvement in the performance a multimodal ASR system;
2) Performed a feasibility study to confirm the presence of speech information in the GEMS signal. This sensor modality was chosen as it is a contact based sensor, similar to the surface electrodes for MES. This sensor is also capable of extracting information of the vocal folds in the trachea. This modality is important as MES can only provide vocal fold information in a very limited manner; and
3) Developed a multimodal ASR using the acoustic signal, the MES, and the GEMS signal. It is anticipated that this multimodal approach to ASR will provide increased accuracy, as compared to the unimodal acoustic ASR system. This multimodal system will be evaluated under acoustically noisy conditions.

The rest of the thesis is organized as follows:

Chapter 2: A brief background review on ASR is provided. This includes non-acoustic ASR methods, such as the visual signal, the MES, and the GEMS signal.

Chapter 3: Various signals classification techniques, including linear discriminant analysis and HMM are discussed, with an emphasis on MES classification. An alternative HMM training algorithm termed the approximated maximum mutual information (AMMI) training is proposed to improve the MES classification accuracy. Performance of this algorithm was evaluated against the conventional *maximum likelihood* (ML) training.

Chapter 4: A new methodology of ASR using the GEMS signal at the trachea area is proposed. Features extraction and classifier structure for the GEMS signal classifier were studied. The feasibility of using the GEMS signal for ASR is confirmed.

Chapter 5: A multimodal ASR system combining the acoustic signal, the GEMS signal, and the MES was developed. Two types of combination methods were evaluated. The first method combines the outputs of three classifiers that operate on the acoustic signal, GEMS signal, and MES independently. Combination is accomplished in by a majority vote or score-combination scheme. In the second method the three types of signals are used as inputs to a single combining classifier for ASR.

Chapter 6: Conclusions and contributions of this thesis are summarized. Recommendations for future work are discussed.

Chapter 2 Background

2.1 Introduction

In this chapter, we will provide some background on augmented ASR. We will start by the definition of ASR and some of its applications. The limitations of conventional acoustic ASR are then discussed; these provide the major motivations for seeking of other non-acoustic speech signal to augment conventional acoustic ASR systems. Next we will review three non-acoustic signals which have been used for ASR; namely, the visual signal, the myoelectric signal (MES), and the general electromagnetic motion sensor (GEMS) signal.

2.2 What is Automatic Speech Recognition

Automatic speech recognition (ASR) is a machine intelligence technique that converts a speech signal to a sequence of words with an algorithm implemented on a computer system. ASR can be used as an alternative interface between humans and machines, where manual hand operation is impossible or undesirable. For example, instead of tediously typing text into a computer, a user can simply dictate verbally to the computer with ASR. Another example is that ASR can be applied to implement a hybrid model that integrates live call center agents with the advance speech processing approach for callers to ensure the correct computerized service is provided. Other applications of ASR include real-time transcripts and speech analytics. With the advancements in digital signal processing (DSP) techniques and increasing computational capabilities of computer systems in last several

decades, ASR has practically found its places in many computerized applications scenario; however, conventional acoustic based ASR suffers many limits and shortcomings, which are discussed below.

2.3 Challenges in Automatic Speech Recognition

There are three major challenges which need to be addressed when implementing an ASR system. First, the acoustic signal is vulnerable to ambient noise interferences. Although some ASR systems claim that they can achieve recognition rates above 98% if operated under "optimal conditions", these optimal conditions often mean a nearly noise free environment, or at least the operation under the same conditions that the system was trained upon. This may be impractical, if not impossible, in certain real-world applications. Second, a significant large number of training samples is typically required to achieve a high performance ASR system. This training can be tedious and time consuming for the user. User-independent systems seek to help overcome this issue; however, the performance of these systems is not as good as user-dependent systems. Third, there are certain situations where the acoustic based ASR is inapplicable. This includes situations where voice production is prohibited, such as communications in critical military mission or in a quiet library, or where voice production is not possible, such as persons with temporary or permanent speech impairments.

In this research, we investigate alternatives to the acoustic signal to extract speech information for ASR. These signals can be used as a secondary source of speech information to enhance conventional acoustic ASR systems. As well, they may also prove useful in standalone systems, such as voice prostheses. For these purposes, the non-acoustic signal

must possess the following attributes: it must contain speech information in a consistent manner, resist a variety of ambient interferences especially acoustic noises, and can be collected in a convenient way. In the next few sections, a number of these non-acoustic speech signals are discussed.

2.4 Visual Speech Signal for Automatic Speech Recognition

The visual speech signal refers to the movement of the lips, tongue, and other facial muscles of the speaker. To perform ASR effectively, even in a noisy ambient environment, the computer may need to rely on visual cues from the speaker. For example, the computer may analyze the speaker's face to extract speech information, such as lip movements, to assist ASR.

Generally, visual speech recognition can be divided into three main steps. First, a sequence of images around the mouth area is recorded simultaneously during the utterance. Second, features which characterize the lip movements are extracted from these images. Third, classification is performed on these features for speech recognition. Several studies have demonstrated the potential of this approach [22]- [27].

In [22], Petajan developed one of the first lip reading systems. A set of visual features, including the perimeter, the height, and the width of the lip, was derived from mouth images. The acoustic signal was first processed by an acoustic classifier to produce a few candidate words. The extracted visual features were then further analyzed by a visual classifier to make final decision among candidate words. By combined the visual and acoustic information, a recognition rate of around 80% was achieved in recognition the 26 English alphabet and the digits zero through nine. In [23], Goldschen extended the

system by using discrete hidden Markov model (dHMM). His work also concluded that time derivative features led to better performance. In other words, the lip movements provided more speech information than the lip positions. Using this scheme, he was able to achieve a recognition rate of 25.3% on 150 test sentences without using any syntactic, semantic acoustic or contextual information. Chiou *et al.* [24] proposed a contour finding of distinct visual features based on active contour model. Visual features were extracted from a sequence of mouth images and a HMM was then applied to process the visual features for word recognition. With the visual information alone, he was able to achieve a 93% recognition rate for an 11-word isolated vocabulary. The effectiveness of this model was confirmed by the works of Sugahara *et al.* [25], where a real-time lip reading system was developed on a commercial personal computer platform. The sampled active contour model was used to extracted lip shapes from series of face images, which was classified using a 4-state left-right HMM. A 93% recognition rate was obtained for a vocabulary consisting of the names of 10 Japanese railroad stations. Meier *et al.* [26] used time-delayed neural networks (TDNNs) for ASR. The acoustic features and the visual features were combined adaptively at several levels of the recognition network. The signal to noise ratio (SNR) was estimated and then was used to determine the weights for both the acoustic features and the visual features in the combination. This combination approach produces better recognition performance compared to the single acoustic recognition, especially in the case of high background noise. By using the additional visual information, the classification error rate was reduced up to 50%. Sergheer *et al.* [27] investigated the visual features representation for lip reading. He used a hyper column model (HCM) to extract visual speech features from the input images. An automatic lip reading was im-

plemented by feeding the HCM features a HMM classifier for recognition, which yielded around 81.5% classification rate in recognition of 9 Japanese sentences.

The visual speech recognition has demonstrated promise either as a standalone recognition approach or as a robust supplement for the traditional acoustic ASR. The major limit of this approach comes from the acquisition of mouth images. Since all of the visual features extractions are based on the images, the images have to be captured with brightness, contrast and luminance, which satisfy the minimal requirements for feature extraction. The image quality must be controlled or adjusted to provide consistency as well. In order to position the mouth accurately for feature extraction, lip markers are sometimes required, or a sophisticated face-tracking algorithm must be employed. In addition, in order to capture the motion of the lip, the camera must be set away from the user, which is not practical for some applications, such as portable units that go with the user.

2.5 Myoelectric Signals for Automatic Speech Recognition

2.5.1 Nature of the Myoelectric Signal

The myoelectric signal (MES), which is also known as the electromyography (EMG) signal, is an electrical signal associated with contraction of a muscle. The schematic representation of the MES generation process is shown in Figure 2-1. The α-motoneuron is an efferent neuron that originates in the spinal cord and synapses with muscle fibres associated with muscle contraction. The motor unit (MU) is the smallest group of muscle fibres that the central nervous system can control individually through a single α-motoneuron. Each terminal branch of a motoneuron innervates a muscle fibre in the MU. When a muscle contracts, the central nervous system controls muscle force by adjusting the number

of recruited α-motoneurons and the firing rate of each motoneuron [28].

When a muscle fibre is active, the firing of a motoneuron triggers an electrical depolarization which is called the single fibre action potential (SFAP). SFAPs originate at the neuromuscular junction where the terminal branch of a motoneuron innervates the muscle fibre, and propagates toward both ends of the muscle fibre in opposite directions. The spatial-temporal superposition of all the SFAPs from muscle fibres within a MU is called the motor unit action potential (MUAP) [29]. Accordingly, the time sequence of the MUAPs is called as the MUAP train. As the SFAP is an all-or-nothing phenomenon, meaning that the shape of the pulse is constant for each innervation, the MUAP is also an

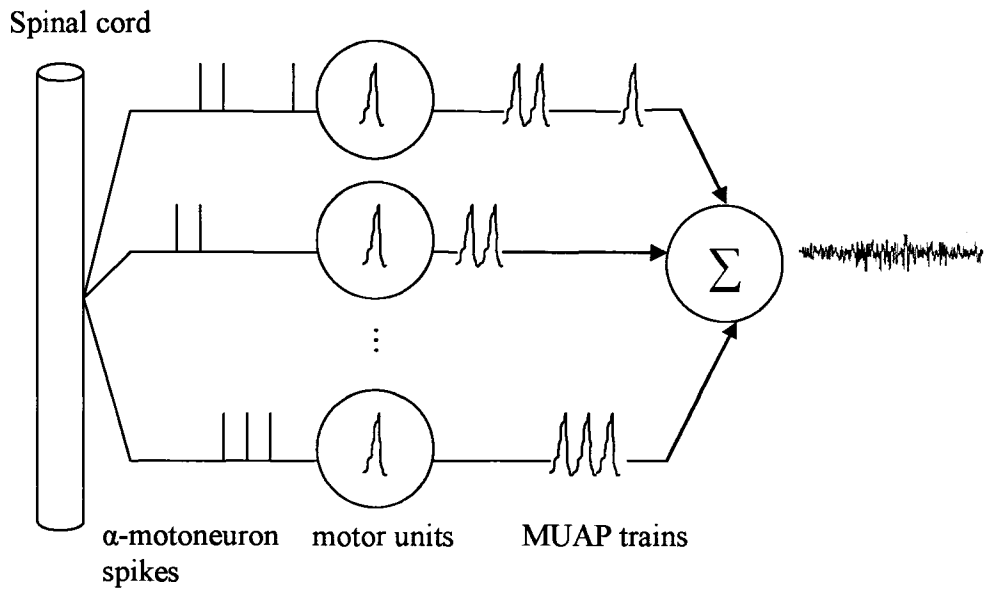


Figure 2-1 Schematic representation of the MES generation process based on [28]

all-or-nothing phenomenon; hence, the MUAP train can be modelled as a filtered train of impulse functions, where the filter function is a shape of the MUAP, as shown in Figure 2-1.

The MES can be non-invasively detected using electrodes on the surface of the skin. The measured MES is the summation of individual MUAP trains, innervated by active motor units, within the electrode pickup area. Because firing rates of α-motoneurons are irregular and asynchronous, the MES appears as a band-limited random signal. If the number of recruited MU is large, the MES can be reasonably represented by a Gaussian distribution function according to the central limit theorem. The amplitude of the surface recorded MES can range from 0 to 10 mV and the frequency can range from DC to 500 Hz [29].

2.5.2 Acquisition of the MES

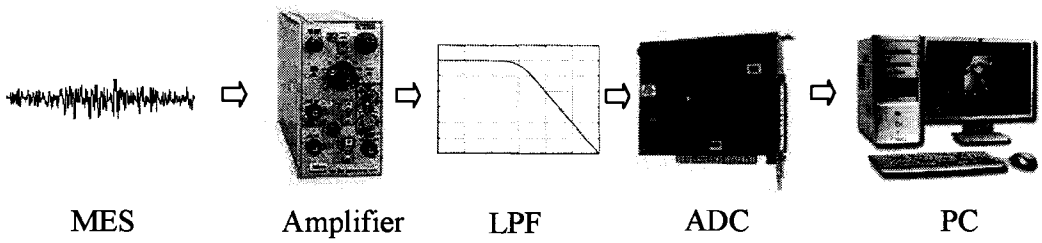


Figure 2-2 A typical MES acquisition system

A typical MES acquisition system consists of the following components: surface electrodes, amplifier, filters, and analog to digital converters, as shown in Figure 2-2.In a bipolar MES measurement configuration, a pair of electrodes is placed on the skin surface, above the muscle of interest. Since the amplitude of the signal is small (i.e. 0 ~ 10 mV), signal amplification is required to meet the input ranges of analog to digital converter.

The amplified signal is fed into a filter to remove various electrical noises, including 60 Hz power line interference. An analog to digital converter samples and digitalizes the MES, which is the used by the PC for further processing. In this thesis, this signal processing is performed off-line.

The MES is vulnerable to various electrical interferences, including:

1) electronic component noise which is inherent in the detection and recording equipment;
2) ambient noise, which is generated from sources of electromagnetic radiation, such as power line, radio transmission, and motor circuitry; and
3) motion artefacts, which are from movement of the electrode that change the electrode-skin interface.

Careful consideration should be taken when designing the acquisition system and during data acquisition. The two major objectives of the MES acquisition system are to maximize the signal to noise ratio, which is defined as the ratio of the energy of the MES to the energy of the noise signal and to minimize the signal distortion; that is, the relative contribution of each frequency component in the MES should be unaltered.

To accomplish the goals, special considerations are taken into account while designing the MES acquisition systems, as described below.

1) Differential amplification. The differential measurement configuration is based on the assumption that any signals that originate far away the detection sites will appear as common signals, whereas the signals that originate in the immediate vicinity of the detection sites will be different and consequently be

amplified. Thus, the relative distant noise signals will be subtracted while the relative local MES will be amplified.

2) Electrode-skin interface impedance. In order to prevent the detected signal attenuation and distortion due to the input loading, the skin-electrode interface impedance should be reduced as much as possible. One possible solution is to use nonpolarizable electrodes (e.g. Ag/AgCl electrodes) and apply conductive gels on the surface of the detection sites before measurement. Conversely, the input impedance of the differential amplifier should be as large as possible. In addition, the balance between the impedance of the detection sites is also of great importance in order to minimize the distortion of the MES.
3) Filtering. A band-pass filter, which consists of a high pass filter (HPF) to get rid of high frequency noise and a low pass filter (LPF) for baseline drift, usually applied after the stage of the differential amplifier to further increase the signal to noise ratio. Additionally, a 60 Hz notch filter is usually employed to eliminate the power line interference.
4) Electrode material. There are two types of common used electrodes for the MES measurement, both with their advantages and disadvantages. Conductive gel Ag/AgCl electrodes, which approach ideal non-polarizable electrodes, provide low noise due to a stable half cell potential, and have bandwidth that nears DC. On the other hand, polarizable electrodes such as stainless steel electrodes, which can be used without conductive gel, are more suitable in applications where the use of gel can dry out for longer term usage or can cause skin irritation. This is at the expense of introducing more noise.

2.5.3 Myoelectric Signals in Prosthetic Control

Since the 1940s, people have been researching upper arm prosthetic control using the MES. Many MES classification techniques, including MES ASR, were derived from such efforts.

The MES can be used as an indicator of muscle activity. For example, when the strength of muscle contraction is increased, more MU are recruited and the firing rate of active motor units increase, which increases the amplitude of the resultant MES. In [30], a myoelectrically controlled upper arm prosthesis was proposed using amplitude measures for parameterization of the MES from residual muscles; however, this simple parameterization could only differentiate three limb motions reliably: hand close, hand open, and rest. To implement a multifunctional prosthesis control system, either additional MES collection channels or more sophisticated MES classification techniques are required. In [31], a large number of electrodes were employed to collect MES at different control sites; however, due to the number of electrode sites required, the approach did not apply to amputees with severe nerve and muscle damage to their stumps. In [32], instead of using large number of electrode collection sites, one of two electrode sites were first used to collect the MES, then the recorded MES were parameterized in terms of the time series autoregressive (AR) model. This method utilized the statistical dynamics of the MES rather than its amplitude to recognize the correlations between the MES and limb functions. A real-time microprocessor system with a response time between 0.15 to 0.2 seconds was built to classify the parameterized MES. A recognition rate of 85% was achieved in discrimination 4 to 5 limb functions. Recently, a MES-based multi-fingered prosthesis was established for hand amputees in [33]. The MES was classified using the

combination of back propagation neural network (BPNN) and AR model. Six types of finger motions were identified in the proposed system with four electrode sites. A recognition rate of around 77% was achieved by the system.

2.5.4 Speech Information in Myoelectric Signals

The correlation between the MES and speech was observed by Morse *et al.* [11]. The MES around the neck was collected during speech using four electrodes. The average amplitude from each channel was then used as an input into a *maximum likelihood* (ML) classifier. The experimental results demonstrated that the recognition rate was five times higher than the *a priori* accuracy, although the recognition rate dropped significantly as the size of vocabulary increased. This decrease in recognition rate with increased vocabulary is of course not unexpected.

A separate study conducted by Sugie *et al.* [12] also confirmed the availability of speech-related information in the MES. MES were collected using three pairs of electrodes positioned around the mouth. An observation window with 40 ms length was used to segment each MES channel. Consecutive overlapping observation windows were used, with a window spacing of 10 ms. The number of times that each MES channel crossed a preset threshold was counted for each observation window. Then a state of either active or inactive was assigned to each observation window according to the number of crossings. The state sequences were then fed into a finite automaton for classification. This approach yielded a recognition rate of 64% for five Japanese vowels.

Chan *et al.* [13] also confirmed the presence of speech information in the MES. Furthermore, he observed that the onset part of the MES shows strong correlations with speech.

In his experiment, MES were collected during speech using 5 pairs of electrodes on facial muscles. The acoustic signal was recorded simultaneously and used to indicate the start of speech. Each MES channel was then windowed with variable pretrigger values into a 1024 ms length segment. The pretrigger was used as the MES precedes the acoustic speech. Four sets of features: simple time domain (TD) features, short-time Fourier transformation (STFT), the wavelet transform (WT) and the wavelet packet transform (WPT), were extracted. Principal component analysis (PCA) was used to reduce the dimensionality of these features sets. A linear discriminant analysis (LDA) classifier was used to classify the MES. The results were encouraging: a recognition rate of around 90% was achieved for a 10-word vocabulary, which demonstrated the potential of the MES as a secondary source of speech information. He also classified the MES using HMM, which was demonstrated being much robust to the time misalignment compared with the LDA classification [14]. It is shown that the MES is immune to acoustic noise and has potential to be implemented as supplement to the acoustic signal in a multimodal ASR system [10].

Bradley *et al.* [15] presented a MES based speech recognition which was intended to be used in acoustically harsh firefight environment. The Kingsbury's dual-tree complex wavelet transform was employed to extract features from a single channel MES. The features were processed with a conjugate gradient neural network classifier.

Jorgensen *et al.* [16] demonstrated a method for silent ASR using the MES. The MES was collected from the larynx and sublingual areas below the jaw using one pair of electrode. Then the complex dual quad tree wavelet transform was applied to the noise filtered MES. A trust region scaled conjugate gradient neural network was trained to recognize a six-word vocabulary.

Manabe *et al.* [17] proposed an idea of unvoiced speech recognition using the MES. Three pairs of electrodes were mounted on three fingers and held against three articulatory muscles on the face to collect the MES. The root mean square (RMS) of the MES was classified using a three-layer neural network. A recognition rate of 90% was achieved in recognizing five Japanese vowels. Manabe *et al.* [18] later proposed using the MES for voice activity detection (VAD) to determine the boundary between speech and silence. The root mean square (RMS) was calculated from each MES channel, then all the RMS values from different channels were multiplied and the result was finally compared with a threshold for VAD decision. In this paper, he confirmed that some MES reliably precede the voice, which can be a very useful property for predication. In addition, he showed that the MES is relatively insensitive to background noise.

Bu *et al.* [19] developed a myoelectric speech synthesiser for phoneme classification. Interestingly, the differential MES between monopolar electrodes on two different muscles, rather than the differential MES between bipolar electrodes on the same muscle, was used for feature extraction. Frequency domain features were extracted using a second order Butterworth filter bank. Then a probabilistic neural network, called log-linearized Gaussian mixture network was applied for recognitions.

Jou *et al.* [20] extended isolated MES ASR to continuous one. There considerations were taken into account: First, the MES was decomposed into a set of time domain feature space which keeps the useful speech information while reducing the noise. Second, three types of contextual filters, including the delta filter, the trend filter and the stacking filter were applied on the extracted features to better model the context. Third, the anticipatory effect was modeled by adding frame based delays to the MES. The features were classi-

fied with a LDA classifier. He then applied this approach on a 108-word vocabulary recognition, the results showed a noticeable reduction in the recognition error rate.

Maier-Hein *et al.* [21] investigated the effects of repositioning electrodes between recording sessions, environmental temperature variations and skin condition differences among different speakers, which have a significant impact on the performance of MES ASR. She proposed a signal normalization and system adaptation method which could be used to mitigate those impacts. In her paper, she also suggested that for MES ASR, applying more than two electrodes is crucial while the usage of more than five electrodes does not lead to significant performance improvement.

2.5.5 Applications of the MES-based ASR

With the rapid development in the digital signal processing (DSP) techniques, MES-based ASR has attracted huge interest in various communication scenarios; including communication in noisy ambient, implementation of MES controlled voice prosthesis and voiceless communications.

A MES-based ASR can be used to recognize speech signals with severe impairment in recognition rates within a changing ambient noise. In [10], a multi-expert ASR system was built using the acoustic signal and MES form five facial muscle sites, then performance of the multi-expert ASR system was tested under different levels of noise. Experimental results show that the MES expert is more resistant to the noise interference. The application for this research was ASR for pilots flying high performance jet aircraft. Another potential application of the MES-based ASR can be found in [15], where MES collected in an acoustically harsh environment (e.g. firefighters) has been used to extract the

speech information. Both examples above demonstrated the promising potentials of MES as an alternative or additional information source for the application of ASR.

A MES-based ASR system can also be used in the implementation of MES controlled voice prosthesis, which would be beneficial for people with permanent or temporary speech impairments. The advantage of such a system over other communication methods, which are often cumbersome or require a significant training time, a user could simply mouth the words as they would in normal speech, and have a computer interpret their MES and speak the words for them.

Another application of a MES-based ASR system is to realize voiceless communications in public place, where speech is not allowed (e.g. public library). The user utters words without acoustic output, and the related speech information is extracted from the collected MES. The advantages of this system are that the communication can be conducted without interference other people acoustically and keep the communication privacy [17].

2.6 The GEMS Signal for ASR

The general electromagnetic motion sensor (GEMS) is a homodyne radiofrequency interferometer that can be used to detect anatomical vibrations associated with the production of speech. The GEMS signal measured at the trachea can detect tracheal vibrations and provides information of vocal fold contact. Since the vibrations at the trachea and the contact of the vocal fold are highly related to the speech production. The GEMS signal is expected to contain speech information that can be very useful for speech processing.

Holzrichter *et al.* [34] argued that the GEMS signal provided positional or motional ar-

ticulator information, when calibrated and statistically validated, may have potentials to be used for ASR. In addition, the GEMS signal can also be used to determine the onset of speech or derive the voiced excitation function. Specially, the vocal tract transfer function can be estimated by first Fourier transforming both the acoustic signal and approximated excitation signal and then deconvolving.

Burnett *et al.* [35] proposed a speech denoising approach using the combination of the acoustic signal and the GEMS signal. Since the GEMS signal was relatively unaffected by acoustical noise, it could be used to determine the onset of speech under acoustic background noise. Once the non-speech periods were determined, the background noise spectral content could be estimated, and then a correlated filter could be built to maximally eliminate the background noise.

2.7 Summary

As we have discussed, the visual signal, associated with the lip motions, the MES, associated with the muscle contractions, and the GEMS signal, associated with the vocal fold vibrations consist of useful speech information. In addition, they exhibit some desirable characteristics in comparison with the acoustic signal (e.g. resistant to acoustic noise). Therefore, if the speech information could be extracted with satisfactory accuracies, these signals can be used as secondary information sources for ASR.

Due to the limitations of data acquisition interface for the visual signal (e.g. properly positioned camera or head tracking, consistent lighting conditions), only the MES and the GEMS signal were investigated in this thesis. In the following chapter, we will discuss the classification techniques that are used to extract the speech information for the MES.

Specially, we will propose a new algorithm termed the AMMI training with an aim to improve the classification accuracy.

A feasible study regarding the presence of speech information in the GEMS signal will be conducted in Chapter 5. While it had been suggested that the GEMS signal has potential for ASR, it had not been attempted before. The GEMS signal will be collected at the trachea area. Different feature extraction schemes and classifier structures will be examined and be used to classify the collected GEMS signal for ASR.

Chapter 3 MES Classification Techniques

3.1 Introduction

In the previous chapter, we have discussed the visual signal, the MES, and the GEMS signal as alternative information source for ASR. Previous studies have demonstrated the availability of speech information in the MES from articulatory muscles. The MES can be used as a supplement for conventional acoustic ASR to improve the overall performance. In this chapter we enhance the standalone MES ASR system, which in turn should enhance any combined MES-acoustic system. Various MES classification techniques have been proposed since the last few years. We begin examining some of these methods; specifically: linear discriminant analysis (LDA) and hidden Markov models (HMM). To improve the accuracies of the classification, we propose the use of the approximated maximum mutual information (AMMI) for HMM training.

3.2 Pattern Recognition

Pattern recognition can be defined as process of identification of raw data (patterns) based on either *a priori* knowledge or on statistical information extracted from the patterns. Pattern recognition generally consists of two main stages: feature extraction and classification.

1) Feature extraction. The goal of feature extraction has generally been to find a representation that is relatively stable for different examples of the same pattern, while simultaneously maximizing the ability to discern between different

categories of patterns. The features that are used typically depend on the character of the raw data. For example, in MES prosthetic control, Zardoshti *et al.* [36] suggested that the IAV of the MES is the most suitable feature for small window sizes, and the AR model of the MES provides greater class separability for large window sizes with the expense of more computational time.

2) Classification. Classification is a description scheme that does the actual job of classifying the raw data relying on the extracted features. Usually, it is based on the availability of a set of patterns that have already been classified. The two following classification approaches are common used: deterministic approaches and statistical approaches. Deterministic approaches are based on the structural interrelationships of features while statistical approaches are based on statistical characteristics of patterns.

3.3 Deterministic Approach – Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a commonly used technique for data classification and dimensionality reduction. The basic idea of LDA is to derive typical feature sequences for a muscle activity pattern through some averaging procedure and then use distance measures to compare patterns. LDA tries to maximize the class separability by maximizing the ratio of between-class scatter to the within-class scatter; the within-class scatter defines the scatter of samples around their respective class centers and the between-class scatter defines the scatter of the expected vectors around the global mean. An optimizing criterion can be found by a combination of with-class scatter and between-class scatter. Then a linear transformation matrix which is used to obtain the maximal class separability can be calculated by the eigenvector decomposition of the optimizing

criterion. One may refer to Duda and Hart [37] for a detailed description of this technique.

Due to its computational simplicity and training efficiency, LDA has been applied on a variety of applications including MES pattern recognitions; however, a LDA classifier requires a non-stationary input pattern to be aligned temporally to the reference pattern. When there is a temporal misalignment between the two patterns, the classification rates of LDA classifiers will decrease. Chan *et al.* [14] used LDA to classify the MES from five facial muscles of two subjects for a 10-word vocabulary ASR. The study has shown that the performance of LDA was sensitive to temporal misalignment. Within a 100 ms range of temporal misalignment, an increase in classification error of over 40% was observed.

Temporal alignment can be difficult when the two patterns to be compared are different in length. For example, in the case of the MES-based ASR, the speech rate may vary, which will cause inconsistencies among the length of MES. When a LDA classifier is applied for classification, the error rate will be high as a result.

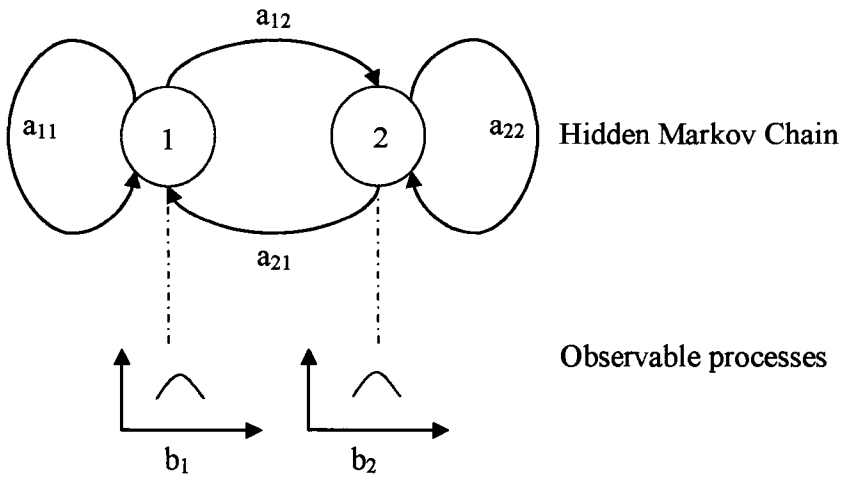
One possible solution is to use a form of "time warping" techniques, such that the input pattern and the reference are aligned temporally. This goal can be accomplished either linearly, which is referred as linear time warping (LTW) or nonlinearly, which is referred as dynamic time warping (DTW). It is noticed that speech rate variation causes nonlinear time fluctuation of MES lengths. Elimination of such fluctuation by time normalization is usually conducted before classification; however, a linear normalization (e.g. LTW) is inherently insufficient for coping with the highly nonlinear MES length fluctuations.

On the other hand, DTW, in which a nonlinear warping function is used to model the temporal fluctuation, can be applied effectively for the purpose of time normalization. The fundamental idea is that DTW finds the warping of the time dimension in one pattern such that a minimal distance between the two patterns is attained. Finding the best alignment between the two patterns is functionally equivalent to finding the best path through a grid mapping the features on one pattern to the features of the other pattern. DTW uses dynamic programming to find the path. The reader is referred to [38] for more details of the algorithm.

The second possible solution is to use another category of classification approaches, namely, statistical approaches as described in the following section.

3.4 Statistical Approach – Hidden Markov Models

Compared with deterministic approaches, statistical approaches are much less sensitive to the accuracy of the temporal alignment. In addition, with their inherent strong mathematical structure, statistical approaches provide a powerful framework for pattern recognition. In this section, we will focus our attention on one fundamental statistical approach in ASR, namely: the hidden Markov model (HMM).



a_{ij}: the probability of transition from state i to state j while being in state i

b_i: the probability density function associated with state i

Figure 3-1 A two-state HMM

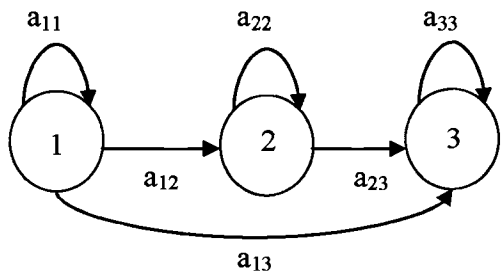


Figure 3-2 An example of 3-state left-right HMM

A HMM is a doubly embedded statistical process. An example of two-state (N = 2) HMM is shown in Figure 3-1. It consists of two components: the hidden Markov chain

and the observable statistical process associated with each hidden state. The hidden Markov chain consists of a set of distinct states and probabilities associate with interstate transition. The second component governs the output of each hidden state, which determines the observable sequences.

A HMM is fully characterized as $\lambda = \langle A, \Pi, B \rangle$, where A is defined as the transition matrix, Π as the initial state probability vector, B as an observation probability distribution associated with each hidden state. Mathematically,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} & a_{22} & \cdots & a_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ a_{N1} & a_{N2} & \cdots & a_{NN} \end{bmatrix} \quad \textbf{(3-1)}$$

$$\Pi = \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_N \end{bmatrix} \quad \textbf{(3-2)}$$

$$B = \langle b_1 b_2 \cdots b_N \rangle \quad \textbf{(3-3)}$$

The structure or topology of a HMM depends on the interstate connections. For example, a fully connected HMM is a model that every state of the HMM can be reached form every other state of the HMM in a single step. Specially, a type of HMM structure is called a left-right HMM as shown in Figure 3-2. The underlying state sequence associated with the left-right HMM has the property that the states only proceed from left to right (the transition matrix A is an upper triangular matrix). The left-right HMM has been found better than other structures (e.g. the fully connected HMM) in modeling acoustic

signals since the left-right HMM can readily model signals whose properties change over time in a particular sequence [39].

For acoustic ASR, in short time intervals, the acoustic signal can be viewed as a quasi-stationary stochastic process. Each state in a HMM represents the statistical characteristics of the quasi-stationary stochastic process, which may be associated with a specific phoneme. The HMM then statistically represents this sequence of phonemes that form a word.

3.5 HMM Training and Evaluation

The basic assumption of HMM for pattern recognition is that the pattern can be modeled as an underlying parametric random process of which the parameters can be learnt using different algorithms on the training patterns; that is, the pattern can be represented as having been generated according to some probability distributions.

There are two problems that must be solved for HMM to be useful in real-world pattern classification: evaluation and model training [39].

1) Evaluation. Given the observation sequence $O = o_1 o_2 \cdots o_T$, and a model $\lambda = \langle A, \Pi, B \rangle$, how to compute $P(O \mid \lambda)$, the probability of the observation sequence given the model?
2) Model Training. How to adjust the model parameters $\lambda = \langle A, \Pi, B \rangle$ to maximize $P(O \mid \lambda)$?

The solution for the first question is relatively straightforward: An algorithm called Viterbi algorithm can be used to efficiently solve this problem [39]. Next, we will focus

on the solutions to the second problem. In particular, *maximum likelihood* (ML) training and approximated maximum mutual information (AMMI) training will be presented in the following two sections.

3.6 Maximum Likelihood Training

The problem of HMM training can be defined to adjust the model parameters $\lambda = \langle A, \Pi, B \rangle$, such that the probability of the observation sequence is maximized; that is, $\lambda^* = \arg\max_{\lambda} P(O \mid \lambda)$.

Instead of finding the value of λ which maximizes $P(O \mid \lambda)$ directly, we define objective function $L(\lambda) = \log(P(O \mid \lambda))$, since $\log(\cdot)$ is a monotonically increasing function, maximizing $L(\lambda)$ is equivalent to maximizing $P(O \mid \lambda)$. We can transfer the product of probabilities to the summation of the logarithm of probabilities. This manipulation is of particular use when using multiple observation sequences for training, which is essential for accurate model training when the left-right HMM is used [39].

It has been shown that there is no analytical solution for finding the model which maximizes the likelihood of the observation sequence [39]; however, we can adjust λ such that $L(\lambda)$ is locally maximized using a recursive method called the Baum-Welch method. The problem of finding the optimal model λ is functionally equivalent to a standard constrained optimization problem. The objective function to be maximized is the logarithmic likelihood of the observation sequence given model λ, i.e., $\log(P(O \mid \lambda))$, the model parameters to be estimated are subject to certain stochastic constraints as explained in [39].

We define Baum's auxiliary function as,

$$F(\lambda,\overline{\lambda})=\sum_{Q}P(Q\mid O,\lambda)\log[P(O,Q\mid\overline{\lambda})] \qquad \textbf{(3-4)}$$

where $\overline{\lambda}$ denotes a re-estimated model and λ is the current model. We then find $\overline{\lambda}$ which maximizes $F(\lambda,\overline{\lambda})$, i.e. $\overline{\lambda}=\arg\max_{\overline{\lambda}}[F(\lambda,\overline{\lambda})]$. It has been proven that $\overline{\lambda}$ increases the likelihood of the observation sequence; that is,

$$P(O\mid\overline{\lambda})\geq P(O\mid\lambda) \qquad \textbf{(3-5)}$$

We then replace the current model λ with $\overline{\lambda}$ and repeat the above steps until a maximum number of iterations has been reached or the change in the likelihood in two sequential iterations is under a certain threshold. It also has been proven that the iterative procedure leads the objective function $L(\lambda)$ to converge to a critical point.

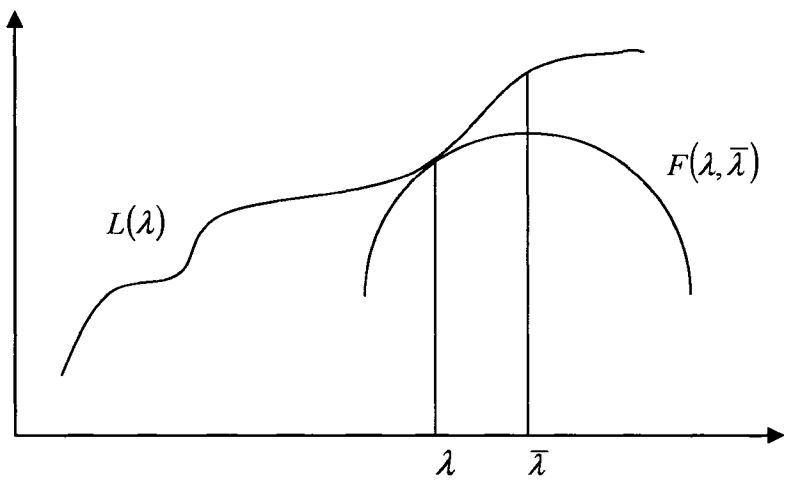


Figure 3-3 Graphical interpretation of a single iteration of the EM algorithm

The Baum-Welch method is a specialty of the general Expectation-Maximization (EM) algorithm. A graphical interpretation of EM algorithm is shown in Figure 3-3. The Baum's auxiliary function $F(\lambda,\overline{\lambda})$ is upper-bounded by the objective function $L(\lambda)$ and

the value of the two functions is equal at λ. The EM algorithm chooses $\overline{\lambda}$ which maximizes $F(\lambda, \overline{\lambda})$. Since $L(\lambda) \geq F(\lambda, \overline{\lambda})$, increasing $F(\lambda, \overline{\lambda})$ ensures the value of the objective function $L(\lambda)$ is increased at each iteration.

Assuming the observation probability function associated with each HMM state is a Gaussian mixture, the re-estimation formulae of $\lambda = \langle A, \Pi, B \rangle$ can be derived from this optimization procedure [39].

Despite its prevailing for real-world pattern recognition, ML training also contains certain limitations. First, the ML training is trying to maximize the likelihood of observation sequences; however, this optimal criterion, the maximum likelihood of observing the training samples given the model, does not associate directly with the minimum classification error rate (MCE), which is a critical performance indicator of a classifier.

Second, the ML training is optimal when the stochastic assumptions on the signal is true, for example, the observation sequences associated with subsequent states are independent with each other and the probability distribution function of each state is a true description of the signal probabilistic characteristic; however, these stochastic assumptions are merely a mathematic approximation of the real MES.

Third, theoretically, the ML training requires an infinite number of training samples to give an optimal estimation on the signal model. When the number of training samples is insignificant, the ML training will introduce bias to the estimated parameters. In practice, the availability of training samples is usually limited.

3.7 Approximated Maximum Mutual Information Training

To overcome the inherent limitations of ML training, a new category of training named the maximum mutual information (MMI) training, based on the concept of discriminative training, may become more appropriate for certain pattern recognition.

The MMI training adjusts parameters of HMM so that the mutual information between observations and correct patterns is maximized, thus producing a model which has an increased capability of distinguishing observations generated by different patterns. MMI training has to estimate parameters jointly across the entire classes, and the gradient descent searching algorithm is sensitive to step sizes: small step sizes take long time for the training to converge, while large step sizes may lead to instability. The AMMI training optimizes the objective function called the approximated MMI criterion. Unlike the MMI training, the parameters for each HMM can be calculated separately in a method similar to the Baum-Welch algorithm.

Let $\lambda = \langle A, \Pi, B \rangle$ be the parameter of the HMM for the pattern specified by γ,

$$J(\lambda) = \sum_{O \in u} \log P(O \mid \lambda) - d \sum_{O \in v} \log P(O \mid \lambda) \qquad \textbf{(3-6)}$$

where u is the set of indices of the training data that were labelled as the pattern specified by γ, v is the set of indices of the training data that were recognized as the pattern specified by γ after the *maximum a posteriori* (MAP) criterion is applied on the training data, and d is a parameter that can be used for adjusting the discrimination rate. The optimization for the objective function $J(\lambda)$ can be implemented in a manner similar to the Baum-Welch algorithm and the re-estimation formulae are given in [40].

The main difference between ML training and AMMI training lies in the objective function to be optimized. ML training maximizes the likelihood of observation sequences given the model. On the other hand, AMMI training maximizes the likelihood of model given the observation sequences which minimizes the CER.

3.8 AMMI Training for MES Pattern Recognition

3.8.1 Data Collection

The MES database used in the section was from the previous study by Chan [10]. The MESs were collected from five articulatory muscles around the mouth: *zygomaticus major* (ZYG), *platysma* (PLT), *depressor anguli oris* (DAO), *anterior belly of the digastricus* (ABD), and *messeter* (MST). These muscles control the movement of the mouth and the lip. Their respective functionalities are as follow:

1) ZYG: draws upper lip upward and outward
2) PLT: controls the movement of the lower lip
3) DAO: draws corners of mouth downward and laterally
4) ABD: elevates and retracts hyoid bone and tongue
5) MST: prime mover of jaw closer

Five pairs of Ag-AgCl Duotrode MES electrodes (Myotronics, 6140) were placed on the five articulatory muscles for the MES collection [10]. One Ag-AgCl Red-Dot electrode (3M, 2259) was placed on the back of the neck to provide a common ground reference.

A systemic depict of the MES acquisition system is shown in Figure 3-4. The MES was pre-amplified by an array of optically isolated differential amplifiers. The outputs of the pre-amplifiers array were the inputs of another array of amplifiers (Tektronics, AM502),

whose gains were set to values between 10 and 20, depending on the amplitude of the MES to be amplified. The gain was determined by attempting to maximum the dynamic range of the analog to digital converter, without over-ranging. The amplified MES were then anti-aliased by a low pass filter with a cut-off frequency of 500 Hz. During the data collection, the acoustic signal was recorded simultaneously using a microphone. The output of the microphone was amplified by a gain of 5 and bandlimited to 5000 Hz. The acoustic signal was used later to segment the data by using it to determine the start of speech. Both MES and acoustic signal were sampled at 10 kHz using a data acquisition board with a sample-and-hold unit (CIO SSH16) and a 12-bit analog to digital converter (ADC) (CIO DAS16/330).

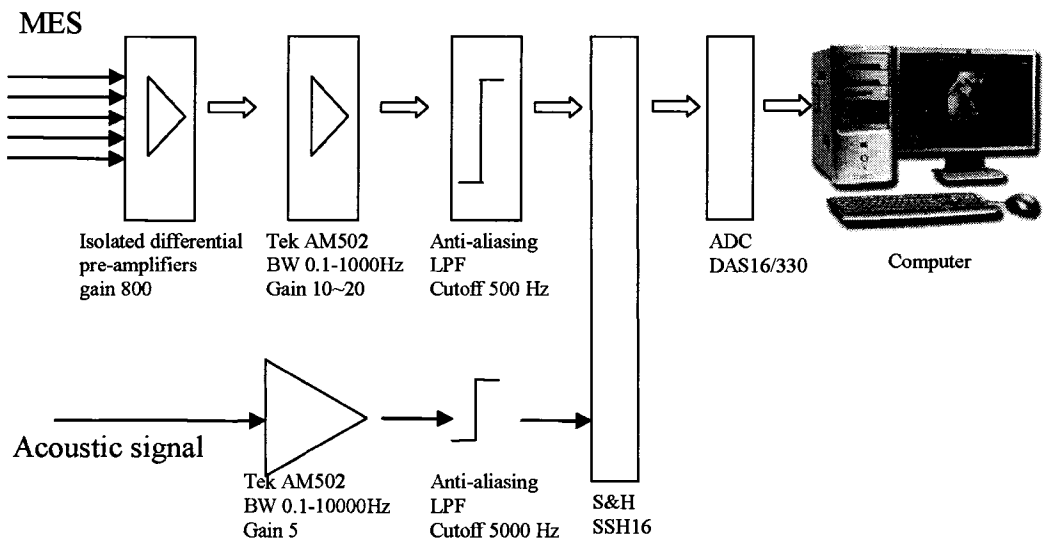


Figure 3-4 Systemic depict of the MES acquisition system

A 10-word vocabulary consists from the words form 'zero' to 'nine' was uttered during the data collection. Five subjects were participated in the data collection. Five channels of

MES from facial articulatory muscles (i.e., ZYG, MST, ABD, PLT and DAO) were sampled when each subject was uttering the words. Each word in the vocabulary was repeated 52 times by each subject. The words were uttered in a random order. In addition, there was a one-second pause between each repetition to reduce any coarticulatory and anticipatory effects.

In the original experiment, MES were collected with the acoustic signal contaminated by six different levels of white Gaussian noise ranges from 0 dB to 18 dB. In order to get a sufficient number of training and test samples, we combined the entire MES from the six noise levels. This combination was justified by the assumption that the MES is unaffected by the ambient noise, which is demonstrated in [10].

3.8.2 Data Processing

Recorded data were processed offline using Matlab® Version 7. The original MES signals were first downsampled to 1000 Hz, then segmented into frames of 1024 ms length using the simultaneously recorded acoustic signal as a trigger signal, as depicted in Figure 3-5. Since the muscular contractions would be expected prior to the acoustic speech, the start of MES activity would occur before the acoustic signal indicates speech. Thus MES data was segmented using a pretrigger. A pretrigger value for the MES was fixed at 500 ms, which has been shown to be empirically optimal range [10].

A set of sliding windows of distinct window size and step was used to frame the segmented MES, as shown in Figure 3-6. Within each window, the signal features were computed; namely, the root mean square (RMS) value and autoregressive (AR) coefficients. The values from each window were combined to form a feature vector. Depending

on how many AR coefficients were selected, the dimensionality of the feature vector would change. For example, if the first AR coefficient was used, the feature vector would have a dimensionality of 3. The number of feature vectors will vary with the window size and step. For example, when a length of 1024 ms signal was feature extracted using a 128 ms length window with a step of 16 ms, the number of feature vectors will be 57.

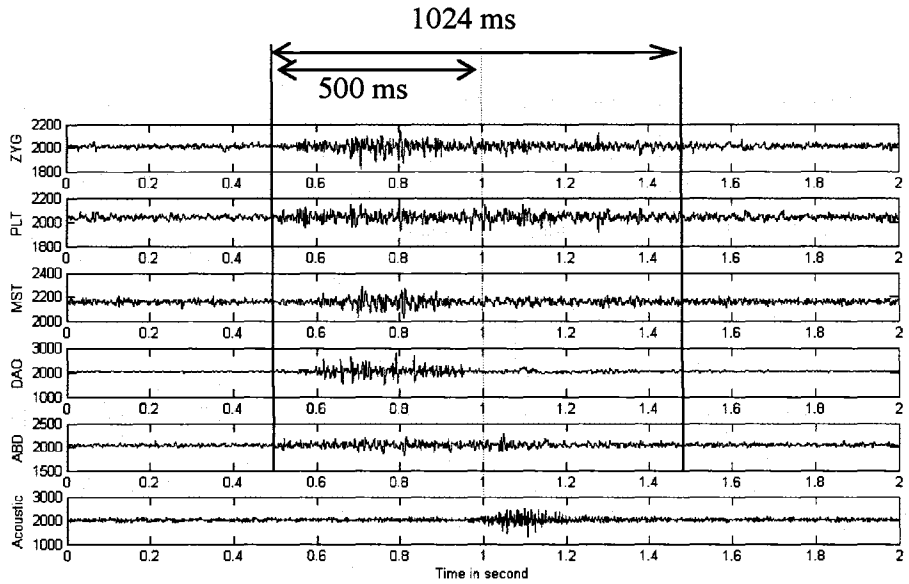


Figure 3-5 Segment of the MES

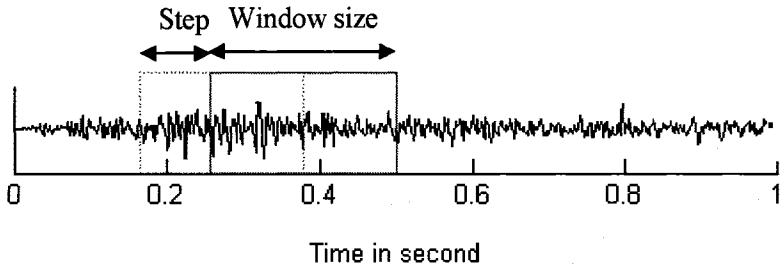


Figure 3-6 Sliding window for features extract

3.8.3 MES HMM Classifier

The construction of a MES HMM classifier involves two phrases: classifier training and MES pattern recognition.

The MES for each word was separated into two separate sets. The odd number samples were used as training samples and the even number as test samples. A left-right HMM with single mixture observation Gaussian densities was built for each word. The HMMs were trained using the training samples.

After the HMM parameters were estimated, the test samples were classified by the HMMs using the Viterbi algorithm. The classification error rate (CER), which was defined as the ratio between the number of wrongly classified test samples and the total number of test samples, was used to assess the performance of the system.

Both the ML training and the AMMI training were used for the HMMs training, The CERs were compared with various model parameters settings: window size, number of HMM states, number of AR coefficients, and the number of training samples.

3.9 Results

3.9.1 Number of Features

In this section, we compared performance of the AMMI training to that of the ML training varying the number of features. A 10-state HMM was used with 128 ms of length observation window size, with a step of 16 ms. The RMS and a various number of the AR coefficients were used as features, with the AR order ranging from 0 to 12.

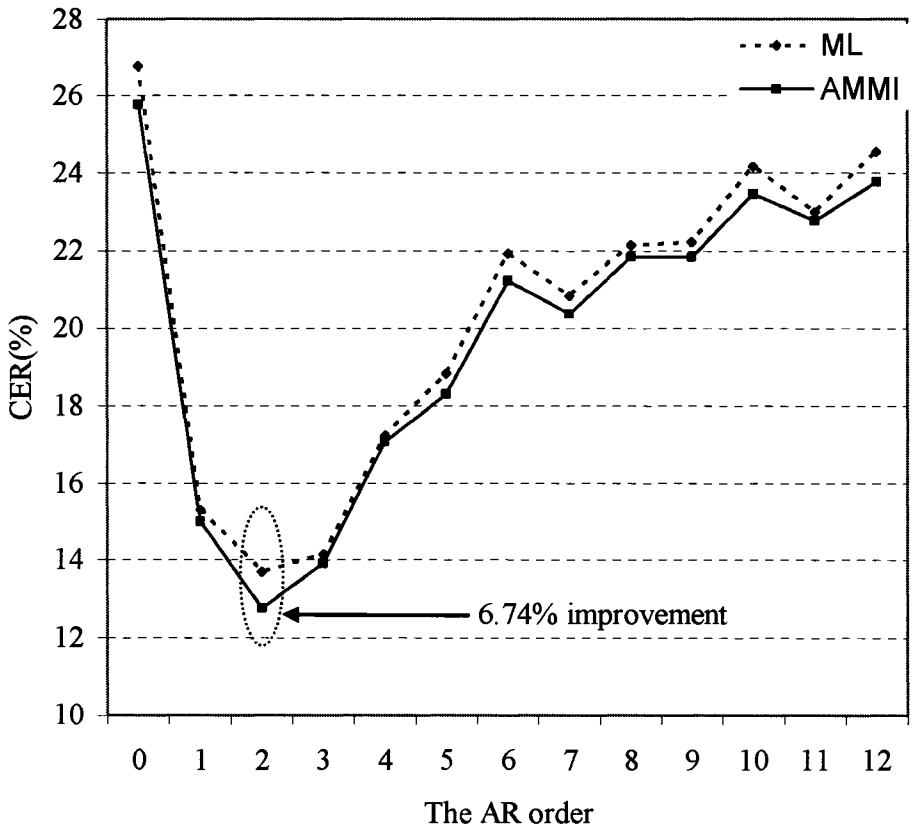


Figure 3-7 CER as a function of the AR order

As shown in Figure 3-7, the AMMI training outperforms the ML training consistently in terms of CER. When only the first AR coefficient was used as a part of feature vectors, the CER obtained by the AMMI training was 15%, when the first two AR coefficients were applied, the CER rate dropped significantly to 12.77%. Using more than 2 AR coefficients, however, does not further reduce the CER. Using small number of AR coefficients does not provide enough discriminative information. On the other hand, introducing more AR coefficients increased the dimension of the signal features, which would require more training data to fully utilize the additional AR coefficients. In this case, with a fixed training data set, the increase dimensionality is likely resulting in an undertrained

system causing the decreased performance. The CER from the ML training was 15.31% and 13.69% for the first and first two AR coefficient cases, respectively. Examining the empirically optimal case where the first two AR coefficients are used, the AMMI training provides a 6.74% improvement over the ML training algorithm.

3.9.2 Observation Window Size

Different size of observation window and step were applied to extract features from the MES data. The window sizes used were: 16 ms, 32 ms, 64 ms, 128 ms, and 256 ms. Observation windows were spaced one eighth of the window size and 10-state HMMs were used. The RMS and first two AR coefficients were used as features.

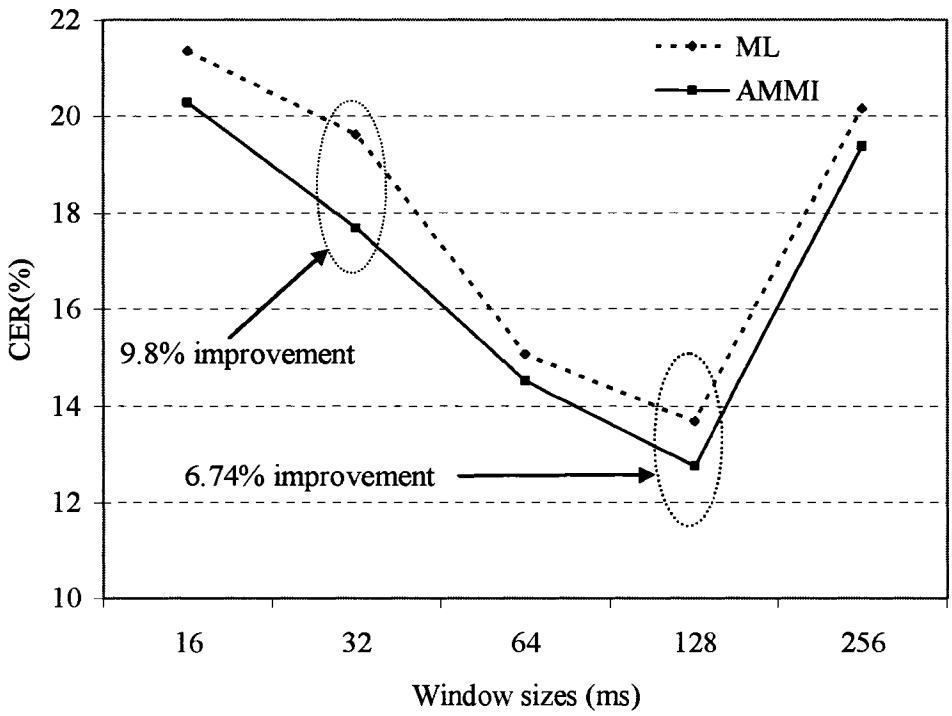


Figure 3-8 CER as a function of window sizes

The error rates with different observation window sizes for both the AMMI training and

the ML training are shown in Figure 3-8. The error rates decreased significantly with increasing window size until the optimal window size of 128 ms was reached, after which the CER increased. As shown in the plot, the HMM model by the AMMI training had a consistently lower error rate than the ML training. With 128 ms observation window, the CER was 12.77% and 13.69% for the AMMI and ML training, respectively. There is an improvement of 9.8% when a window size of 32 ms was used. The improvement for the optimal window size was 6.47%. An increase on the CER on small and large observation window size was observed. A small observation window would lead to high variance and introduce noise in the observed features. On the other hand, too large window size would smooth the data and would be unable to capture the short time change of the MES. Thus, it was reasonable that the empirically optimal observation window size was at a point between the two extremes.

3.9.3 Number of States

In this experiment, a left-right HMM with the number of states varying from 3 to 15 was used. The observation window size was fixed at 128 ms, with 16 ms window spacing, and the RMS and first two AR coefficients, which were found to yield the lowest CER in the previous two experiments, were used as features. The CERs with the changing state sizes are plotted in Figure 3-9. The CER varied from 31.69% to 13.69% for the ML training, and 28.53% to 12.77% for the AMMI training. The CER increased when using a HMM with too few or too many numbers of states. The optimal state size was 10 for both the ML and AMMI training. The largest improvement of around 13% occurred when a 5-state HMM was used.

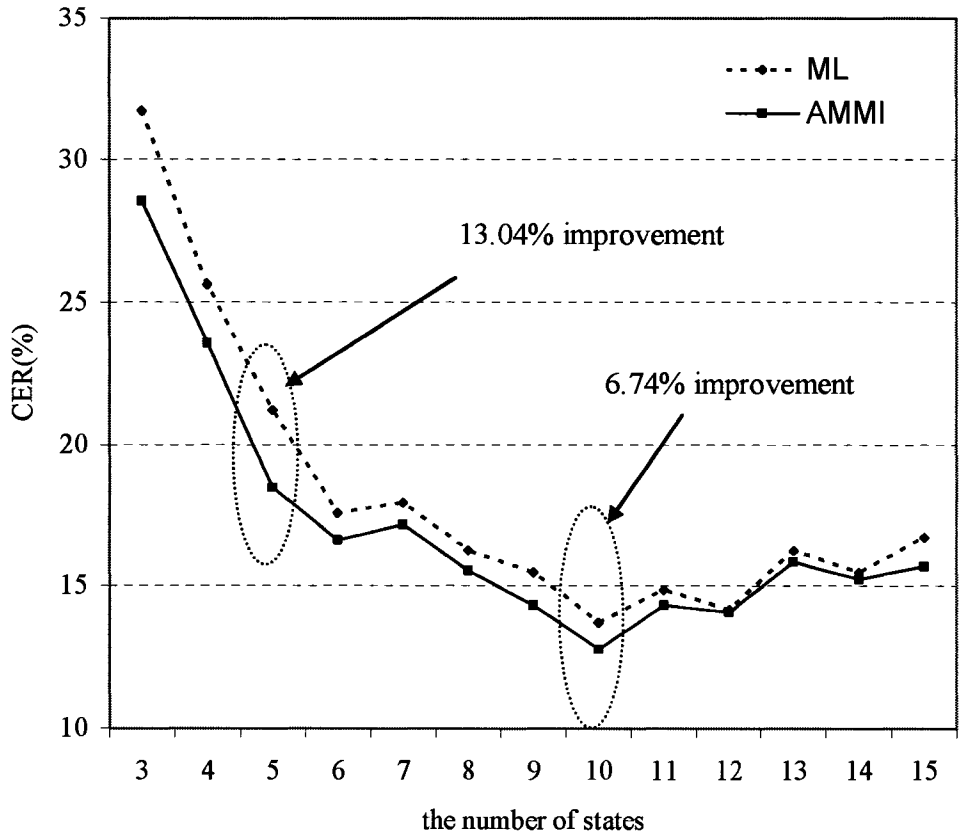


Figure 3-9 CER as a function of the number of states

3.9.4 Training Size

Using the empirically optimal HMM and feature settings found from the previous sections, the effect of the number of training samples was evaluated. The number of training samples was decreased from 260 to 100 by steps of 10. The CER for the AMMI training

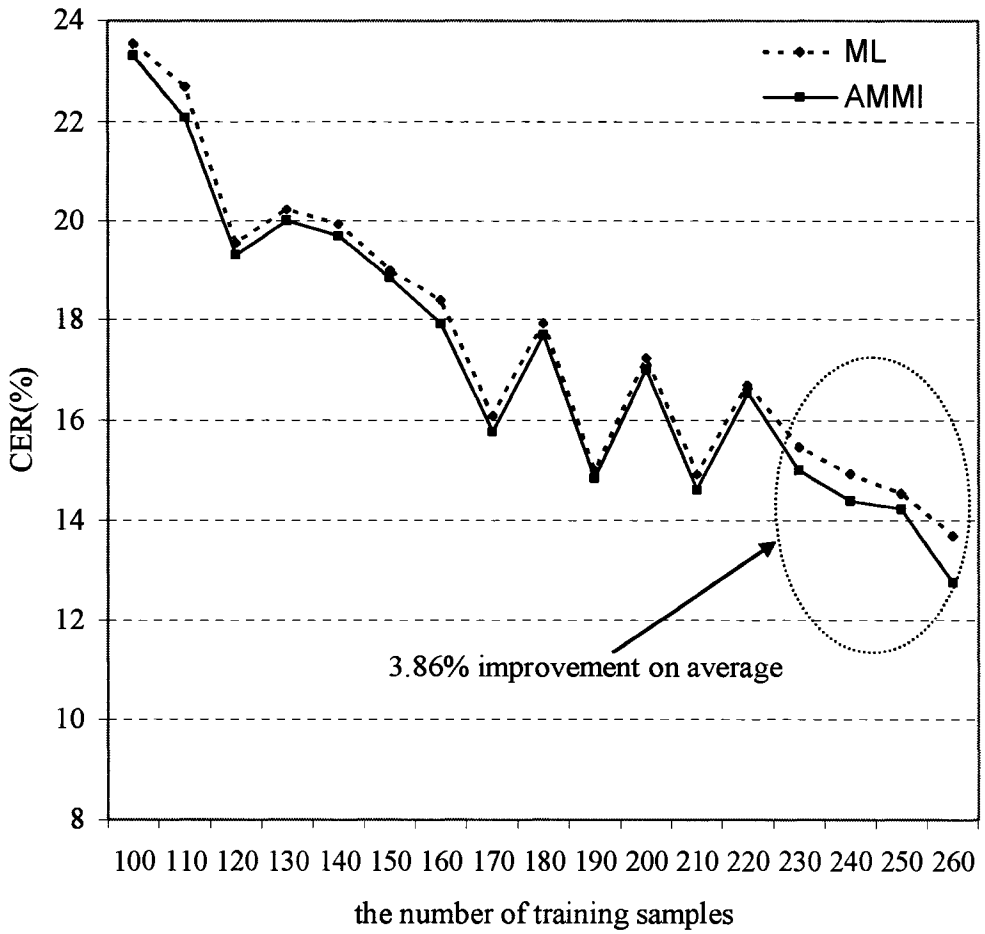


Figure 3-10 CER as a function of the number of training samples

and the ML training as a function of the number of training samples is shown in Figure 3-10. The CER trended to increase with the reduction of the number of training samples, as expected. The CER for the AMMI training was 12.77% (13.69% for the ML training)

when the number of training samples was 260, and rises to 23.31% (23.54% for the ML training) when the number of training samples was reduced to 100.

3.10 Discussion

It has been shown that the AMMI training can be used as an alternative in the estimation of HMM parameters to reduce CER for MES ASR. The experimental results demonstrated that the AMMI training obtained a HMM model with a reduction in recognition error rates when compared to that with the ML training. The average improvement was approximately 3% on the test database. At the empirically optimal operating point, the AMMI training provides a 6.74% improvement over the ML training. The increase in classification accuracy may appear modest; however, one must acknowledge that the improvement is from an already high classification accuracy (80% ~ 85%). In addition, this increase in classification accuracy was consistent when changing the window size, number of states, number of features, and training size.

ML training seeks to optimize the HMM parameters that maximizes the likelihood of observation, whereas the AMMI training aims to minimize the classification error rate. The parameters of HMM were adjusted so that the mutual information between observations and correct classes is maximized, thus producing a model which has more capability in distinguishing observations generated by different classes.

The trade-off for an increase in the classification accuracy by the AMMI training was an increase in the computational complexity. This increase in complexity is only in the training phase. Once the classifier is trained, there is no impact on the testing phase.

An empirically optimal operating point for the MES classification in the context of ASR was proposed. When the vocabulary changes, the empirically optimal operating point may change; the change is not expected to be large, so that the results and the search scheme can serve as a good guideline for future research.

It was noted that the usage of more AR coefficients does not achieve smaller CER. This result can be explained by the fact that the introduction of more features increases the number of free parameters in the HMM and hence the dimensionality of the problem. Additional training data would be required to properly estimate the HMM parameters. Unless the additional features are providing sufficient additional discerning information, they can increase the error rate.

As expected, the CER decreased as the number of training samples increased for both ML training and AMMI training. However, when a larger number of training samples was used, the AMMI training provided more improvement over the ML training, with a gain of 4% on average, which may imply the AMMI could be more desirable when large number of training samples was used.

Chapter 4 GEMS Signals for Automatic Speech Recognition

4.1 Introduction

In this chapter, we propose a new modality for ASR using the general electromagnetic motion sensor (GEMS) signal. We will provide a brief description of the GEMS first. Then describe a measurement setup to collect the GEMS signal at the trachea area. We proceed to the analysis of the GEMS signal for ASR in the following section. The process of classification first decomposes the GEMS signal into a feature space. We attempt to find a feature space which produces a high recognition rate, examining a number of different features and their combinations. The structure of the HMM which has the highest performance is also investigated. This chapter will demonstrate that the GEMS signal contains useful speech information. If this information can be extracted appropriately, the GEMS can be used as alternative information source or as another supplement for ASR.

4.2 GEMS Signals

Radarlike sensors have been used to measure the organ motion of the human vocal system during acoustic speech by transmitting electromagnetic (EM) waves [41]. A pulse of very low power and high frequency EM waves train is transmitted into the neck or the jaw where articulatory movements associated with speech are monitored. The GEMS is a type of EM sensor which is optimized for glottal electromagnetic sensing [42] [43].

The GEMS is a vibration detection sensor that operates at radio-frequency (RF) of 2.4

GHz and is capable of measuring very small relative motions over a wide frequency range. The operating frequency range allows the EM waves penetrate approximately 10 cm into the human tissue and reflect back to the sensor [44]. Fundamentally, the GEMS is a homodyne phase interferometer that detects moving objects by transmitting electromagnetic (EM) waves at a target and mixing the reflected waves with the original transmitted wave [41]. All dielectric and conductivity discontinuities caused by tissue-air or tissue-tissue interfaces in the path of the propagating EM waves will reflect the EM waves, which changes the phase of the original transmitted waves. The reflected waves are then detected and mixed with the transmitted waves. The position or motion information of tissues is derived by calculating the phase difference between the received and transmitted waves. In addition, a high pass filter is employed to remove the low frequency signal which is associated with the low frequency clutter from the skin-air interface and other stationary or slowly moving tissue interfaces. Under certain conditions, the output of GEMS, which is called the GEMS signal, is a measure of target position from the sensor as a function of time [43].

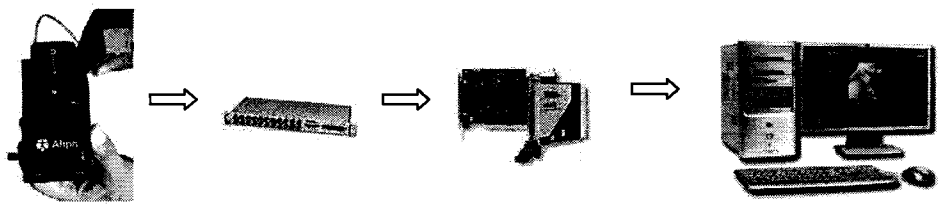
The production of human speech involves in a sequence of coordination of the articulatory organs that includes the tongue, the lip and the vocal fold etc. For example, to produce the sound of \E\ as in "beat", the middle of the tongue has to be raised close to the hard palate while the mouth is kept slightly open. Furthermore, in order to produce voiced sounds (vowels and some consonants such as /b/ /d/ /g/ /v/), the vocal fold has to vibrate while for unvoiced sounds (e.g. /p/ /t/ /k/ /s/ /h/) it does not. The GEMS signal is a measure of the organ motions, which exhibit different patterns when different sounds are produced; therefore, if acquired appropriately, the GEMS signal may be used as a new

modality of ASR by associating the distinct pattern of the GEMS signal with the sound produced.

4.3 Method

The experimental research in this section was reviewed and approved by the Carleton University Research Ethics Committee. Five subjects participated in this experiment. The GEMS signal was collected at the trachea area during speech production of a ten word vocabulary consisting of the digits "zero" to "nine". The GEMS signal was classified using HMMs trained with the AMMI algorithm. First, a combination of features that produces the highest recognition error rate was searched from a group of five candidate features: root mean square (RMS) value, integrated absolute value (IAV), zero crossing (ZC), slope sign change (SSC), and autoregressive (AR) coefficients. The optimal settings for the GEMS signal classification, (i.e., the window size, the window space, and the number of HMM states) were also investigated.

4.3.1 Instrumentation



GEMS Aliph ARV Revision B

I/O Connector NI BNC-2090

ADC PCI-6071E

Computer

Figure 4-1 Instrumentation for the GEMS signal acquisition

The GEMS signal was acquired using the Aliph Radio Vibrometer (Aliph, ARV, revision B) with a neck interface. The neck interface was firmly attached against the skin to eliminate gaps between the antennae and the skin during the measurements. This configuration will enable the near field operation of the GEMS, which will yield the best GEMS signal. The output analogue signal of the GEMS was fed to the data acquisition board (NI PCI-6071E I/O) through an I/O connector (NI BNC-2090) for digitalization. Finally, the digitalized GEMS signal was fed into a computer for further analysis. The sampling rate that was used during the data acquisition was set to be 8000 Hz. The instrumentation settings for the GEMS signal acquisition system were shown in Figure 4-1.

4.3.2 Experimental Method

A 10-word vocabulary consisting of the words "zero" to "nine" was used for this experiment. Data were collected from five subjects (5 males; age range 24 to 35). Each word was repeated 60 times by each subject such that a sufficient number of training and test samples were available for data processing. The daqSPEECH software was used to facilitate the data collection as shown in Figure 4-2. For each repetition, the word to be uttered was displayed on the screen. The start of each signals recording was controlled by an external trigger, which the subjects can press when they are ready to utter the word. The word list was randomized and a two-second pause was inserted between each repetition to reduce any coarticulatory and anticipatory effects. During the data collection, the subjects were instructed to repeat each utterance in a consistent manner such that variations in the utterances (e.g. speaking rate, speaking volume) were minimized.

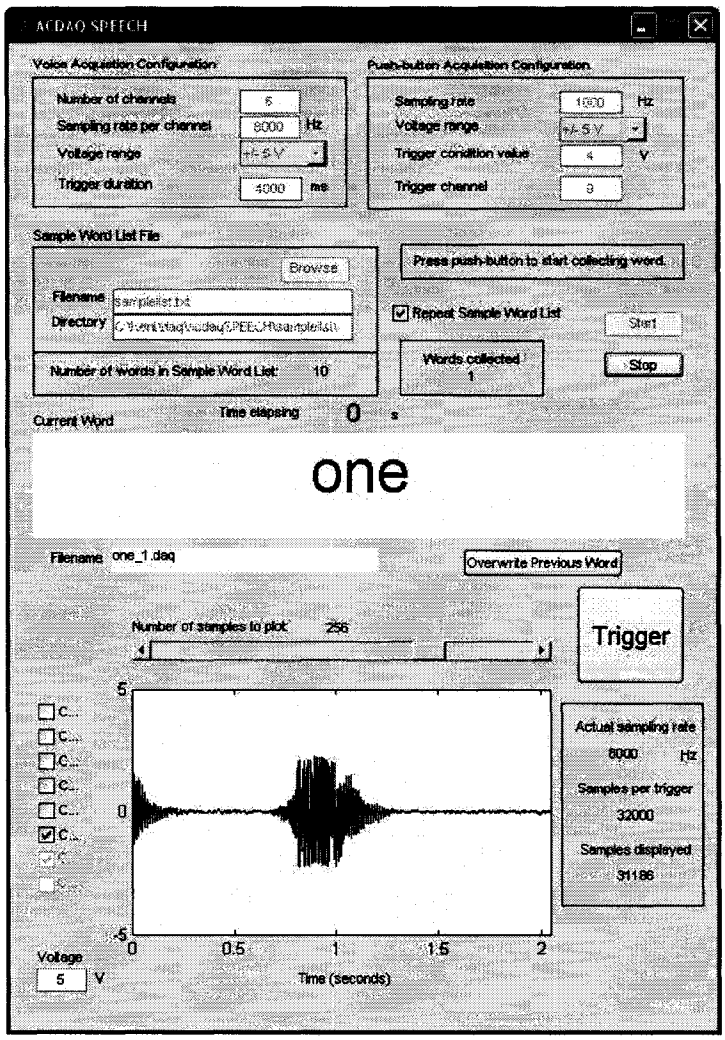


Figure 4-2 Screenshot of the daqSPEECH data acquisition software

4.3.3 Data Processing

The recorded signals were processed offline using Matlab® Version 7.1. The GEMS signal was segmented into a frame of 1000 ms in length using the simultaneously recorded acoustic signal as the trigger signal as shown in Figure 4-3.

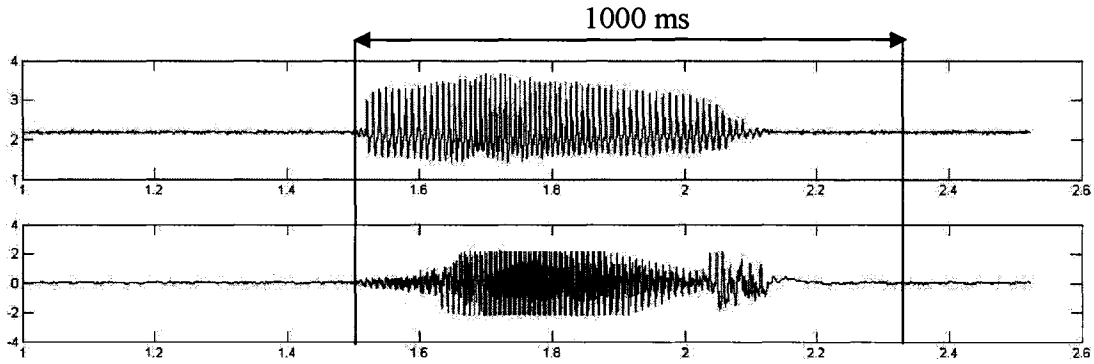


Figure 4-3 Segment of the GEMS signal

The segmented signal was further partitioned into consecutive frames using a sliding observation window. The number of frames depended on the size of the observation window that used and the space between two consecutive observation windows. Within each frame, a set of features, including the RMS, the IAV, the ZC, the SSC and the AR coefficients, were computed.

These features were classified for each subject individually. A left-right HMM with single mixture observation Gaussian densities was used as the signal model for each word. Every other repetition of a particular word in the vocabulary was used as in the training samples; the rest of the repetitions were used as the test samples.

The GEMS signal classification proceeds in the following four steps. First, only AR coefficients were used for classification. We examined the impacts of AR coefficient order on

CER. Second, an optimal features combination, which produced the lowest recognition error rate, was chosen from a set of five time-domain features and their combinations. Third, using the optimal features combination, we investigated the effects of the observation window size and the window spacing on performance. Finally we tried to find the number of HMM state that yielded the lowest CER.

4.4 Results

In this experiment, performance of classification was evaluated by the average CER over five subjects by classifying the test samples.

4.4.1 Order of AR Coefficients

To examine the effect of the order of AR coefficients on CER, We used the first n AR coefficients for classification, where n varied from 1 to 10. During this running, a size of 110 ms observation window with 100 ms spacing and 6-state left-right HMMs were employed. The size of observation window and the number of HMM state were chosen empirically. In each running, a different number of AR coefficients were extracted from the GEMS signal. The training and testing samples was partitioned using the hold-out method as used in previous experiment: half of the total samples were used for training and the rest half for testing. The CER of classification the test samples was recorded after each running. The plot of CER as a function of the number of AR coefficients is shown in Figure 4-4.

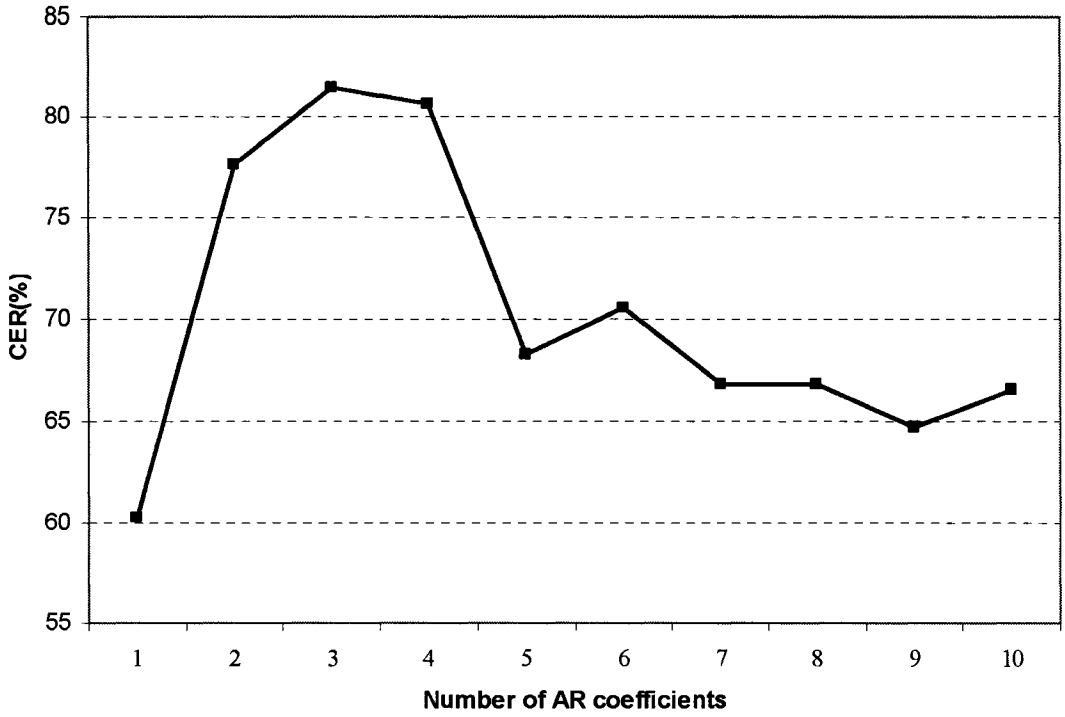


Figure 4-4 CER as a function of the number of AR coefficients

The CER increased significantly from 60.3% to 81.4% then dropped to around 65% when the number of AR coefficients was increased 1 to 10. A peak of CER was observed when 3 AR coefficients were used. The lowest CER was achieved when one AR coefficient was extracted. The result showed that using more AR coefficients does not necessarily yield a lower CER. This may be explained by the non-stationary statistic property of the GEMS signal. In this case the observation window was 110 ms in length, within which the stationary assumption of the AR model did not hold anymore. In this scenario, introducing more AR coefficients would essentially introduce more errors in estimation. Based on this result, it was reasonable to use one AR coefficient in the following experiments.

4.4.2 Feature Selection

In this experiment, we were trying to classify the GEMS signal by a different set of feature combinations. 5 time domain features (e.g. RMS, IAV, ZC, SSC and AR coefficient) and their combinations were used as candidate. In each running, classification was performed using the selected feature combination. The CER on the testing samples was used as criterion to distinguish the effectiveness of different feature combinations in the representation of the GEMS signal. The features were extracted using a 110 ms non-overlapping observation window. The GEMS samples were separated into halves for training and testing. A 10 left-right HMM of 6-state with single mixture observation Gaussian densities was trained and used for testing in each running.

Since 5 features were used, there were only 31 possible combinations. Exhaustive searching through all possible combinations was used in this experiment. The CER attained by using different feature combinations were recorded for each running. As shown in Figure 4-5, depending on the feature combination, the CER fluctuated from 37.6% (RMS, ZC and SSC) to 60.3% (AR). The result showed that the CER tended to drop when more features were extracted for classification: the average CER if more than 2 features were used was 39.2%, the CER increased to 45.2% if the number of features was less than 3. It also showed that the RMS, ZC and SSC combination produced the lowest CER of 37.6%. Based on this experimental result, it seems that among the 5 features, the speech information of the GEMS signal was better captured by the following 3 physical characteristics associated with the GEMS signal: power (RMS), frequency (ZC) and change of amplitude (SSC). In the following sections, we used the 3 features for the GEMS signal classification.

When determining the optimal parameters for classification, including features, window sizes, spacing, and the number of HMM states, there is no analytical solution. Thus, we determine an empirically optimal set of parameters by experimentation: different size of observation window and HMMs with different number of states were used to classify the GEMS signal. Classification performance was evaluated by CER. Due to the complexity and size of the search space, the searching scheme we employed was rudimentary. The point was to show the potential of the GEMS in ASR and the classification accuracy can be improved by carefully tuning the classifier and to determine a reasonable operating point.

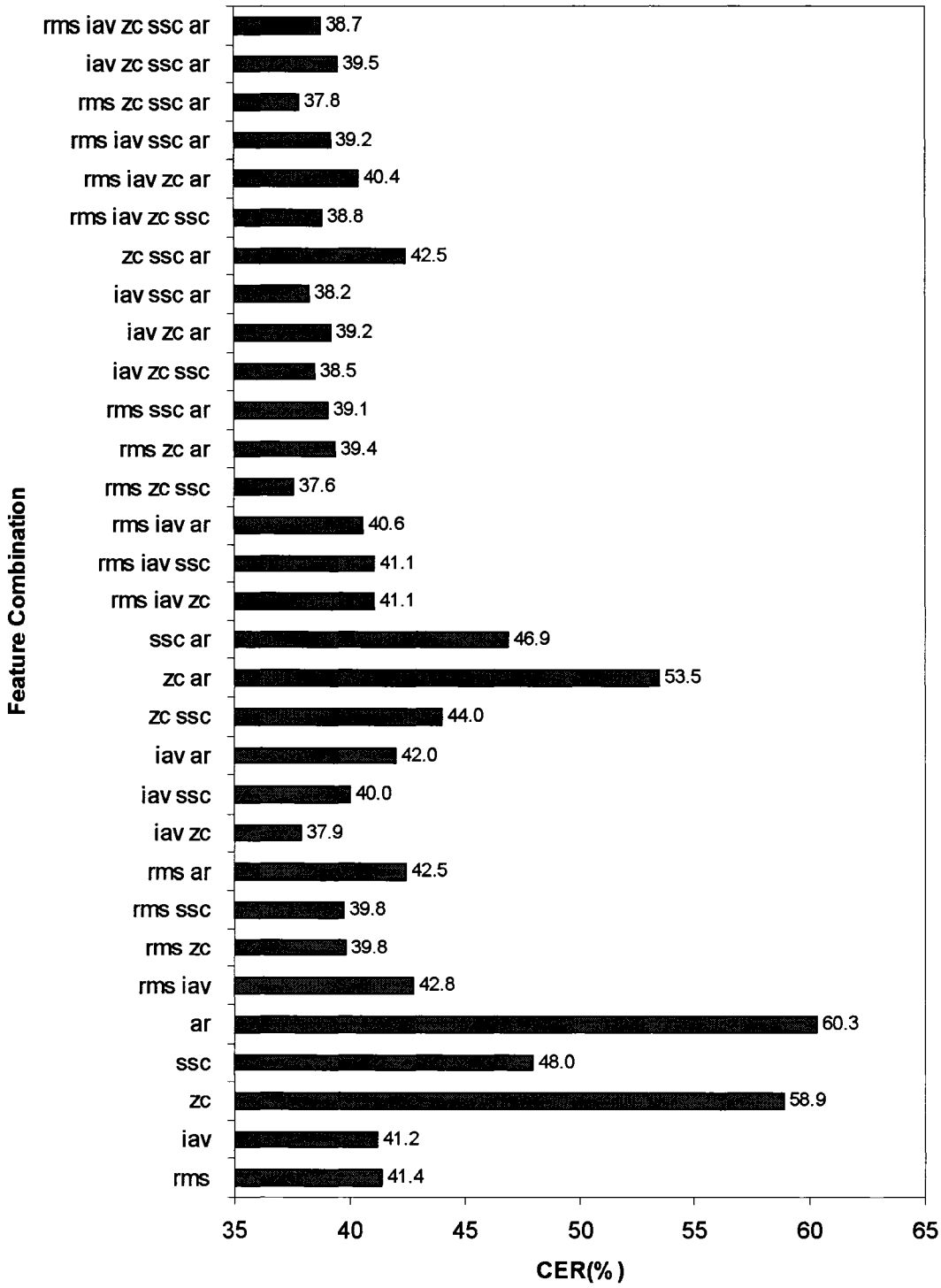


Figure 4-5 CER as a function of feature combination

4.4.3 Observation Window Size

The effect of observation window size on CER was investigated in this section. Based on the results from previous experiments, the window spacing was set to be 100 ms, the size of observation window was changed from 10 ms to 200 ms with a step of 10 ms. RMS, ZC and SSC were extracted within each observation window. A 6-state right-left HMM with single mixture observation Gaussian densities was used for classification.

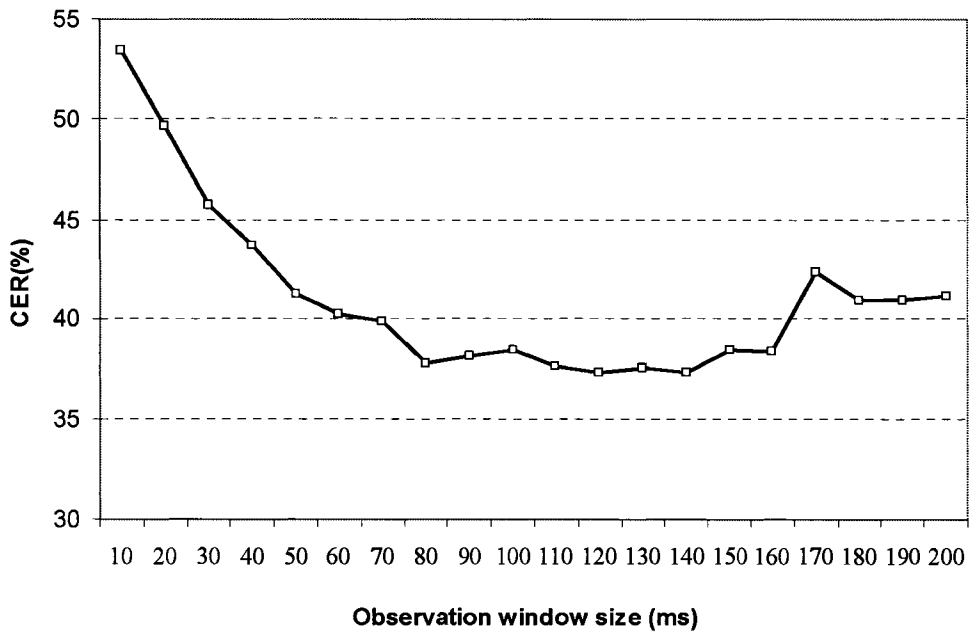


Figure 4-6 CER as a function of observation window size

The CER as a function of observation window size was showed in Figure 4-6. The CER decreased gradually from 53% to 40% when the window size increased from 10 ms to 70 ms. The rate levelled off around 38% until the window size was increased to 160 ms. The lowest CER of 37.3% was achieved with a 120 ms observation window. The performance deterioration on both sides was not unexpected. When the window spacing was fixed, a small observation window would have too much variability relative to the useful

information (i.e. essentially a poor SNR); contrarily, too large window size would be unable to capture the local variations of the GEMS signal.

4.4.4 Observation Window Spacing

The observation window spacing on the CER was examined in this section. The observation window size was set to be 120 ms based on the result of Section 4.4.3. The window spacing was increased from 20 ms to 110 ms with a step of 10 ms. RMS, ZC and SSC were calculated within each observation window. The GEMS samples were separated in half for training and testing. 6-state right-left HMMs with single mixture observation Gaussian densities were used as classifier.

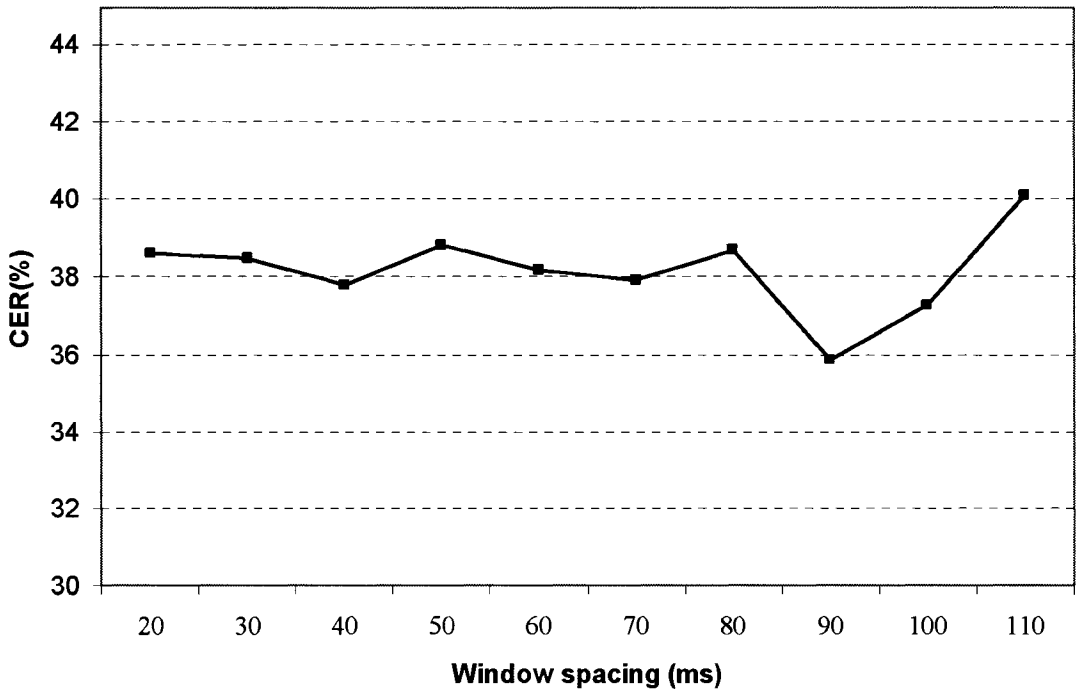


Figure 4-7 CER as a function of window spacing

Figure 4-7 showed the result of CER as a function of window spacing. The CER fluctuated around 38% when the spacing increased from 20 ms to 80 ms. A smallest CER of 35.9% was observed at window spacing of 90 ms. The CER increased when large window spacing was used after 90 ms. The relatively smoothness of the CER between window spacing of 20 ms and 80 ms may be explained by the domination of the observation window size (120 ms) when a relatively small spacing was used for feature extraction.

4.4.5 Number of HMM States

The number of HMM states also has impact on the classification accuracy. In this section, the observation window size and window spacing was set based on the results from previous searching: the observation window was 120 ms in length and the window spacing was set to be 90 ms. RMS, ZC and SCC were calculated within each observation window. Right-left HMMs with single mixture observation Gaussian densities were constructed as classifier. Half of the samples were used for training and the rest for testing. This experiment started with a 2-state HMM. After each running, the number of states was increased by 1 until 10 states were used. The CER as a function of the number of HMM states was depicted in Figure 4-8. The CER decreased monotonically as the number of HMM states increased, from 48.9% (2-state HMM) to 31.1% (10-state HMM). HMM classification essentially was used the HMM to model the statistical distribution of the time series, which may be time-variant in this case, thus the number of HMM states should be enough to model the statistical variation over time. It indicated from the result that a sufficient large number of HMM states (> 7) was required in order to achieve a satisfactory classification accuracy on the given database. However, the number of states cannot be increased

arbitrarily: as the number of states increases, the parameters to be estimated will also increase significantly, which, in turn, requires more training samples available for training the HMM accurately.

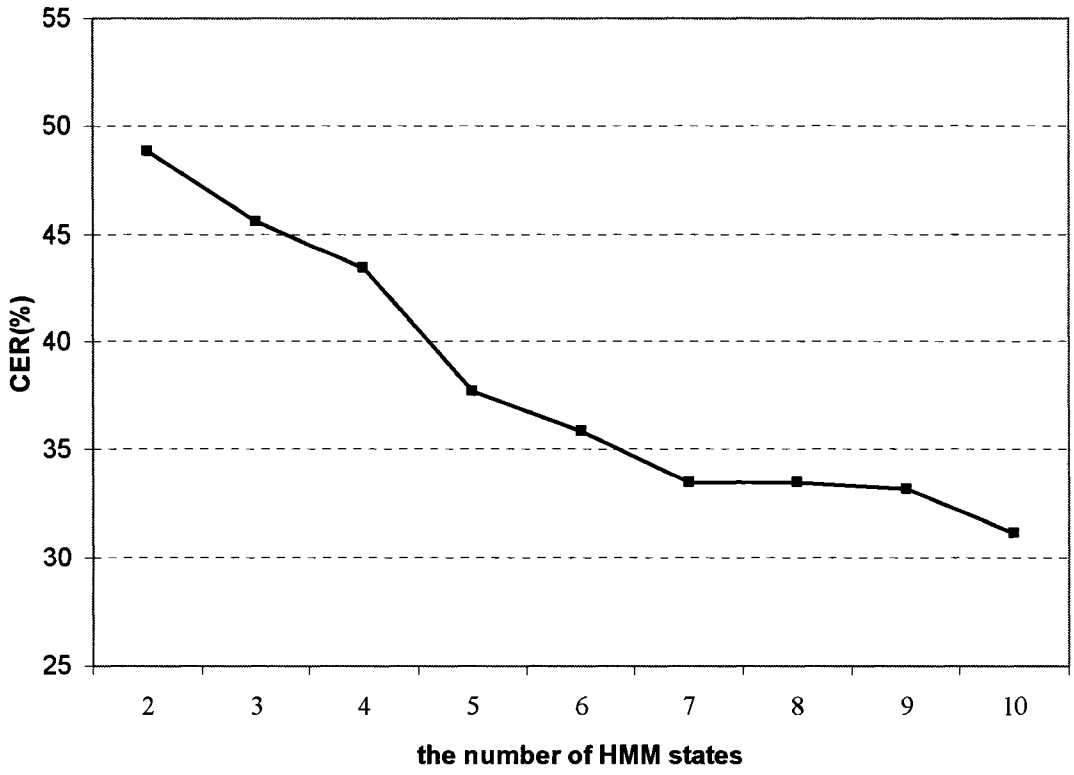


Figure 4-8 CER as a function of the number of HMM states

4.5 Discussion

In this chapter, the GEMS signal at the trachea area was collected from five subjects for ASR. Five time-domain features were extracted from the recorded GEMS signal. Several feature combinations were attempted for classification. A HMM classifier was constructed using the AMMI as described in pervious chapter. A CER of 31.1% was achieved through an empirically analysis.

The CER was much smaller than the *a priori* rate, obtained by random guessing (i.e. 90% for ten-word vocabulary), which implies the availability of speech related information in the GEMS. There were two reasons that might account for the experimental results. First, the GEMS signal collected from the neck essentially was a summation of the various tracheal tissue vibrations during uttering if it was properly filtered. When a different word was uttered, a different pattern of tracheal tissue vibrations was presented, which, in turn, changed the characteristics of the GEMS signal (e.g. shape, amplitude, power level etc.). These sequential variations in the GEMS signal were then captured by a well-trained classifier to identify the word uttered. Second, the vibration of the trachea is associated with the utterance of a voiced sound. Since different words consists of different voiced sounds which appear in different positions in the word, these voiced sound sequences are likely the main source of the speech information.

To extract the speech information which contain in the GEMS signal, a classifier must be constructed. Considering the prowess of HMM in the effectively modeling of various complex time series, we also adopted the HMM to construct the classifier. It can be anticipated that the feature extracted scheme (feature set, observation window size, etc.) and the structure of HMM have significant impact on the classification accuracy.

While the higher orders of AR coefficients showed its effectiveness for classification of certain signals, e.g. the acoustic signal and the MES signal. It seemed that the usage of higher orders AR coefficients failed to improve the classification performance, which might imply that the higher orders AR coefficients did not provide extra information regarding speech. This is not unexpected as the GEMS signal is quite sinusoidal in nature, which implies that only a low order AR model is required.

There were 5 time domain features and their combinations were examined. The CER trended to decrease when more features were used for classification. The average CER when more than 2 features were used was 39% compared to 45% when less than or equal to 2 features were used. The combination of RMS, ZC and SSC was founded to yield the lowest CER compared with other combinations. Having additional features caused an increase in the dimensionality of the problem, raising the CER. With a larger training set, it is possible that a lower CER could be achieved.

The size of observation window also had significant effect on the classification performance. Inappropriate observation window size will decrease the classification accuracy. A window size between 80 ms to 160 ms was founded to be suitable for the database. Too small window size was unable to fully extract the information contain in the GEMS signal while too large window size failed to characterize the variations of the signal.

The number of HMM states should be suffice to classify the GEMS signal with satisfactory accuracy. However the number of states was upper bounded by the training samples which were available for the estimation of HMM parameters.

A rudimentary approach was introduced for feature extraction and construction of HMM

classifier. The experimental results showed that the lowest CER was obtained using a feature combination of RMS, ZC and SSC, which was extracted by a 120 ms observation window with a 90 ms spacing, and classifying by 10-state left-right HMMs with single mixture observation Gaussian densities.

The settings (i.e. feature combination, observation window size, the number of HMM states, etc.) for the GEMS signal classification, which were found by the proposed approach in previous sections, however, may not be globally optimal. For example, frequency domain features, which were not mentioned so far, may be useful as well. In addition, if the searching strategy that was employed or the initial search point were changed, the results may differ. The objective of this section is not to find global optimal CER which is beyond the scope of this study. The point was to demonstrate the potential of the GEMS signal for ASR. Some optimization was performed to provide an indication of the performance level of this modality. However, the results from the experiment are encouraging. If measured and processed properly, the GEMS signal has potentials to be used for ASR, which could become another promising supplement of the conventional acoustic ASR other than the MES.

Chapter 5 Multimodal Speech Recognition

5.1 Introduction

Performance of a conventional acoustic ASR system degrades severely in the presence of noise. Despite ongoing research improving unimodal ASR systems, which relies solely on acoustic information, these will eventually saturate in performance and arguably already have begun saturating. One of feasible solution to this problem is to supplement the acoustic signal with other speech related signals, which are immune to acoustic noise, resulting in a multimodal ASR system [45]. In this chapter, we propose a multimodal ASR system using the acoustic signal, the MES, and the GEMS signal. Two types of multimodal ASR systems were constructed as shown in Figure 5-1. Type I was constructed by fusing the classification outputs of the three HMM classifiers. Type II was constructed using a single HMM classifier that used all three types of signals as inputs. Relying on more information for decisions, the multimodal ASR system is expected to outperform the unimodal acoustic ASR system.

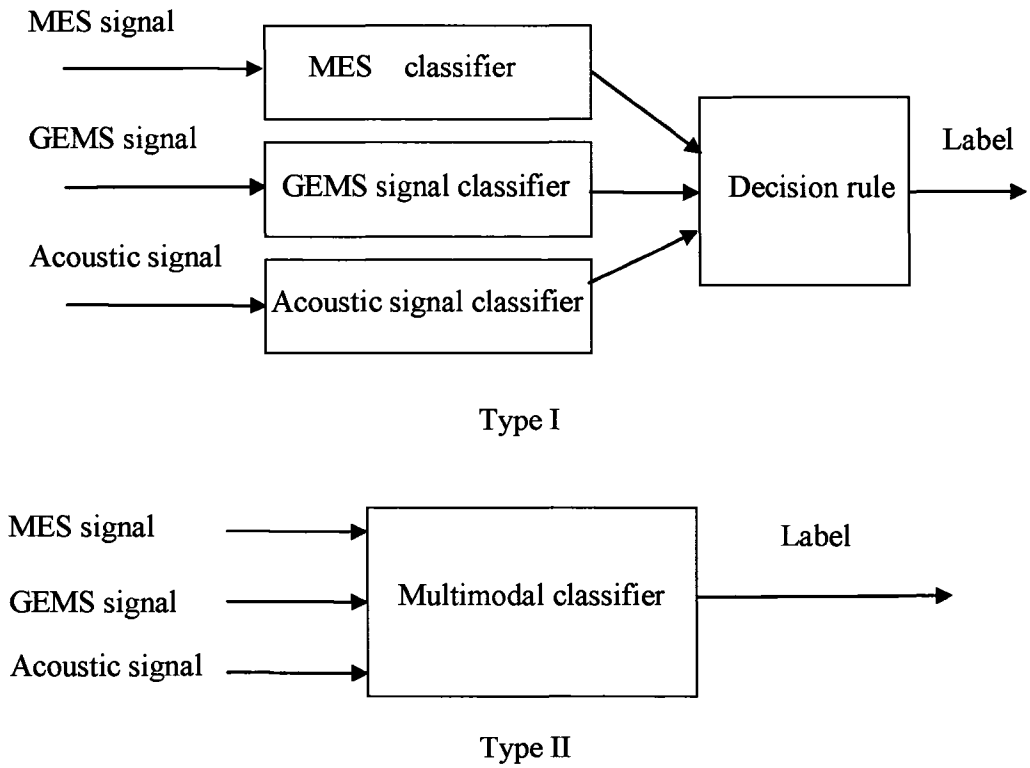


Figure 5-1 Type I and Type II multimodal ASR system

5.2 Measurement Setup

The experimental research in this section was reviewed and approved by the Carleton University Research Ethics Committee. Three types of speech relevant signals were collected from five subjects: the acoustic signal, the MES, and the GEMS signal.

MES were collected from three articulatory muscles around the mouth: *zygomaticus major* (ZYG), *platysma* (PLT), and *depressor anguli oris* (DAO). Three pairs of Myo-Tronics Duo-Trode disposable Ag/Ag-Cl electrodes were placed on the muscles for the MES collection. One Red-Dot Ag/Ag-Cl electrode from 3M (Montreal, QC) was placed on the left wrist to provide a common ground reference. Before positioning the elec-

trodes, a small coating of electro-medical gel from MyoTronics (Tukwila,WA) was applied on the surface of electrodes to reduce the electrode-skin impedance. The GEMS signal was collected at the trachea area using the Aliph Radio Vibrometer (Aliph, ARV, revision B). The acoustic signal was recorded simultaneously using a microphone during the data collection. The acoustic signal was used for both acoustic ASR and data segmentation.

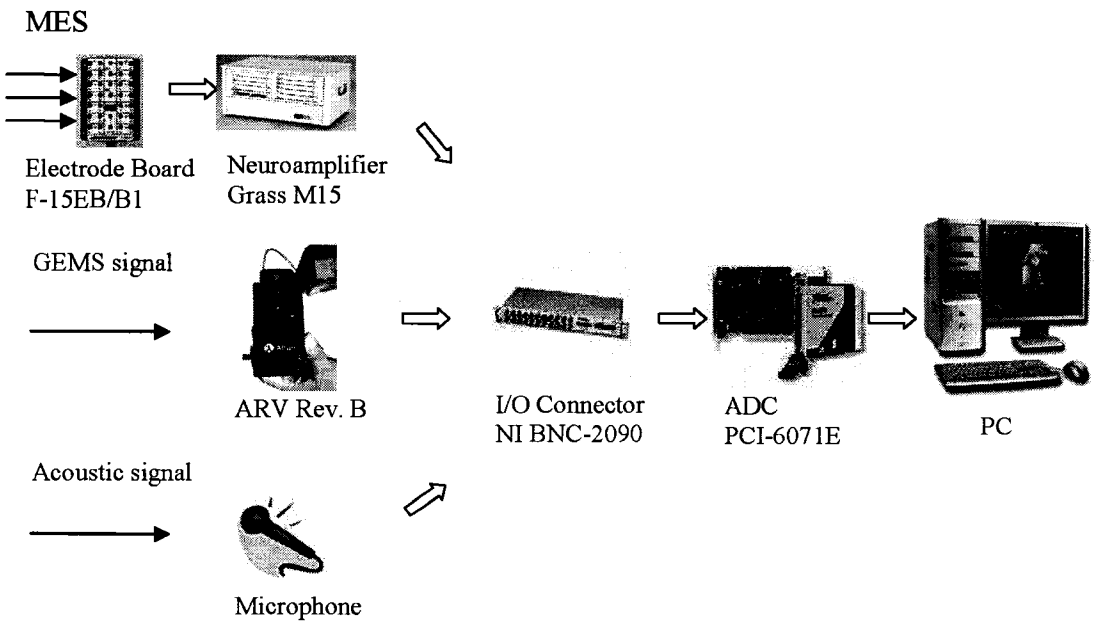


Figure 5-2 Measurement setup for acoustic, MES, and GEMS signal

MES were fed into the data acquisition system using an electrode board (Model F15EB/B1 from Grass, West Warwick, RI). The MES were amplified differentially using a programmable multi-channel amplifying system (M15A54, Grass, West Warwick, RI). The MES were bandlimited between 30 Hz to 300 Hz to eliminate the electrical noise outside the MES bandwidth. In addition, a 60 Hz notch filter was applied to each MES channel to eliminate power line noise interference. The GEMS signal was collected using

the neck interface which enables the ARV to operate in near field. The National Instruments PCI-6071E I/O (Austin, TX) data acquisition board was used to digitalize the analogue signals; that is the MES, the GEMS, and the acoustic signals. The sampling rate that was used in this data collection was 8000 Hz per channel. A system diagram of the data acquisition setup is shown in Figure 5-2.

5.3 Data Collection

A 10-word vocabulary consisting of the words "zero" to "nine" was used for this experiment. Three subjects participated in the data collection. Five channels of signals, including the three MES, the GEMS signal, and the acoustic signal were collected from each subject. Each word was repeated 60 times by each subject such that a sufficient number of training and test samples were available for data processing. The daqSPEECH software was used as user interface for the data collection as in Chapter 4 (Figure 4-2). Subjects were instructed to utter the word 1 second after the start of data recording such that the MES prior to the utterance could be recorded. The word list was randomized to reduce any coarticulatory and anticipatory effects. In addition, a two-second pause was inserted between each repetition such that the articulatory muscles could start contraction from relax status at the beginning of each utterance. During the data collection, the subjects were instructed to repeat each utterance in a consistent manner such that the variations were minimized (e.g. speaking rate, speaking volume).

The recorded signals were processed offline using Matlab® Version 7.1. The raw data were segmented into 1000 ms of length. The acoustic signal was used to position the start of each utterance for segmentation. The acoustic signal and the GEMS were segmented

from the utterance start point. The MES were segmented 500 ms prior to the utterance start point as the MES activity precedes the acoustic speech signal. The data segment scheme is depicted in Figure 5-3.

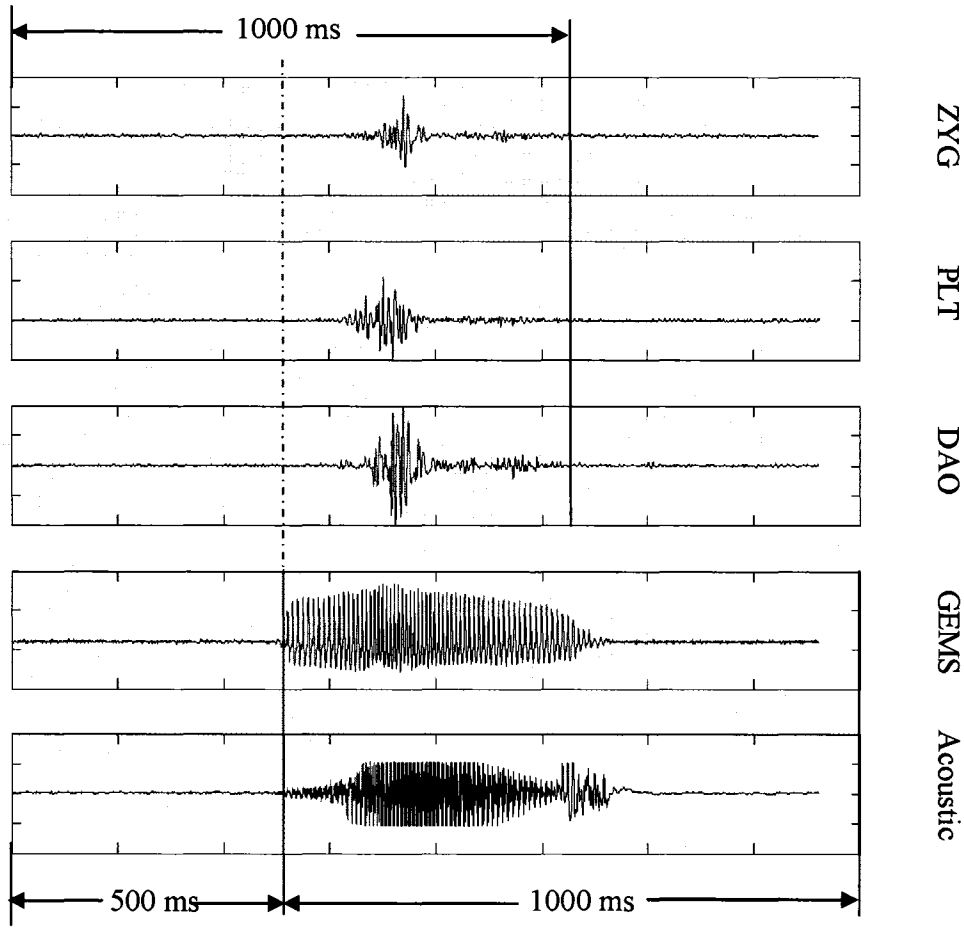


Figure 5-3 Segmentation of the MES, the GEMS signal and the acoustic signal

5.4 Type I Combination

5.4.1 Classifiers

Three HMM classifiers were built to classify the GEMS signal, the MES and the acoustic signal respectively. The feature extraction scheme and HMM classifier construction were

based on the results from previous chapter.

Classification of the MES was based on the results from Chapter 4. Two types of time-domain features including RMS and two AR coefficients were extracted using a 128 ms observation window which spaced by 16 ms. The MES data were separated into two parts: half of the data was used for training and the rest for testing. 10-state left-right HMMs with single mixture observation Gaussian densities were employed for the MES classification.

Classification of the GEMS was performed on the base of the results from Chapter 5. Three time-domain features including RMS, ZC and SSC were extracted using a 120 ms observation window which spaced by 90 ms. Each other recorded GEMS was used for training and the rest for testing. 10-state left-right HMMs with single mixture observation Gaussian densities were used for the GEMS classification.

To construct a classifier for the acoustic signal classification, a searching process similar to Chapter 3 was applied to find an empirically reasonable operating point for an HMM classifier. As a result, RMS and two AR coefficients were extracted using a 25 ms observation window which spaced by 10 ms. 6-state left-right HMMs with single mixture observation Gaussian densities were used for the acoustic signal classification.

5.4.2 Combination of Classifier Outputs

The outputs of the three classifiers were combined for the final decision based on certain combination method. Two types of combination: majority vote (MV) and score-combination (SC) were used to fuse the classifier outputs in this experiment. A mathematic description of classifier combination was provided in the following sections. De-

fine Θ as a pattern space which is the union of the M mutually exclusive classes $C_i \left(i \in \Lambda = \{1\,2 \cdots M\}\right)$,mathematically,

$$\Theta = \bigcup_{i=1}^{M} C_i \qquad \textbf{(5-1)}$$

5.4.2.1 Level of Classifier Outputs

Suppose there are K classifiers, denoted as $e_k \ (k = 1\,2 \cdots K)$ to classify an unknown pattern $X \in \Theta$. Generally, the output of a classifier can be one of the three levels [10].

Abstract Level: The unknown pattern X is assigned an index $i \in \Lambda$ if a classifier e_k believes $X \in C_i$.

Rank Level: The unknown pattern X is assigned by a classifier e_k a rank vector $V_k = \left[v_k(1)\, v_k(2) \cdots v_k(M)\right]^T$, which is a permutation of the set Λ. The order of $v_k(j)$ in the permutation indicate the degree of confidence that the classifier e_k believes $X \in C_i$,where $i = v_k(j)$.

Measurement Level: The unknown pattern X is assigned by a classifier e_k a measurement vector $S_k = \left[s_k(1)\, s_k(2) \cdots s_k(M)\right]^T$. Element $s_k(i)$ is a numerical value which indicates the degree of confidence that the classifier e_k believes $X \in C_i$.

5.4.2.2 Majority Vote

In the majority vote, the assignment of an unknown pattern X to class C_i by classifier e_k can be denoted mathematically by the following binary relationship.

$$T_k(X \in C_i) = \begin{cases} 1 & \text{if } e_k(X) = i, \text{ where } i \in \Lambda \\ 0 & \text{otherwise} \end{cases} \quad \textbf{(5-2)}$$

Denote $E(X)$ be the combination of classifier outputs for an unknown pattern X. A common form of the majority vote rule is to assign $E(X) = i$ if the majority number of the classifiers believe $X \in C_i$ and the number exceeds a certain percentage of the total number of the classifiers K, that is,

$$E(X) = \begin{cases} \arg_i \left\{ \max \left[\sum_{k=1}^{K} T_k(X \in C_i) \right] \right\} & \text{if } \max \left[\sum_{k=1}^{K} T_k(X \in C_i) \right] \geq \alpha K \\ M+1 & \text{otherwise} \end{cases} \quad \textbf{(5-3)}$$

The constant α is between 0 and 1 and is used to control the threshold for a valid decision. Specifically, for the majority vote combination in this section $E(X)$ is assigned an index i if more than one classifier agrees upon a proposition. If no agreement is present, then the majority vote assigns a value of M+1, indicating that a decision was no possible; however, considering the relatively high accuracy of the acoustic signal classifier, when none of the three classifiers were agree upon, the proposition of the acoustic signal classifier was selected instead.

5.4.2.3 Sore-combination

Sore-combination can only be applied to classifiers which have an output on the measurement level. In order to combine the outputs of different classifiers, the measurement vector $S_k = [s_k(1)\, s_k(2) \cdots s_k(M)]^T$ first must be normalized so that each measurement is in the same range. Various nonlinear transformations can be applied for the normalization.

The normalized scores can then be fused with basic mathematical operations such as sum, difference and weight. In this study, the likelihood values were used as scores and a logarithmic transformation is used for measurement vector normalization, with the sum operation used for combination, which was selected based on the results of a previous paper [10]. Since score-combination makes full use of each classifier's output, it is expected to be superior in performance compared with major vote, which only considers the ranking of individual classifier and simply ignores the score information.

5.4.3 Results

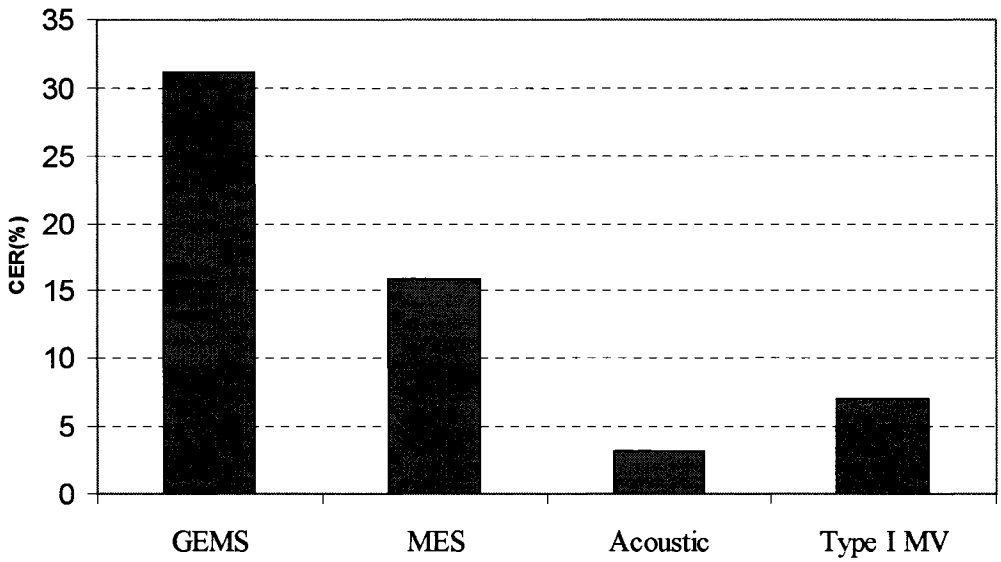


Figure 5-4 CER by unimodal classification and by Type I majority vote

Figure 5-4 showed the CER of the unimodal system and the multimodal system by Type I majority vote. The CER for the unimodal system was 31.1%, 16% and 3.1% when the GEMS signal, the MES and the acoustic signal was used respectively. The CER for the MES classification is slightly higher than the result obtained in Chapter 3 (12.8%). Con-

sidering only 3 channels of MES were collected, compared to 5 channels in the Chapter 3 data, the results are reasonable. The acoustic classifier had the lowest CER of 3.1% partly due to the acoustic signal was recorded in quiet lab environment. The judgments of the three classifiers were combined by the majority vote, resulting in a CER of 7.1%. While the combination yielded a CER better than the standalone GEMS signal or MES classifier, the rate was still higher than the acoustic signal classifier. This result also confirmed the fact that majority vote does not necessarily yield an accuracy better than any of the individual classifier. In the worst case, when the pattern of failure occurs, the accuracy of majority vote can be worse than an individual classifier.

In majority vote, the combination was performed in a way such only the class which each classifier assigned the highest score was considered. Remember the HMM classifier assigned a belief (score) to each class which it considered an unknown pattern belonged to. Thus, it was reasonable to utilize that information for more accurate final decision. This can be achieved by a method called score-combination.

Figure 5-5 showed the CER comparison of unimodal classifiers and multimodal classifier by score-combination. The CER dropped to 1.4% when the results from individual classifier were combined using score-combination method. Performance of the multimodal classifier was better than any of the individual standalone classifier. Score-combination also outperformed majority vote (7.1%), an improvement around 80%. Compared with majority vote, it would seem that score-combination fused the judgment of each classifier more effectively by utilizing the information which was ignored by majority vote. It was also noticed that the score-combination multimodal classifier had a lower CER than the unimodal acoustic classifier; this would be particularly useful when the acoustic signal

was contaminated by noise. It could be anticipated that the combination method would provide a robust mechanism for maintaining classification accuracies against ambient noise.

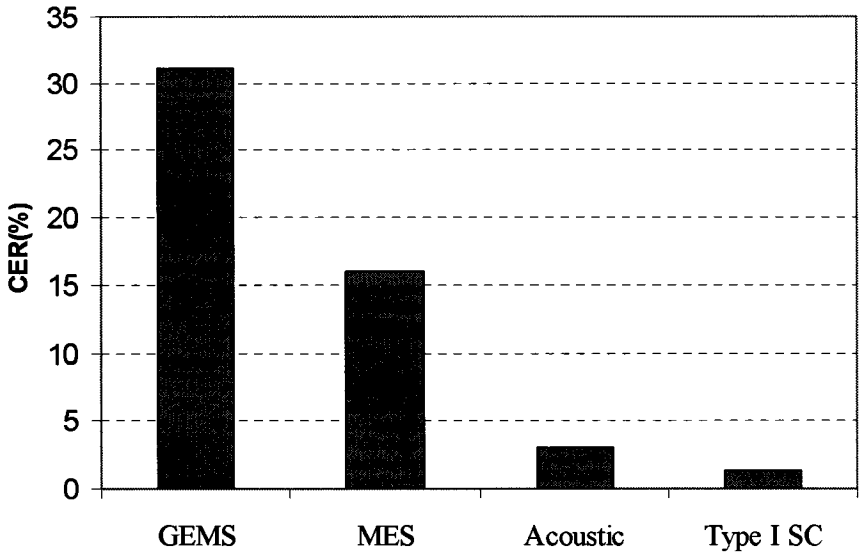


Figure 5-5 CER by unimodal classification and by Type I score-combination

5.5 Type II Combination

5.5.1 Implementation

Due to the differences between measurement mechanism and physical characteristics, the MES, the GEMS signal and the acoustic signal can be assumed to contain orthogonal information regarding speech. Combined together, these signals would provide more information for ASR. Another reasonable approach to utilize these information sources is to build a classifier which takes the three signals (the MES, the GEMS and the acoustic signal) as input based on the assumption that the three signals represent a distinct speech pattern.

To construct a classifier for this propose, the feature extraction scheme and the structure of the HMM must be determined. One major difficulty lies in the fact that the HMM classifier can not be optimized for all signal types at the same time in this case. Rather, the goal was to construct a classifier which produced the best overall performance.

It had been shown that 10-state left-right HMMs with single mixture observation Gaussian densities were adopted for both the GEMS signal and the MES classification for optimal performance. We also noticed that there was only slight difference in CER for the acoustic signal classification with 6-state HMM and 10-state HMM. Thus, it made sense to use 10-state left-right HMMs with single mixture observation Gaussian densities as the "all-in-one" classifier.

Special considerations also must be taken into account for feature extraction scheme. Since only one classifier was used, when performed feature extraction, the observation window spacing must be kept the same among all the signals to construct the feature ma-

trix; however, previous experiments had shown that the optimal window spacing was different for different signals, i.e., 16 ms for the MES, 10 ms for the acoustic signal and 90 ms for the GEMS. Bearing in mind that the goal is to achieve the overall performance, again, we conducted experiments to determine an empirically optimal value for the window spacing, resulting in a 16 ms window spacing.

When the value of the window spacing was fixed, different sizes of observation window can be used to extract feature from different signals. Based on the results from previous Chapters, an observation window of size of 128 ms, 120 ms and 25 ms was used for the MES, the GEMS signal and the acoustic signal, respectively. In addition, within the observation window, different feature sets can be extracted. Based on the results from Chapter 3 and Chapter 4, RMS and two AR coefficients were extracted for the MES; RMS, ZC and SSC for the GEMS signal, RMS and two AR coefficients for the acoustic signal. The segmentation process for a single window is illustrated in Figure 5-6. The next window follows the same segmentation process, with each window offset by the 16 ms window spacing.

All types of signals were partitioned in half, the odd number samples were used for training and the even number samples were used for testing. Performance were evaluated by the CER in classification the test samples.

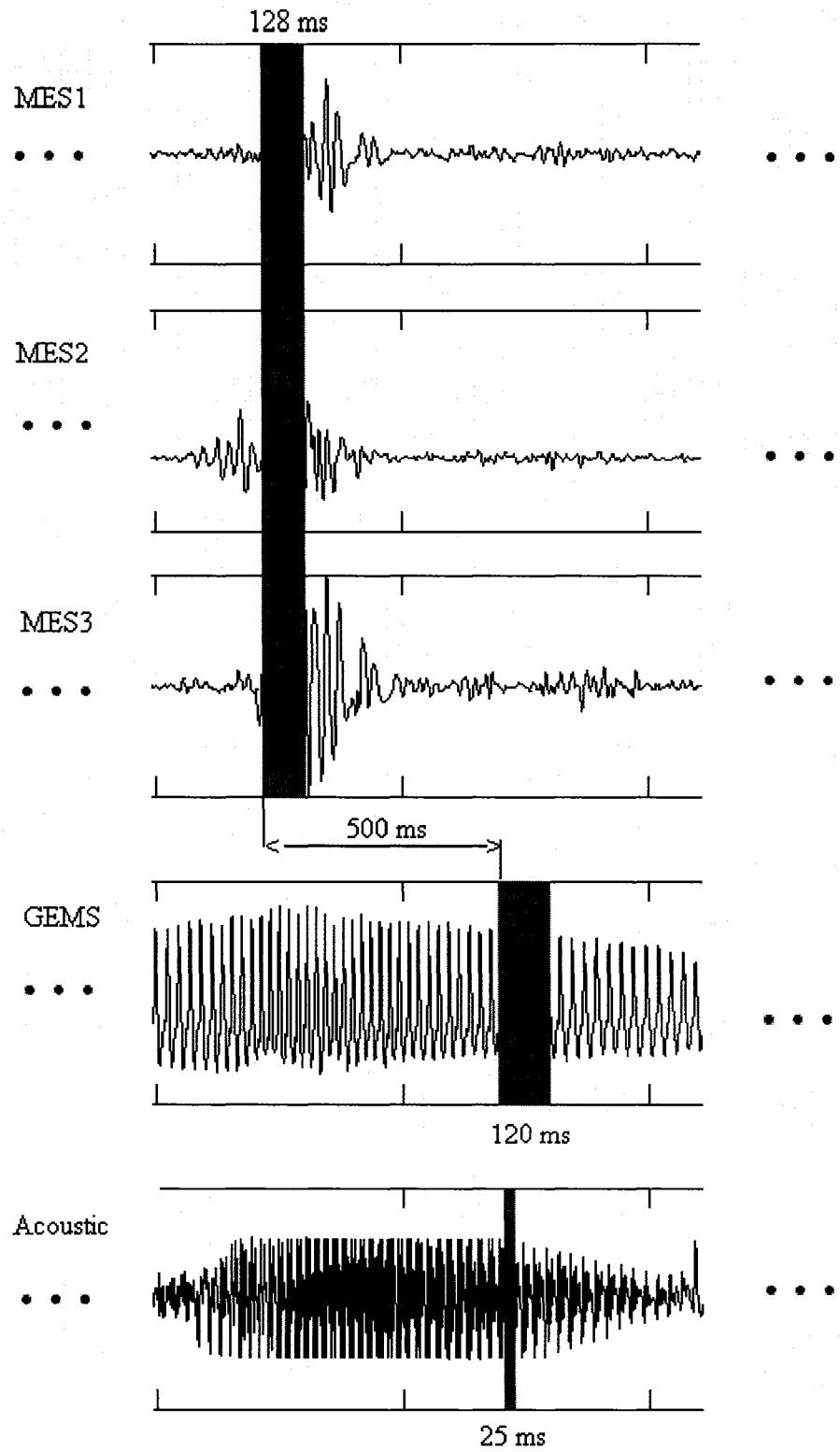


Figure 5-6 Segmentation process for the combining classifier

5.5.2 Results

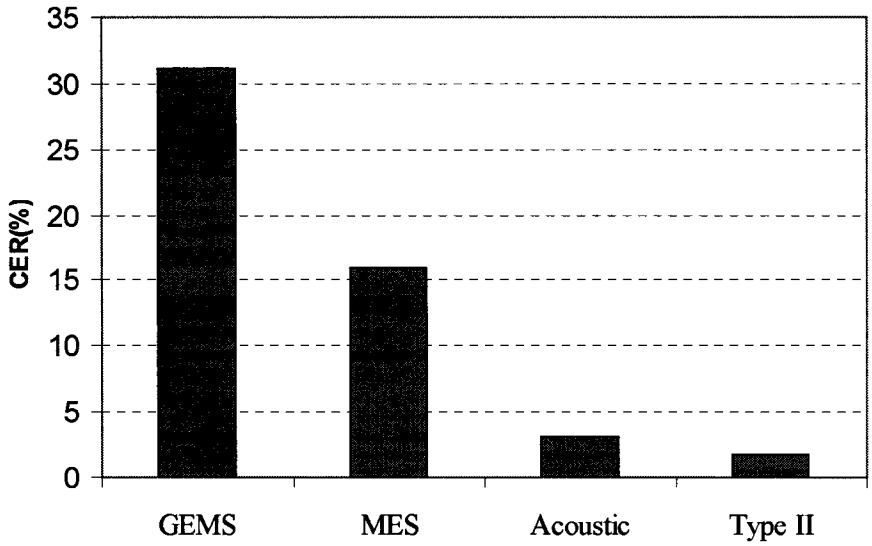


Figure 5-7 CER of unimodal classification and Type II combination

The CER of the unimodal classification and the multimodal classification by Type II combination were showed in Figure 5-7. The CER dropped to 1.8% by the "all-in-one" classifier. The CER was superior to any of the three unimodal classifier. A nearly 42% improvement over the acoustic classifier was observed. By using the MES and the GEMS signal as supplement to the acoustic signal for ASR, the proposed classifier seemed being able to discriminate the speech the unimodal acoustic classifier had difficult with, thus attaining a gain in the recognition accuracy. In addition, the CER was comparable with the rate obtained by score-combination method in previous section (1.4%). However, since this approach of combination used only one classifier instead of three in score-combination, there was the advantage of reducing the computational overhead and easing

classifier implementation.

5.6 Multimodal ASR in Noisy Environment

5.6.1 Method

Two types of multimodal ASR systems had been proposed by combining the judgments of standalone ASR systems or by constructing an "all-in-one" classifier which used the MES and the GEMS signal as supplement to the acoustic signal. It was shown that the multimodal systems provided more accurate classification if an appropriate combination method was chosen. Indeed, the CER could be lower than the rate attained by a stand-alone acoustic system even under quiet lab environment. In this section, performance of the multimodal ASR systems was evaluated under noisy conditions.

The samples were separated in half. The odd numbered samples were used for training and the rest for testing. Specially, test samples of the acoustic signal were contaminated with different levels of additive white Gaussian noise (AWGN). The intensity of AWGN was measured by signal-to-noise-ratio (SNR) which was defined as the ratio of the power of the signal to the power of the AWGN. The SNR used in this experiment ranged from 0 dB to 50 dB. When the SNR was equal to 0 dB, the power of the signal was equal to the power of the AWGN.

A unimodal acoustic classifier and two types of multimodal classifiers (Type I and Type II) were constructed as described in previous sections. The classifiers were trained using training samples which were not contaminated by AWGN, and then were used to classify the test samples contaminated by different levels of AWGN. Performance was evaluated using the CER.

5.6.2 Results

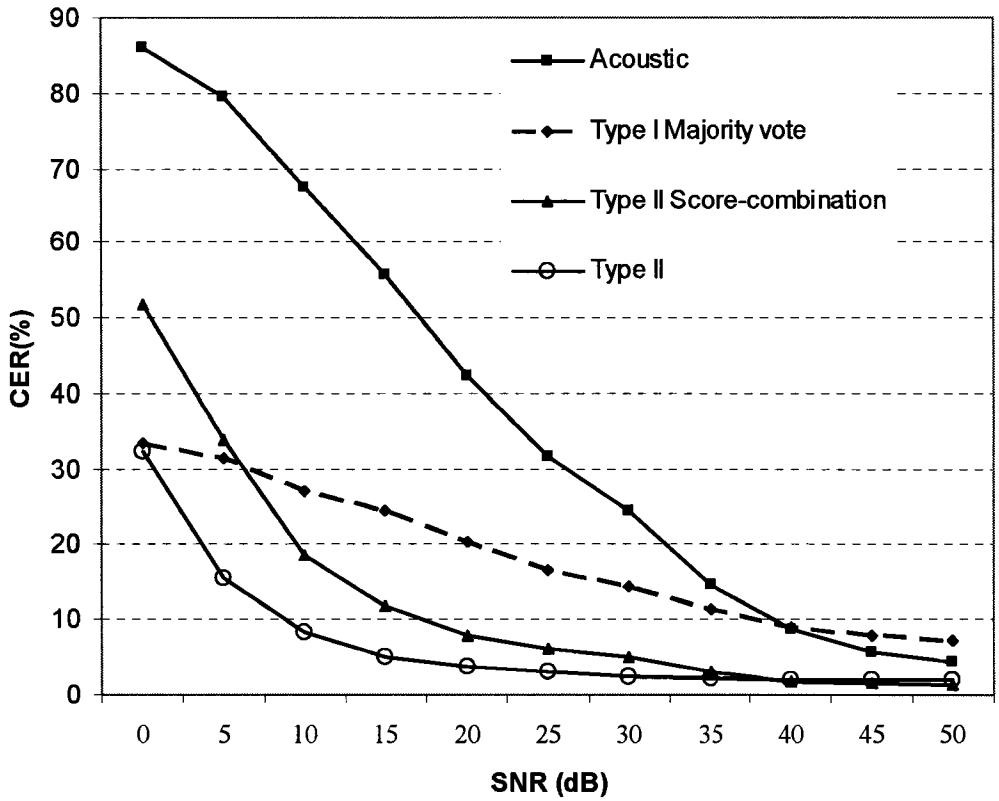


Figure 5-8 CER as a function of SNR

Figure 5-8 showed the plot of CER as a function of SNR for unimodal acoustic classifier and two types of multimodal classifiers. As expected, the CER for the acoustic classifier degraded dramatically from 4.4% to 86% as the SNR decreased from 50 dB to 0 dB. When the SNR was as small as 0 dB, the CER of the acoustic classifier was close to random guessing (i.e. 90%).

The CER for majority vote decreased from 33.3% to 7.3% as SNR increased from 0 dB to 50 dB. In terms of CER, majority vote outperformed the acoustic classifier when the SNR was less than 40 dB from where it seemed that the MES and the GEMS signal com-

plemented the contaminated acoustic signal in providing speech information. However, when the SNR was high, performance of majority vote was slightly worse than the acoustic classifier. This may partly due to the relatively higher CER of the MES and the GEMS signal classifier under high SNR environment which presented themselves as outliers to the acoustic classifier.

The CER for score-combination decreased from 51.8% to 1.4% when SNR increased from 0 dB to 50 dB. When the SNR was larger than 10 dB, score-combination was superior to both majority vote and the acoustic classifier. When the SNR was 0 dB, however, the CER was 51.8%, worse than the rate of majority vote (33.3%).

The CER for Type II combination decreased from 32.4% to 1.9% as SNR increased from 0 dB to 50 dB. This type of combination yielded the lowest CER among all classifiers when the SNR was small (less than 40 dB). For high SNR (greater than 40 dB), its performance was also close to the score-combination method, which was the method that yielded the lowest CER under these conditions.

When the SNR was less than 10 dB, the rate of performance degradation of score-combination and Type II combination was much great than that of majority vote. Notice that the rate was comparable to that of the acoustic classifier, the deteriorating acoustic signal accounted for the degradation. The noise in the acoustic signal will propagated to the score output of HMM classifiers. The majority vote essentially acted as a filter to these noisy scores, which had a smooth effect on the degradation rate. This could also account for the larger CER for score-combination compared to that of majority vote.

Although a significant performance improvement in terms of CER was observed for mul-

timodal ASR systems when the SNR was small. The CER was increased with the decrease of the SNR. The main reason was that the judgment of the acoustic classifier or the acoustic signal was treated equally even when contaminated by noise under current combination scheme. One alternative is to adaptively adjust the weight of the acoustic signal when the SNR was small such that the acoustic signal has less domination for final decision. As a result, the MES and the GEMS signal, which are more resistant to noise, will gain more weight in making decision, and thus increases the chance of making correct decision.

5.7 Discussion

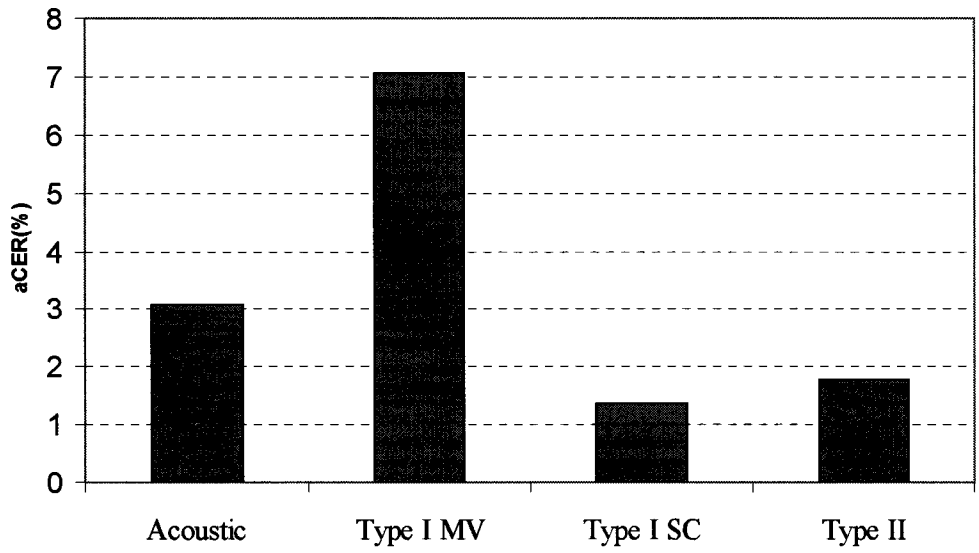


Figure 5-9 A comparison of the unimodal system and multimodal systems without adding AWGN

Two types of multimodal systems have been constructed in this chapter. The CER comparison among the acoustic unimodal ASR system and the multimodal systems without adding AWGN was plotted in Figure 5-9. The CER for the unimodal acoustic ASR system was 3.1%. With Type II score-combination, the CER dropped to 1.4%. The CER

dropped to 1.8% when Type II combination was used. The noticeable reduction in CER (except majority vote) showed that the multimodal ASR system could be constructed to enhance the unimodal acoustic ASR system by properly fusing the results of individual classifier (Type I) or by constructing a combining classifier (Type II).

An observation of the confusion matrices of the classifiers could be of assistance to understand the reason that accounts for the improvement. The confusion matrix for the MES HMM classifier and for the acoustic HMM classifier was shown in Table 5-1 and Table 5-2 respectively. As indicated in the tables, the acoustic HMM classifier had difficulty in recognizing the word '5': 9 out of 21 words spoken were misclassified; however, the MES HMM classifier was more efficient, with only 1 misclassification. On the other hand, the MES HMM classifier has difficulty recognizing the word '9', 7 misclassifications out of 30 words spoken, which could be recognized more accurately by the acoustic HMM classifier, only 1 misclassification out of 30 repetitions of the word '9'. Thus, it can be anticipated that with optimal combination methods, the multimodal ASR system would be able to utilize the capacities of individual classifier, reducing the CER.

The CER was 1.8% when using Type II combination, a rate that was comparable with the result by Type I score-combination. Since only a single classifier was be used, there was no need to train each classifier separately like in Type I combination, which may be very time consuming and computational intension. Furthermore, there was no any decision rule required, which may be somehow subjective. However, the single one structure also suffered certain limitations: For example, the classifier structure may not be optimal for all signals to be classified. In addition, the classifier considered all the input signals equally, ignoring the fact that some signals may convey more information for classifica-

tion.

Table 5-1 MES HMM classifier confusion matrix for subject 5

		HMM Classification									
		0	1	2	3	4	5	6	7	8	9
Word Spoken	0	23	2	1	0	0	0	3	0	0	0
	1	0	29	0	0	1	1	0	0	0	0
	2	1	0	26	1	1	1	0	0	0	0
	3	0	2	0	26	0	1	0	1	0	0
	4	0	0	0	0	29	1	0	0	0	0
	5	0	0	0	0	0	29	0	1	0	0
	6	0	0	0	0	0	0	27	0	1	2
	7	2	0	0	0	0	0	0	24	0	4
	8	0	1	0	0	0	0	3	0	25	1
	9	1	0	0	1	0	1	1	1	2	23

Table 5-2 Acoustic HMM classifier confusion matrix for subject 5

		HMM Classification									
		0	1	2	3	4	5	6	7	8	9
Word Spoken	0	26	0	1	0	1	1	0	0	0	0
	1	0	30	0	0	0	0	0	0	1	0
	2	1	0	28	0	0	1	0	0	0	0
	3	0	0	0	29	0	1	0	0	0	0
	4	2	0	0	1	27	0	0	0	0	0
	5	2	0	0	6	0	21	0	0	0	1
	6	0	0	0	0	0	0	30	0	0	0
	7	0	0	1	0	0	0	0	29	0	0
	8	0	0	0	0	0	0	0	0	30	0
	9	0	0	0	0	0	1	0	0	0	29

Despite of the computation complexity, Type I combination, on the other hand, offers the advantage that each classifier can be tuned for optimal performance. Also, different classifier structures even different types of classifiers can be used depending on the characteristics of the signals to be classified. In addition, the fusing can be adjusted according to

applications.

However, the multimodal system does not always produce superior performance. For example, the CER with majority vote was 7.1%. Although the CER was much less than the CER with the MES or the GEMS standalone classification (16% and 31.1% respectively), it was worse than the CER with acoustic standalone classification (3.1%). Two reasons may account for this deterioration. First, two wrong decisions made by the MES and the GEMS signal classifiers, when coincidently being unanimous, could avert the correct decision by the acoustic signal classifier. Again, these two classifiers do not have a high performance even under ideal conditions compared to the acoustic classifier. Second, when all classifiers failed to agree upon a decision, the decision of the acoustic signal classifier would be chosen based on the decision rule. On the other hand, the decisions from the MES and GEMS signal classifiers would be discarded, without making any use of the two additional information sources, which is desirable when the SNR is high but poor when the SNR is low. Despite its simplicity, the major vote treats each member classifier equally without considering the ability of individual classifier. In addition, the major vote does not utilize the probability information (e.g. score) resulting from individual classifier. Therefore, the score-combination method, taking score information into account, seems more appealing in this scenario.

It was arguable that the performance of the acoustic signal classifier was not optimal with a CER of around 3.1%. While constructing a superior acoustic signal classifier is possible with carefully selected features and types of classifiers, it is sufficient to illustrate the effectiveness of the combination approach which resulted in a multimodal ASR system. In addition, an improved acoustic signal classifier or other signal classifiers should simply

enhance the overall performance of the multimodal ASR system. Take the major vote for instance, since the decision was make by the acoustic classifier when there was no agreement among the three classifiers, an acoustic classifier, with higher recognition accuracies, could reduce the CER.

Performance of the proposed multimodal ASR systems had been evaluated using different levels of AWGN. Results showed that the multimodal ASR systems had consistent improvements over the unimodal acoustic ASR system under various levels of SNR. As expected, the CER increased dramatically when the samples were contaminated by AWGN. For example, at 15 dB SNR, the CER for the unimodal acoustic ASR was nearly 57%. The rate for the Type II multimodal ASR was 5%, an improvement of 91% was observed.

5.8 Conclusions

A multimodal ASR system using the GEMS, the MES and the acoustic signal has been proposed. With appropriate selection of classifier structure and combination methods, performance of the multimodal ASR system was demonstrated to be superior to a unimodal acoustic ASR system. It was also demonstrated through an experiment that, when applied to a noisy environment, where performance of the unimodal acoustic ASR system would deteriorate dramatically, the multimodal ASR system can be used to perform recognitions much more accurately.

Chapter 6 Conclusions and Future work

6.1 Conclusions

Conventional acoustic ASR methods have many limitations. Relying solely on the acoustic signal, a conventional acoustic ASR system will suffer a severe degradation in recognition performance when operating with background noise and the system will fail when no acoustic communication was allowed. A new methodology of ASR was proposed in this research to overcome such drawbacks. The MES and the GEMS signal were used as alternative information sources for ASR. Thanks to their immunity to the acoustic noise, a MES or GEMS signal ASR system could be used to improve performance in noisy environments. In addition, without relying on the acoustic signal, they could also be used in applications where acoustic speech is prohibited.

A new method of training the HMM classifier for the MES classification have been presented, the new method, using the AMMI training, estimates the HMM parameters by utilizing the misclassification information during the training stage to maximize the separability between patterns. The AMMI training reduced the classification error rate by 3% consistently over the widely used ML training, with an improvement of 6.74% observed at the empirically optimal operating point.

The GEMS signal at the neck has been proposed as a new modality for ASR. The RMS, the ZC and the SSC have been selected among 5 time domain features. A classification error rate of 31.1% was achieve for a 10-word vocabulary with a 10-state HMM classifier

using the extracted features. Similar to the MES, the GEMS signal is resistant to various acoustic noises. A GEMS signal recognizer can be used to perform ASR under severely background noise. If the classification accuracies could be improved further, it would make many for promising unimodal applications. For example, similar to the MES ASR, the GEMS signal ASR does not required the signal presents acoustically, which can be used to realize the silent communication or be used as voice prosthesis for patience who loses voice temporarily.

Two types of multimodal ASR systems were developed. A significant reduction in the classification error rate was observed by both proposed multimodal systems. Compared to the unimodal acoustic ASR system, the classification error rate of the unimodal acoustic system dropped from 3.1% to 1.4% when the classification of each individual classifier was fusing using the score-combination. The classification error rate dropped to 1.8% when a single HMM classifier was constructed to perform recognition, using the three types of signals as inputs.

Performance of the multimodal ASR systems was simulated under noisy environment by adding 11 levels of AWGN to the acoustic test samples. Results showed that the multimodal ASR systems had a consistent reduction in recognition error rate under various noise levels. Type II combination yielded the best overall performance among the three combination methods. It was not unexpected that the performance of the acoustic ASR degraded severely with the presence of AWGN. At 20 dB SNR, for example, the CER increased tremendously to 42.3%. By contrast, Type II combination yielded a CER of 3.6% at the same SNR level. Advantages of this multimodality, shown in the results of previous experiments, demonstrated the applicability of this new method for ASR.

6.2 Contributions

6.2.1 Major Contributions

1) Introduced the AMMI training in the context of MES ASR. This was demonstrated to improve the recognition accuracies of MES ASR, as the AMMI training uses the misclassification information in the training stage to adjust the HMM parameters adaptively. The reduction in the error rate was shown to be consistent over the ML training. This was published and presented at the 2006 International Conference of the IEEE Engineering in Medicine and Biology Society.
2) Implementing and testing of a new modality of ASR using the GEMS signal. This has been demonstrated of having the potential to be implemented as a unimodal ASR system or as a supplement to the conventional acoustic ASR system. Testing and evaluation of the GEMS ASR system was done using experimental data. While this idea has been previously proposed [34], this work represents the first implementation and confirmation of feasibility of ASR using the GEMS signal.
3) Developed a multimodal ASR system using the GEMS signal, the MES and the acoustic signal. A majority vote, score-combination, and signal multimodal classifier were implemented. These implementations were tested and evaluated using experimental data. The multimodal ASR system demonstrated superior performance than the unimodal acoustic ASR system if using a proper combination method. This performance increase was shown to be especially true under acoustically noisy conditions.

6.2.2 Minor Contributions

1) Observed the AMMI training is sensitive to the training samples size, which implies the AMMI training could yield a better performance when there is significant large number of samples available for training.
2) Optimized a set of time-domain features and a structure of the HMM for the MES which provides a lower recognition error rate.
3) Optimized a set of time-domain features and a structure of the HMM for the GEMS signal which provides a lower recognition error rate.
4) Proposed two types of combination methods to construct a multimodal ASR system, that is, a multimodal ASR system that fusing the classification of individual signal classifier and a multimodal ASR system that classifying three types of signals using one signal classifier.
5) Evaluated the multimodal ASR system with the majority vote and the score-combination, showed that the multimodal system does not necessary always improve the performance of the unimodal ASR system, the decision rule should be set up with caution.

6.3 Future Work

1) The availabilities of speech information in the GEMS signal have been proven in this study. Potentials of this new methodology for ASR using the GEMS signal in various real-world applications should be fully explored, including unimodal applications.
2) The GEMS signal was currently collected at the trachea area using a neck interface that firmly attached the antenna to the skin, which enable the GEMS

operate in the near field area. The GEMS signal from other locations should be studied. For example, the GEMS signal could be collected to detect the motion of the jaw, the chin or the mouth, which may convey the information for ASR. The major difficulty lies in positioning the antenna for the near field operating, producing the GEMS signal with better SNR.

3) Classification accuracies of the GEMS signal should be further improved. Performance of various types of classifiers on the GEMS signal classification should be evaluated.
4) Only time domain features were extracted and used for classifications in this study. The effects of frequency domain features on the recognition performance should be studied.
5) Techniques to reduce computational intensities should be applied. For example, the vector quantization could be performed on the extracted features to reduce the feature dimension.
6) More efficient instrumentation for the MES and the GEMS signal measurement should be developed, which could be implemented by designing wearable measurement equipment or using wireless sensors.
7) The possibilities of training and operating the classifiers in offline, real-time, continuous speech and user independent manner should be further explored.
8) Other possible combination methods to construct a multimodal ASR system should be investigated. For example, the decision rules could be trained adaptively to enhance the robustness of the multimodal ASR system or a series of sequentially arranged classifiers could be used to improve the recognition rate.

9) Synchronization among the three types of signals should be performed. This could be achieved by employing a synchronized signal segment scheme or by modifying the structure of the classifier.

10) Using contextual information is another method of improving classification accuracy and should be integrated into future work.

References

[1] S. Das, A. Nadas, D. Nahamoo M. Picheny, "Adaptation Techniques for Ambience and Microphone Compensation in the IBM Tangora Speech Recognition System," Proceedings of the IEEE CASSP, vol. 1, pp. 21-23, 1994.

[2] L. Barbier, C. Mokbel, G. Collet, "Trainable Noise Subtraction Filters for Speech Enhancement in Car," Proceedings of Fifth European Conference on Signal Processing, vol. 2, pp. 1111-1114, 1990.

[3] A. D. Berstein, I. D. Shallom, "A Hypothesized Wiener Filtering Approach to Noisy Speech Recognition," Proceedings of the IEEE CASSP, pp. 913-916, 1991.

[4] S. Furui, "On the Use of Hierarchical Spectral Dynamics in Speech Recognition," Proceedings of the IEEE CASSP, pp. 789-792, 1990.

[5] P. Alexandre, P. Lockwood, "Root Cepstral Analysis: A Unified View. Application to Speech Processing in Car Noise Environments," Speech Communication, vol. 12, no. 3, pp. 227-288, 1993.

[6] Y. M. Cheng, D. O'Shaughnessy, P. Kabal, "Speech Enhancement Using a Statistically Derived Filter Mapping," Proceedings of International Conference on Spoken Language Processing, vol. 1, pp. 515-518, 1992.

[7] Y. Ephraim, "A Bayesian Estimation Approach for Speech Enhancement Using Hidden Markov Models," Proceedings of the IEEE ASSP, vol. 40, no. 4, pp. 725-735, 1992

[8] Y. Anglade, D. Fohr, J. C. Junqua, "Speech Discrimination in Adverse Conditions Using Acoustic Knowledge and Selectively Trained Neural Networks," Proceedings of the IEEE ICASSP, vol. 2, pp. 279-282, 1993

[9] Y. Gong, "Speech Recognition in Noisy Environments: A Survey," Speech Communication, vol. 16, pp. 261-291, 1995.

[10] A. D. C. Chan, K. Englehart, B. Hudgins, D. F. Lovely, "Multi-expert Automatic Speech Recognition Using Acoustic and Myoelectric Signals," IEEE Transactions on Biomedical Engineering, vol. 53, no. 4, pp. 676-685, 2006.

[11] S. Hartzog, M. S. Morse, B. Trull, C. Alegre, P. Harris, "Recognition of Speech from Signals Secondary to Speech," Proceedings of the IEEE EMBS, pp. 1188-1189, 1988.

[12] N. Sugie, K. Tsunoda, "A Speech Prosthesis Employing a Speech Synthesizer-Vowel Discrimination from Perioral Muscle Activities and Vowel Production," IEEE Transactions on Biomedical Engineering, vol. BME-32, no. 7, pp. 485-490, 1985.

[13] A. D. C. Chan, K. Englehart, B. Hudgins, D. F. Lovely, "Myoelectric Signals to Augment Speech Recognition," Medical and Biological Engineering and Computing, vol. 39, no.4, pp. 500-504, 2001.

[14] A. D. C. Chan, K. Englehart, B. Hudgins, D. F. Lovely, "Hidden Markov Model Classification of Myoelectric Signals in Speech," IEEE Engineering in Medicine and Biology Magazine, vol. 21, no. 4, pp. 143-146, 2002.

[15] B. J. Betts, C. Jorgensen, "Small Vocabulary Recognition Using Surface

Electromyography in an Acoustically Harsh Environment," Technical Memorandum, NASA/TM-2005-213471, National Aeronautics and Space Administration, Ames Research Center, Moffett Field, California, November, 2005.

[16] C. Jorgensen, D. D. Lee, S. Agabon, "Sub Auditory Speech Recognition Based on EMG Signals," Proceedings of the IEEE IJCNN, vol. 4, pp. 3128 - 3133, 2003.

[17] H. Manabe, A. Hiraiwa, T. Sugimura, "Unvoiced Speech Recognition Using EMG – Mime Speech Recognition," Short Talk: Brains, Eyes and Ears, pp. 794 – 795, 2003.

[18] H. Manabe, M. Fukumoto, "Robust and Preceding Speech Detection Using EMG," Proceedings of the IEEE EMBS, pp. 5812 – 5815, 2005.

[19] N. Bu, T. Tsuji, J. Arita, M, Ohga, "Phoneme Classification for Speech Synthesiser Using Differential EMG Signals between Muscles," Proceedings of the IEEE EMBS, pp. 5962-5966, 2005.

[20] S. C. Jou, T. Schultz, M. Walliczek, F. Kraft, A. Waibel, "Towards Continuous Speech Recognition Using Surface Electromyography," ICSLP, pp. 573 - 576, 2006.

[21] L. Maier-Hein, F. Metze, T. Schultz, A. Waibel, "Session Independent Non-audible Speech Recognition Using Surface Electromyography," Proceedings of the IEEE ASRU, pp. 331 - 336, 2005.

[22] E. D. Petajan, "Automatic Lipreading to Enhance Speech Recognition," IEEE Global Telecommunications Conference, Atlanta, GA, pp. 265-272, 1984.

[23] A. J. Goldschen, "Continuous Automatic Speech Recognition by Lipreading," Ph.D. dissertation, George Washington University, Washington, DC, September, 1993.

[24] G. I. Chiou, J. N. Hwang, "Image Sequence Classification Using a Neural Network Based Active Contour Model and a Hidden Markov Model," Proceedings of the IEEE ICIP, vol. 3, pp. 926 – 930, 1994.

[25] K. Sugahara, M. Kishino, R. Konishi, "Personal Computer Based Real Time Lip Reading System," WCCC ICSP, vol. 2, pp. 1341 – 1346, 2000.

[26] U. Meier, W. Hurst, P. Duchnowski, "Adaptive Bimodal Sensor Fusion for Automatic Speechreading," Proceedings of the IEEE ICASSP, vol. 2, pp. 833 – 836, 1996.

[27] A. Sagheer, N. Tsuruta, R. I. Taniguchi, S. Maeda, "Visual Speech Features Representation for Automatic Lip-reading," Proceedings of the IEEE ICASSP, vol. 2, pp. ii/781 - ii/784, 2005.

[28] M. Knaflitz, G.Balestra, "Computer Analysis of the Myoelectric Signal," IEEE Micro. vol. 11, no. 5, pp.12 - 15, 48-58, 1991

[29] CJ De Luca, "Physiology and Mathematics of Myoelectric Signals," IEEE Transactions on Biomedical Engineering, vol. 26, no. 5, pp. 539-553, 1967.

[30] D. S. Dorcas, R. N. Scott, "A Three State Myoelectric Control", Medical and Biological Engineering, vol. 4, pp. 367-372, 1966.

[31] T. Lymaq A. Freedy, and M. Solomonow, "Studies toward a Practical Computer-

aided Arm Prosthesis System," Bulletin of Prosthetic Research, pp. 213-225, 1974.

[32] D. Graupe, J. Magnussen, A. Beex, "A Microprocessor System for Multifunctional Control of Upper-limb Prostheses via Myoelectric Signal Identification," IEEE Transactions on Automatic Control, vol. 23, no. 4, pp. 538-544, 1978.

[33] J. Z. Wang, R. C. Wang, F. Li, M. W. Jiang, D. W. Jin, "EMG Signal Classification for Myoelectric Teleoperating a Dexterous Robot Hand," Proceedings of the IEEE EMBS, pp. 5931-5933, 2005.

[34] J. F. Holzrichter, G. C. Burnett, L. C. Ng, W. A. Lea, "Speech Articulator Measurements Using Low Power EM Wave Sensor," Journal Acoustic Society America vol. 1, pp. 622, 1998.

[35] L. C. Ng, G. C. Burnett, J. F. Holzrichter, T. J. Gable, "Denoising of Human Speech Using Combined Acoustic and EM Sensor Signal Processing," Proceedings of the IEEE ICASSP, vol. 1, pp. 229 - 232, 2000.

[36] M. Zardoshti, B.C. Wheeler, K. Badie, R. Hashemi, "Evaluation of EMG Features for Movement Control of Prostheses," Proceedings of the IEEE EMBS, pp.1141-1142, 1993.

[37] D. Duda, P. Hart, "Pattern Classification and Scene Analysis," Wiley-Interscience, 1973

[38] H. Sakoe, S. Chiba, "A Dynamic Programming Approach to Continuous Speech Recognition," Proceedings of International Congress on Acoustics, Budapest, Hungary, Paper 2OC-13, 1971.

[39] L. Rabiner, "A tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989.

[40] A. Ben-Yishai, D. Burshtein, "A Discriminative Training Algorithm for Hidden Markov Models," IEEE Transactions on Speech and Audio Processing, vol. 12, no.3, pp. 204-217, 2004.

[41] J. F. Holzrichter, "New Ideas for Speech Recognition and Related Technologies," Lawrence Livermore National Laboratory Report, UCRLUR-120310.

[42] T. E. McEwan, U.S. Patent Nos. 5,345,471, 5,361,070, and 5,573,012, 1994.

[43] G. C. Burnett, "The Physiological Basis of Glottal Electromagnetic Micropower Sensors (GEMS) and their Use in Defining an Excitation Function for the Human Vocal Tract," Thesis UC Davis, ProQuest Digital Dissertations, Inc., Ann Arbor, Michigan, document #9925723, January 15, 1999.

[44] S. Gabriel, R. W. Lau, C. Gabriel, "The Dielectric Properties of Biological Tissues: III. Parametric Models for the Dielectric Spectrum of Tissues,'' Physics in Medicine and Biology, vol. 41, pp. 2271–2293, 1996.

[45] R. S. Kumaran, K. Narayanan, J. N. Gowdy, "Myoelectric Signals for Mutimodal Speech Recognition," INTERSPEECH, pp. 1189-1192, 2005.