Speech Synthesis from Surface Electromyogram Signals

LAM Yuet Ming

A Thesis Submitted in Partial Fulfillment
of the Requirements for the Degree of
Doctor of Philosophy
in
Computer Science and Engineering

August, 2006

To my parents, for their love and care.

Speech Synthesis from Surface Electromyogram Signals

Submitted by

Lam Yuet Ming

for the degree of Doctor of Philosophy
at the Chinese University of Hong Kong in August 2006

Abstract

Although speech is the most natural means for communication among humans, there are situations in which speech is impossible or inappropriate. Examples include people with vocal cord damage, underwater communications or in noisy environments. To address some of the limitations of speech communication, non-acoustic communication systems using surface electromyogram signals have been proposed. However, most of the proposed techniques focus on recognizing or classifying the SEMG signals into a limited set of words. This approach shares similarities with isolated word recognition systems in that periods of silence between words are mandatory and they have difficulties in recognizing untrained words and continuous speech.

A method for synthesizing speech from surface electromyogram (SEMG) signals in a frame-by-frame basis is presented. The input SEMG signals of spoken

words are blocked into frames from which SEMG features were extracted and classified into a number of phonetic classes by a neural network. A sequence of phonetic class labels is thus produced which was subsequently smoothed by applying an error correction technique. The speech waveform of a word is then constructed by concatenating the pre-recorded speech segments corresponding to the phonetic class labels. Experimental results show that the neural network can classify the SEMG features with 86.3% accuracy, this can be further improved to 96.4% by smoothing the phonetic class labels. Experimental evaluations based on the synthesis of eight words show that on average 92.9% of the words can be synthesized correctly. It is also demonstrated that the proposed frame-based feature extraction and conversion methodology can be applied to SEMG-based speech synthesis.

基於表面肌電信號的語音合成

作者　林粵明

香港中文大學　二零零六年八月

摘要

雖然語音是人類最自然的交流方法，但在有些情況下，使用語音是不可能或者不適合的。例如聲帶組織損壞的人，在噪音或者水下環境裡通信。針對語音通信限制的問題，許多學者提出了使用表面肌電信號的非聲學通信系統這種方法。但是，被提出的技術大多集中於把表面肌電信號分類到一組有限的詞匯裡。這種情形類似獨立單詞語音識別系統，每個單詞之間必須留有一段的停頓時間，他們在識別未受訓練的單詞和連續語音上有一定的困難。

本論文提出一種由表面肌電信號合成語音的方法。表面肌電信號先被分成小塊並提取特徵，然後使用一個神經網絡將之分類到一組語音類。由此產生一連串的語音標籤，然後本論文使用一個錯誤清除方法來修正神經網絡分類的錯誤。基於這些修正後的語音標籤，通過連接預先記錄的話音小塊而合成語音。實驗結果顯示神經網絡能準確的分類 86.3%的表面肌電信號特徵。通過錯誤清除方法準確率更被進一步提高到 96.4%。基於一組八個詞匯的合成實驗，結果顯示 92.9% 的單詞能被正確地合成。這顯示本論文提出的表面肌電信號小塊特徵提取和轉換的方法能夠適用於基於表面肌電信號的語音合成。

Acknowledgments

I would like to take this opportunity to offer my heartfelt acknowledgements to some people.

I would like to especially thank my supervisor, Dr. Philip Heng-Wai Leong, who was also my final year project and Master Degree supervisor. I deeply appreciated his guidance and encouragement in these years, this thesis would not have been possible without him.

Additionally, I would like to thank Dr. Man-Wai Mak and Dr. Tan Lee for their help and suggestion. They gave me many valuable comments for improving this work.

I thank my colleagues, Mr. Y.H. Cheung and Mr. K.H. Tsoi for bringing me a comfortable working atmosphere in these years.

I would like to thank my parents for their endless support. This thesis is dedicated to them.

Finally, I thank Sze and ChingChing. My life is more meaningful because of them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Speech is the most natural way of communication among humans. The speech production process involves the contraction of the lungs, the vibration of the vocal cords and the resonance of the air stream in the vocal tract. Unfortunately, there are situations in which communication through speech is impossible or inappropriate. For example, people suffering from the side effect of laryngectomy surgeries or vocal cord damage are not able to produce normal speech because the vocal cord vibration plays a vital role in the speech production process; and speech production can be problematic in some physical environments such as underwater. Moreover, speech communication can also be affected by a number of factors. For instance, background noise can degrade the quality of the produced speech, and results in poor intelligibility. The performance of conventional speech recognition systems can be degraded drastically in a noisy environment such as in restaurants, factories, or trains. In addition, communication through speech is undesirable in some situations such as when very high privacy is desirable. For example, in the military or some public places that require silence, e.g. in theaters or libraries.

To address some of these limitations, many solutions have been proposed. To help the people without vocal cords, using a keyboard as input interface, the typed text can be transformed into speech by conventional text-to-speech systems [AHK$^+$87];

another method is to use a prosthetic device to simulate the vibration of the vocal cord, e.g. an electrolarynx [GHK+04], which is a battery powered handheld device, that can transmit a humming sound to the throat or mouth. For such a device, training is required and the produced speech is robotic. Non-acoustic communication systems that recognizing speech without using acoustic signals have also been proposed. Rather than using acoustic signals to perform recognition, alternative information sources are employed. One example of an information sources is through visual images, where the speech recognition is done based on video images of lip-rounding [CH97].

Recently, there has been an increased interest in using surface electromyogram (SEMG) signals [KM96], which are measured muscle activities from the skin surface. SEMG signals have been used to perform speech recognition [CEHL01, MZ04, KKAB04], supplement conventional speech recognition systems [CEHL02a], and construct computer-human interfaces [JB05].

Although previously proposed SEMG-based speech recognition systems show the feasibility of recognizing speech based on SEMG signals, limitations exist in these systems. Most of the proposed methodologies focus on recognizing or classifying the SEMG signals into a limited set of words. This approach share similarities with isolated word recognition systems in that periods of silence between words are mandatory. These systems have difficulties in recognizing untrained words and in order to recognize a new word, the recognition model needs to be retrained. It also has difficulties to recognize continuous speech, and the recognition accuracy can be affected by the duration of the words.

1.2 Objectives

The extensibility and applications of the previously proposed SEMG-based speech recognition systems is limited because of the congenital deficiencies of the isolated word recognition approach. The main objective of this research work was to address

the limitations of the previous proposed SEMG-based speech recognition systems. The detailed research aims were:

- Explore a methodology for continuous speech synthesis from input SEMG signals.
- Explore the feasibility of unlimited vocabulary synthesis.

Even in conventional speech recognition, large vocabulary continuous speech recognition is still a challenging task. This problem is made even more difficult in SEMG-based speech recognition, since the information available in SEMG signals is not as rich as speech signals, and the collected SEMG is weak and noisy.

1.3 Contributions

The main contributions of this dissertation are:

- To select the suitable SEMG features, a detailed comparison was made between non-overlapping and overlapping frequency band features. The utility of overlapping frequency band features was demonstrated quantitatively for the first time.
- By utilizing knowledge concerning the medium-term stationarity of speech, an error correction technique was proposed to post-process the outputs of the neural network and enhance the classification accuracy. This significantly improves the quality of the synthesized speech.
- This is the first time a concatenative speech synthesis technique with overlap-and-add being applied to SEMG-based speech synthesis.
- The proposed approach differs from the conventional one in that training data

is obtained via simultaneous SEMG and audio recordings and, although training is done based on phonemes, in principle, arbitrary speech can be generated. The feasibility of synthesizing speech directly from SEMG signals using this approach was demonstrated.

1.4 Thesis Organization

Chapter 2 gives an introduction to the fundamentals of SEMG signals. The chapter begins with an introduction to the nature and recording methods of SEMG signals, and shows the commonly used SEMG features in ergonomics. A review of previous proposed SEMG-based speech recognition systems and their applications are also presented.

Chapter 3 gives an introduction to the SEMG-based speech recognition systems employing isolated word recognition approach. Two classification techniques: hidden Markov models and neural networks are described.

Chapter 4 introduces the speech synthesis techniques. It begins with an introduction to the speech production mechanism. Two kinds of speech synthesis methods are then described: the linear predictive coding vocoder and the concatenative method. The computation of the linear predictive coding coefficients is also introduced.

Chapter 5 describes the proposed SEMG-based speech synthesis methodology. It begins with a discussion of the limitations of previous work, the implementation details are then described. This is followed by a discussion of the design considerations.

The feature selection process conducted in this work is described in Chapter 6. This chapter begins with an introduction to a separability measure which is used to measure the quality of SEMG features. Two types of spectral features of SEMG signals are compared and the spectral features chosen in this work are presented.

Experimental results are presented in Chapter 7. This includes the classification

performance of the neural network, the effects of the positioning of SEMG sensors, the performance of the error correction technique, and the speech synthesis results.

A summary of this work and directions for future research are presented in the final chapter.

In Appendix A, the circuit diagram of the front end amplifier used for SEMG collection is shown. Appendices B and C describe the K-Means and vector quantization algorithms used in this work.

Chapter 2

SEMG signal processing

2.1 Introduction

Surface electromyogram (SEMG) signals are widely used to analyze muscular activities, including SEMG-based speech recognition. In this chapter, the nature and terminology associated with SEMG signals relevant to this thesis are introduced, including the cause, measurement, and analysis of SEMG signals.

This chapter is organized as follows. Section 2.2 introduces the physiological basis for SEMG signals. In Sections 2.3 and 2.4, acquisition and feature selection techniques of SEMG signals are described respectively. A review of previously proposed SEMG-based word recognition systems is given in Section 2.5. Finally, some applications of these systems are presented in Section 2.6 and a summary is given in Section 2.7.

2.2 Nature of electromyogram signals

The change in relative position of filaments arranged in the interior of a muscle results in muscle contraction and force production. This phenomenon is triggered by an electrical pulse known as an action potential that traverses along the muscle fiber. This action potential is induced by a potential difference between the interior of a muscle cell and the external space. This is also known as the membrane

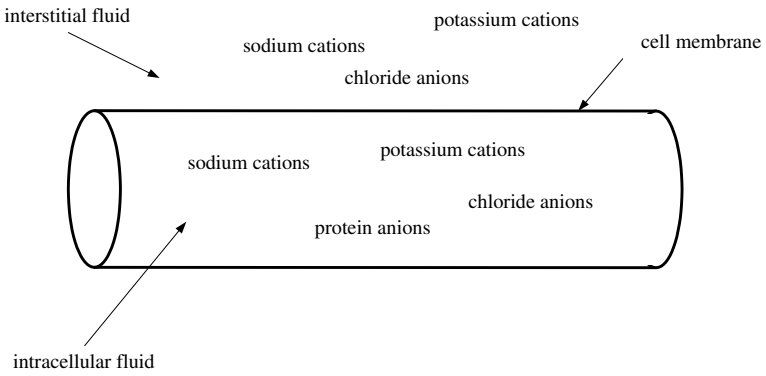


Figure 2.1: Muscle fiber and ions distributions.

potential. The recorded pattern of muscle action potentials is called an electromyogram (EMG) [KM96]. The EMG signal can be used to analyze muscle activities as there is a close functional relationship between them. Muscle activities, such as contraction speed and force, can be derived by analyzing the extracted features from the EMG signals in both temporal and spectral domains. For example, in ergonomics, researchers can understand the level of muscular strain by analyzing the EMG signal and apply this knowledge to reduce occupational fatigue [WJJ96].

Muscle fiber in the human body is surrounded by a cell membrane (Figure 2.1) which divides the intracellular fluid from the interstitial fluid. The distribution of ions in both compartments is uneven because the membrane proteins can transport ions from one side to the other side. Uneven distribution of ions results in a potential difference induced between the intracellular and extracellular space. Muscle contraction is triggered when the potential difference exceeds a certain value.

The normal range of this potential difference is between -60 to -90 mV. The negative sign means the intracellular space is negative compared with the extracellular space. Normally, this potential difference remains fairly constant, and is called the resting state. If the potential difference increases and reaches the so-called threshold value, an excitation state is triggered and muscle contraction is

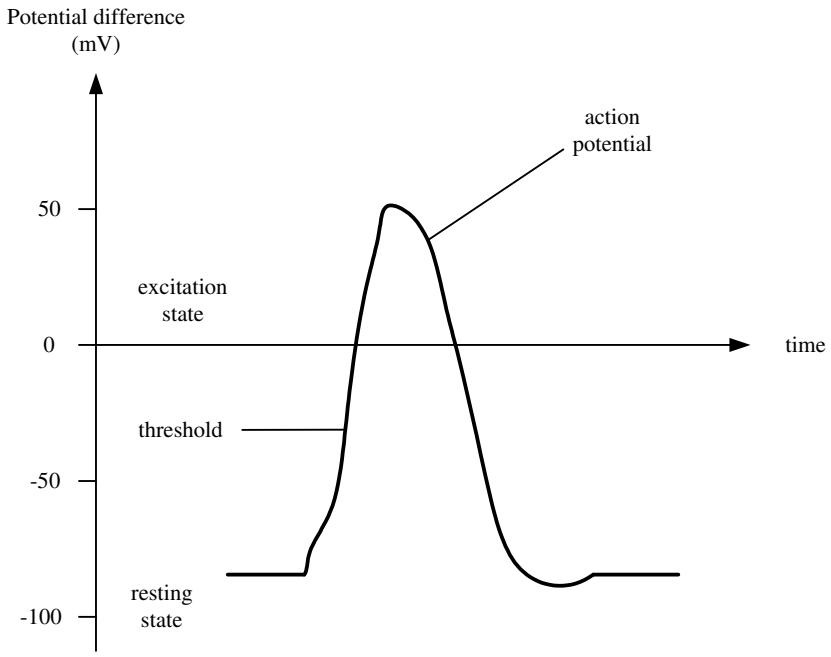


Figure 2.2: Resting and excitation potential of muscle fiber.

induced. This potential difference is called an action potential as the muscular contraction is triggered by the propagation of this potential along the muscle fiber. A figure showing the change in potential difference from resting to excitation and back is shown in Figure 2.2.

The propagation of the action potential and muscle fiber activation is shown in Figure 2.3. Muscle fiber is composed of different filaments arranged regularly. The action potential can trigger the sliding of myosin and actin filaments, where myosin filaments move towards actin filaments. In this figure, the action potential is propagated from the left to right, causing sliding of these filaments from the left to right and resulting in muscle contraction.

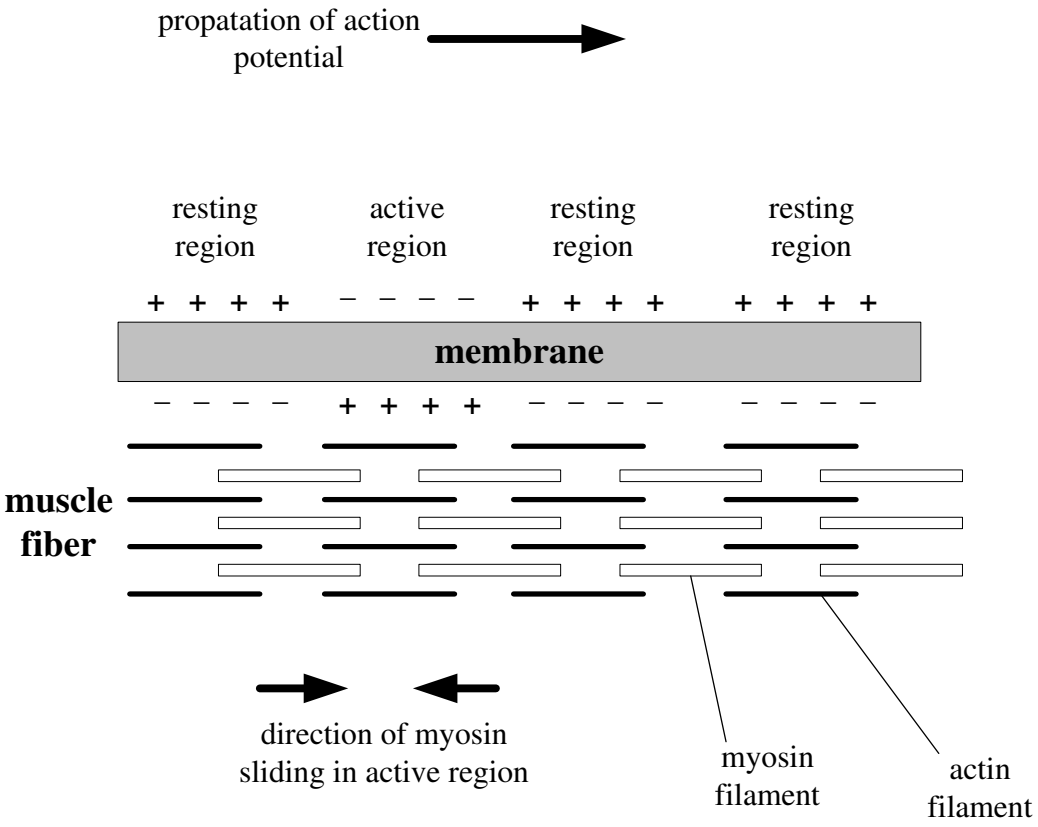


Figure 2.3: Muscle structure and the muscle fiber activation. Positive sign: higher potential, negative sign: lower potential.

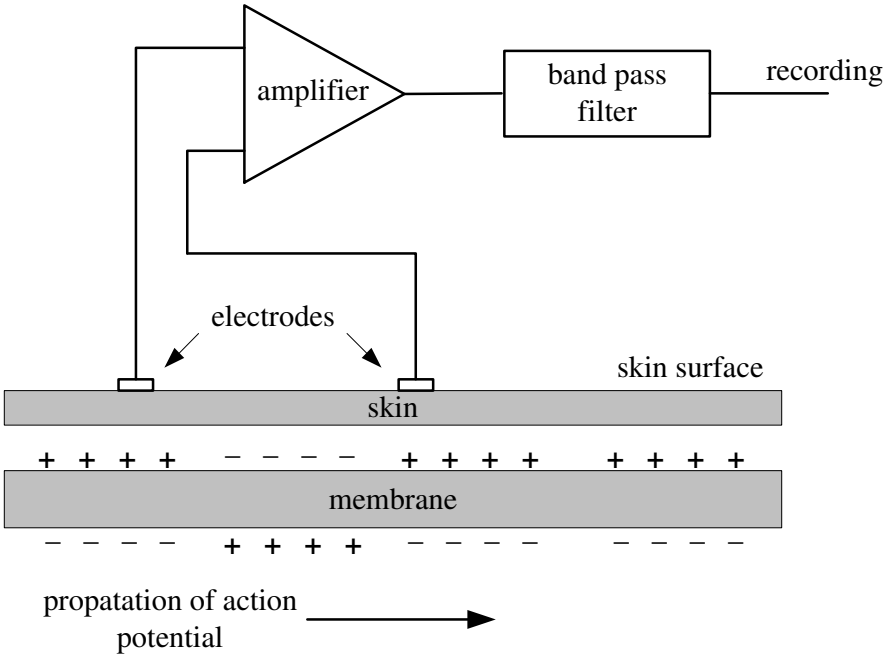


Figure 2.4: Surface electromyogram signal recording using differential amplification.

2.3 Recording of SEMG signals

To measure an electromyogram signal, a method that involves inserting a wire electrode into the muscle is often employed in clinical medicine. However, such an invasive method is not practical for ergonomic studies [KM96]. The surface electromyogram, records the muscle activity via electrodes on the skin surface and is widely used.

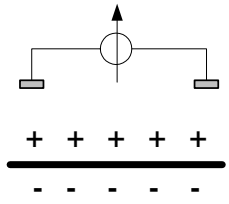
Figure 2.4 shows an SEMG recording. For simplicity, only one fiber is shown. A pair of electrodes are placed on the skin's surface to record the muscle activity. The potential difference at the amplifier's two inputs will be amplified, common mode noise being reduced by the differential configuration. Since the maximum peak-to-peak amplitude of the SEMG signal is 5 mV, the gain of the amplifier should be between 1000 and 2000.

The amplified signal is then passed through a $15 - 500$ band pass filter [KM96].

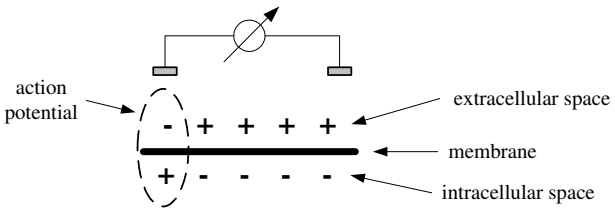
The high frequency cut-off corresponds to the maximum frequency of the SEMG signals. Moreover, because of the movement of electrodes and cables, low frequency components in the collected signal should be removed. This low frequency cut-off ranges from 0 to 15 Hz, the value depending on the condition of skin surface and quality of electrodes. A properly prepared subject, e.g. cleaned using alcohol, can reduce artifacts due to movement of electrodes and a smaller low frequency cut-off can thus be used. A suitable circuit that includes the amplifier and filter is given in Appendix A.

An action potential measured using this scheme is shown in Figure 2.5. In this figure, sub-figures (a)-(e) illustrate the simplified membrane structure and five instants during the propagation of the action potential. The amplitudes of the action potential measured corresponding to the five instants are indicated in the time curve in the lower section. The action potential measured from the two electrodes is zero during the unexcited state (sub-figure (a)). A potential difference is measured when an action potential reaches the left electrode (sub-figure (b)). The potential difference becomes zero once the action potential reaches the middle of the two electrodes (sub-figure (c)). A potential difference is measured again when the action potential progresses further to the right electrode (sub-figure (d)) but with reversed sign comparing when the action potential reaches the left electrode. Potentials at the two electrodes are the same when the action potential passes through and a zero potential difference is measured (sub-figure (e)). The maximum amplitude of the SEMG signals is proportional to the potential difference between the intracellular and extracellular space in the action potential, but the period of the time curve is inversely proportional to the propagation speed of the action potential. As the characteristics of the action potential, e.g. propagation speed of amplitude, is directly correlated to different muscle activities, the SEMG signal recorded can be used to analyze different muscle activities.

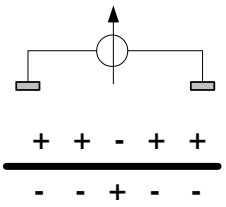
Figure 2.5 only shows the SEMG time curve of a single muscle fiber. Since the excitation of multiple muscle fibers may occur at different times, in real recordings,



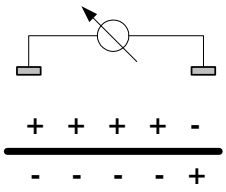
(a) Action potential at t1.



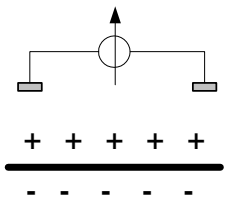
(b) Action potential at t2.



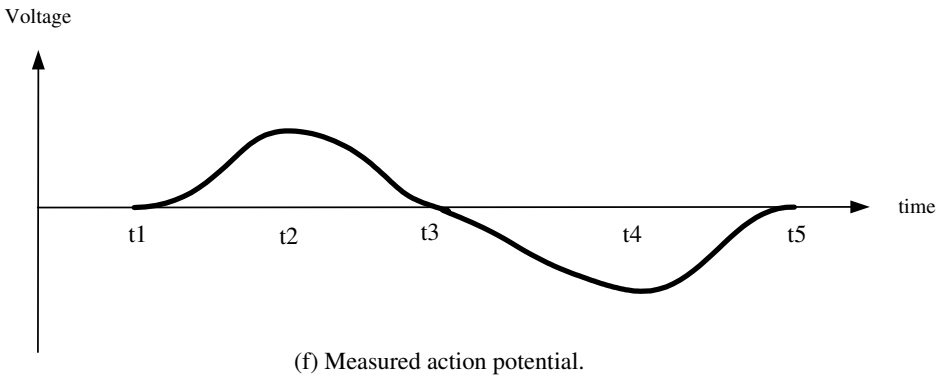
(c) Action potential at t3.



(d) Action potential at t4.



(e) Action potential at t5.



(f) Measured action potential.

Figure 2.5: (a) - (e) are the propagation of action potential at time t1 - t5. (f) is the measured SEMG signals during the propagation of action potential.

the SEMG signal recorded is the superposition of all muscle fibers.

2.4 Interpretation of SEMG signals

Although SEMG signals can be used to analyze muscle activities, the raw SEMG signal is too complex to analyze without considerable data reduction. SEMG analysis is often based on features derived from the original signals rather than the raw time domain signals themselves, and these features can be in both temporal and spectral domains.

2.4.1 Temporal feature extraction

In the temporal domain, amplitude information and zero-crossing rate (ZCR) are often used. These features are derived from the raw SEMG signal over a certain window to reduce the influence of artifacts, i.e. noise in the raw SEMG signal. Lippold e.g. [Lip67] maintained that using features extracted from a window of electromyogram signal, e.g. mean absolute amplitude, are more accurate than using a single amplitude at an instant for muscular contraction analysis.

Zero-crossing rate

ZCR counts the number of times that the raw SEMG waveform intersects the time axis (zero line) within a window:

$$\text{ZCR} = \sum_{i=1}^{N} z(i), \qquad \text{where } z(i) = \begin{cases} 1, & \text{if s(i)s(i+1)} \leq 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

In the above equation, the window size is N and $s(i)$ is the raw signals at time i. Zero-crossings are abundant at rest since the SEMG signals are weak and have a large amount of background white noise. Upon an impending motion, zero-crossings decrease because of the low frequency characteristic (below 500 Hz) of

SEMG signals. Another feature, the number of directional changes in raw SEMG signal per unit time, has similar characteristics to zero-crossing rate.

Amplitude

To extract the amplitude features, the raw SEMG signals are often full-wave rectified and then accumulated over a certain window. The reason for full-wave rectification is that the SEMG signals are quasi-random around zero and simple averaging results in a zero value. Mean absolute amplitude (MAA) and root mean square amplitude (RMSA) are two of the most commonly used amplitude features.

Mean absolute amplitude (MAA):

$$\text{MA} = \frac{1}{N}\sum_{i=1}^{N}|s(i)|. \tag{2.2}$$

Root mean square amplitude (RMSA):

$$\text{RMSA} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}s(i)^2}. \tag{2.3}$$

In the above equations, N is the window size and $s(i)$ is the raw SEMG signal at time i. RMSA is chosen in this work as it is more accurate than MAA for muscular activity analysis [FC86].

2.4.2 Spectral feature extraction

Frequency domain analysis

Any signal can be reconstructed from a series of sine waves having different amplitudes, phases, and periods, and a frequency representation describes a signal by these means. A discrete signal can be transformed into its frequency representation using the well known discrete Fourier transform (DFT):

$$F[k] = \sum_{n=0}^{N-1} x[n] e^{j\frac{-2\pi kn}{N}}, \qquad k = 0, 1, ..., N-1 \tag{2.4}$$

where $x[n]$ is the discrete signal and N is the length of the signal.

Time frequency domain analysis

Although the discrete Fourier transform has been of great value in many areas of engineering and science, a spectrogram representation is often used to represent both time and frequency domain characteristics. The spectrogram representation applies the discrete Fourier transform to a short-time window and moves this window along the time axis to capture the variation of the spectrum. This techniques has been successfully used to analyze biological phenomena [PWS96], and in particular speech [NQL83] [DN93]. One assumption made in applying the discrete Fourier transform on a limited window is that the signal is stationary over this window. The continuous formulation is as follows from the short-time Fourier transform (STFT) [Mit01]:

$$STFT[n, k] = \sum_{i=0}^{Q-1} x[n-p] w[p] e^{j\frac{-2\pi kp}{N}}, \qquad k = 0, 1, ..., N-1 \tag{2.5}$$

where Q is the window length and $w(p)$ is a window function, a commonly used method being the Hamming window:

$$w[p] = 0.54 - 0.46\cos[\frac{2\pi p}{Q}]; \quad where \quad 0 \leq p \leq Q-1. \tag{2.6}$$

The purpose of performing the Hamming window is to minimize the discontinuities at the border of each windowed segment [RJ93], the Fourier transform is thus performed on the windowed segment.

Applying the discrete Fourier transform on a limited window can capture the variation of spectrum along the time axis, however, it suffers from a window size selection problem. If a small window size is used, better time resolution can be

obtained, but this results in poor frequency resolution. On the other hand, a larger window size can improve the frequency resolution, but results in a loss of information between adjacent windows, leading to poor time resolution [Wil98].

2.5 Review of SEMG-based speech recognition

The research on SEMG-based non-acoustic speech recognition is still at a preliminary stage, compared with work on conventional speech recognition. Most of the proposed SEMG-based speech recognition systems were targeted towards recognizing isolated phonemes or words, the recognition models were built for recognizing the whole phonemes or words.

2.5.1 Isolated phoneme recognition

Recently, Jorgensen et al. reported their work on isolated phoneme recognition using two SEMG channels recorded from the chin [JB05]. A phoneme set contains twenty-three consonants and eighteen vowels were used, a 33% recognition rate was achieved. Their work also showed that SEMG signals may be inadequate for recognizing alveolars, where the tip of the tongue touches the alveolar ridge. By removing the six alveolars (*/t/, /d/, /s/, /z/, /ch/, /j/*) from the phoneme set, a recognition rate of 50% was obtained. They estimated the performance can be further improved by excluding */n/, /l/, and /r/*. The authors suggested two future working aspects to improve the performance. One is analyzing the effects of sensor positioning to detect the problematic features, the other is applying context-sensitive techniques used in conventional speech recognition.

By reducing the phoneme size, the recognition rate can be significant improved. A five-vowel SEMG-based speech recognition system was presented in [KKAB04]. Three facial muscles, mentalis, depressor anguli oris and massetter, were involved in the experiment. The SEMG signals for the five English vowels, */a/, /e/, /i/, /o/,*

/u/, were recorded in an isolated manner. An average classification rate of 88% was obtained using a neural network.

Similar work was reported in [MHS03], they achieved an average recognition rate of 94.7% for a phoneme set of five Japanese vowels by using a neural network as the classifier.

2.5.2 Isolated word recognition

Morse and O'Brien [MO86] studied the availability of information within the SEMG signals which is related speech sound. SEMG signals were collected from three positions near the neck and one position over the temporoparietal muscle of the head. Average amplitude was chosen as SEMG features and classified using a maximum likelihood algorithm. For a subject dependent test, their work reported a 97% recognition accuracy on a two-word set. The recognition accuracy deteriorated for larger word sets, being less than 70% for a six-word set and 35% for a seventeen-word set. Their experimental results also showed that the performance was improved when more SEMG channels were used. The authors also conducted an experiment to investigate the correlation between the data width and the recognition accuracy. In the experiment, words were classified using the features extracted from a portion of a word's SEMG signal, e.g. in Figure 2.6, instead of using the whole SEMG signal (W), recognition was done using portion of the word's SEMG signal ($W1$). The results showed that using larger portions can achieve better accuracy.

Manabe and Zhang [MZ04] employed conventional acoustic speech recognition techniques to an SEMG-based ten-Japanese-digit speech recognition system. The ten digits were $0, 1, 2, ..., 9$. Three SEMG channels, the cheek, the chin and the upper lip, were used in this work. Isolated SEMG signals were recorded when each digit was mimed silently. Different SEMG features were analyzed including filter band coefficients, Mel frequency cepstral coefficients, and linear predictive coefficients. Their experimental results showed that filter band coefficients are the

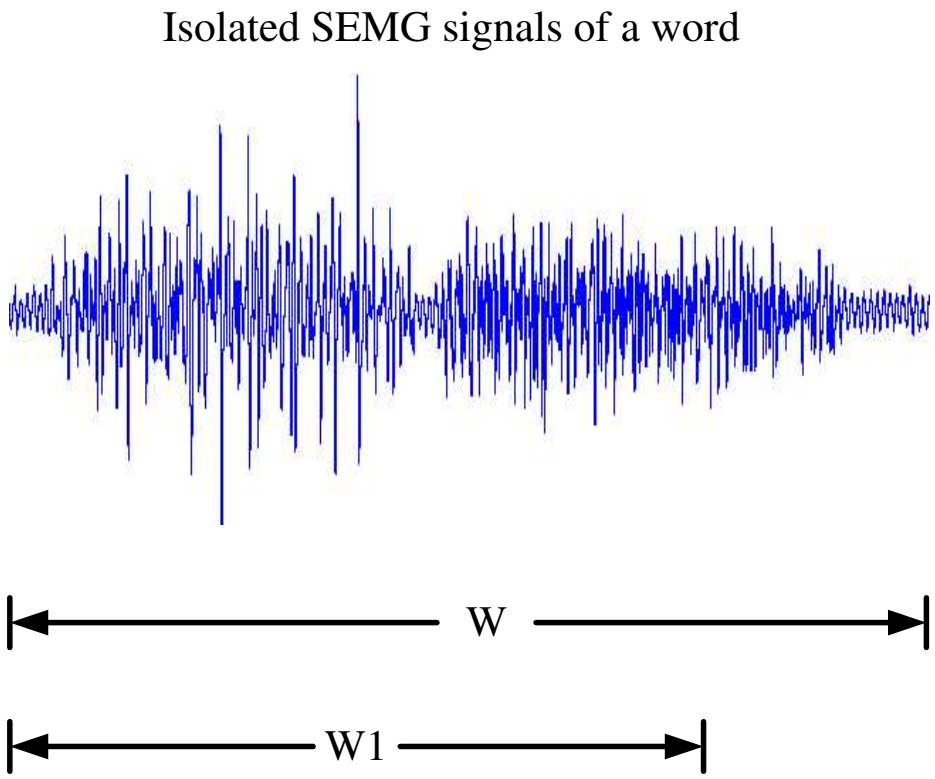


Figure 2.6: Using a portion of a word's SEMG signal to perform recognition.

best and a recognition rate of 63.7% was achieved. In this work, a multi-stream hidden Markov model was used as the classifier, and each SEMG channel was assigned with a weight which determined the contribution of each SEMG channel to the classification. The authors showed that recognition accuracy can be improved by optimizing the channel weights, a 4.0% improvement was obtained on average. This experimental result showed that optimizing the SEMG positioning was critical to improve the recognition accuracy. Jorgensen et al. also presented a SEMG-based six-word (*stop, go, left, right, alpha, omega*) recognizer with a recognition rate of over 90% [JLA03].

2.5.3 Frame-based phoneme recognition

Sugie and Tsunoda proposed to recognize five Japanese vowels (*/a/, /e/, /i/, /o/, and /u/*) using a frame-based approach [ST85]. The SEMG signals were collected from three positions of the face, blocked into frames and each frame was classified into one of the five vowels using a finite automaton. A recognition rate of 64% was achieved. In this work, the active/inactive status of each channel was used as SEMG features. For each SEMG frame, the number of crossings of a threshold level was counted, if the sum exceeded a certain threshold, the channel was regarded as active at that frame, otherwise, the channel was regarded as inactive. Each channel was assigned either a '1' or '0' for an active or inactive state. As a result of this coding, there were only eight possible outputs as there were only three SEMG channels. This may explain why the recognition rate was quite low.

2.6 Applications of SEMG-based speech recognition

2.6.1 Using SEMG to augment conventional speech recognition systems

As the recognition accuracy of conventional speech recognition systems are significantly degraded by a noisy environment, Chan et al. [CEHL02a] proposed to supplement a conventional speech recognition system by using the SEMG signals as a secondary information source. They proposed a scheme that integrated a conventional speech recognizer and an SEMG-based speech recognizer, concurrently recorded SEMG signals and speech signals being used for recognition. Their system was tested under various environments with different noise levels and the experimental results showed that the recognition capability of the SEMG-based recognizer was immune to noise, while the recognition accuracy of the conventional speech recognizer was significantly degraded at increased noise levels. Experimental results also showed that the recognition accuracy of the integrated system was higher than using either individual recognizer. As there are muscle activities prior to the acoustic signals, researchers suggested including certain durations of SEMG signals prior to the acoustic signals to perform recognition by using the acoustic signals as a trigger [CEHL01, CEHL02b]. The experimental results showed that including 500 ms of SEMG signals prior to the acoustic signals achieves the best recognition rate. An average classification rate of 83% was obtained using the hidden Markov model to classify SEMG signals of ten English digits (*zero, one, two, ..., nine*).

2.6.2 Human-computer interface

Jorgensen and Binsted demonstrated the feasibility of applying an SEMG-based isolated-word-recognition system to construct a human-computer interface [JB05]. In their work, a SEMG-based word recognizer that can recognize ten English digits

and six words (*stop, go, left, right, alpha, omega*) were presented, a recognition rate of 73.13% was achieved. The system was then applied to control a web browser. Instead of using a keyboard to input hyperlinks, in their system, an alphabet was constructed from the ten English digits and used to generate inputs. Simple commands were indicated by the six words.

2.7 Summary

This chapter began with an introduction to the physiological nature of the SEMG signals. The SEMG signals are induced by an action potential that propagates along muscle fibers causing muscle contraction and can be recorded via a high gain differential amplifier. The signals can be analyzed in temporal and spectral domains. Examples of SEMG features include zero-crossing rate, amplitude information, and the short-time Fourier transform. Some SEMG-based speech recognition systems and their applications were reviewed. Most of the proposed systems employed techniques of isolated word recognition in conventional speech recognition. In the next chapter, an introduction to the SEMG-based speech recognition will be given.

Chapter 3

SEMG-based speech recognition

3.1 Introduction

In the previous chapter, feature extraction from SEMG signals was described. Features commonly used include zero-crossing, amplitude, and frequency spectrum. Based on the extracted features from the SEMG signal recorded from facial muscles, researchers have developed algorithms to recognize speech. From the reviews, it is found that most of these SEMG based systems employ an isolated word recognition approach, in that features are mapped into a limited set of words, and neural networks and hidden Markov models are two of the most widely pattern recognition techniques. In this chapter, an overview of SEMG-based speech recognition systems will be given.

This chapter is organized as follows. A brief introduction to hidden Markov models and the application of hidden Markov models to SEMG-based speech recognition is given in Section 3.2. Section 3.3 describes artificial neural networks, the use of neural networks to SEMG-based speech recognition, and the advantages of neural networks. A summary is given in the last section.

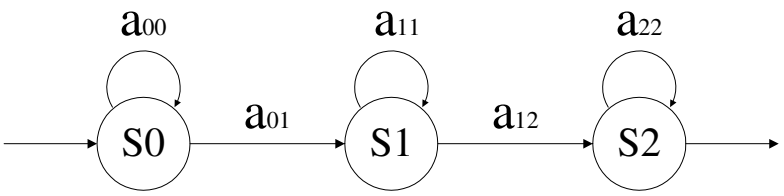


Figure 3.1: Left-to-Right HMM

3.2 Hidden Markov models approach

3.2.1 Hidden Markov models

Figure 3.1 shows a basic three-state left-to-right hidden Markov model (HMM) [RJ86] [Rab89], which is a probabilistic finite state machine (FSM) with a set of state transition and observation probabilities. The state transition probability is the probability of a state transition from one state to another, and the observation probability is the probability that a state emit a particular observation. A HMM calculates a likelihood score for an input observation sequence. In Figure 3.1, $S0, S1, S2$ are the states and a_{ij} is the probability of state transition from i to j. Figure 3.2 is the trellis representation which shows all possible state transition paths.

Given an observation sequence $O = (o_1, o_2, ..., o_T)$, HMM decoding calculates $P(O|\lambda)$, which is the probability of the input observation sequence for a given model λ. The result is the probability that the utterance represented by model λ will produce the observation sequence O.

$P(O|\lambda)$ of course can be calculated by enumerating all possible paths in the trellis diagram (see Figure 3.2) over the entire observation sequence. In total, there are N^T possible paths, where N is the number of HMM states and T is the length of the observation sequence. Assuming $q = (q_1, q_2, q_3, ..., q_T)$ is one of the state

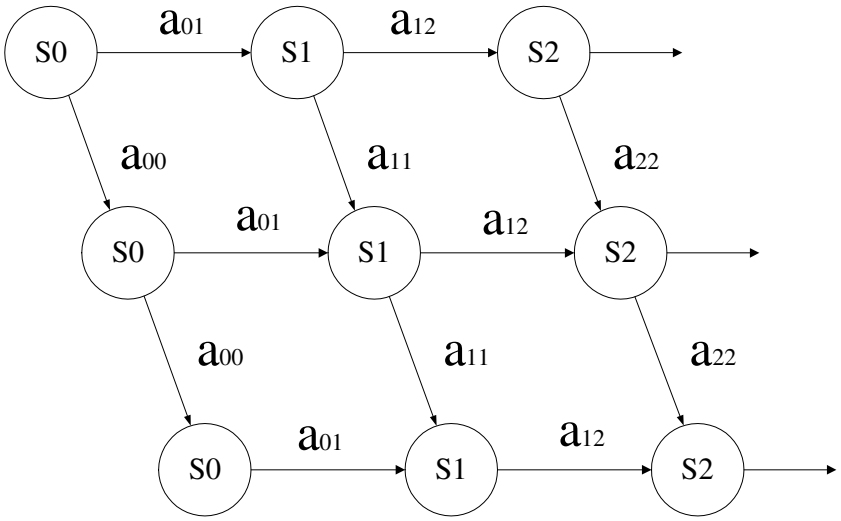


Figure 3.2: Trellis representation of Left-to-Right HMM

traversal paths, the probability can be calculated as follows:

$$\begin{aligned} P(O|\lambda) &= \sum_{all\ q} P(O|q,\lambda)P(q|\lambda) \\ &= \sum_{all\ q} P(O,q|\lambda) \end{aligned} \tag{3.1}$$

In practice, an alternative approach, called the Viterbi algorithm associated with a max function is used [RJ93]. $P(O|\lambda)$ is approximated by the maximum $P(O,q|\lambda)$, which generates a best state sequence q_{best}.

The iterative process to calculate the score for an observation sequence $O = (o_1, o_2, ..., o_T)$ is shown below. Assuming the observation probability for an input symbol o_t at state j is $p_j(o_t)$, the score along the best state sequence at time t that ends in state i is $h_t(i)$, the number of HMM states is N.

1. Recursion for each element o_t in the observation sequence:

$$h_t(j) = \max_{1 \le i \le N} [h_{t-1}(i)a_{ij}]p_j(o_t) \tag{3.2}$$

2. Termination:

$$H = \max_{1 \leq i \leq N} [h_T(i)] \tag{3.3}$$

where H is the probability of the best state sequence.

3.2.2 SEMG-based speech recognition using HMMs

Figure 3.3 shows an SEMG based isolated word recognition system using a hidden Markov model. The collected SEMG signals for words are blocked into frames, features are extracted from each frame and concatenated to form feature vectors. As a result, a sequence of feature vectors can be produced for each isolated word. In some approaches, these feature vectors are vector quantized before being presented to a HMM decoder. As shown in this figure, a separate HMM is used for each word in the target set and scores for the feature vector sequence for all HMMs are calculated. The word with the maximum score is chosen as the output word.

3.3 Neural network approach

Neural networks, also called artificial neural networks, are an architecture for computing inspired by our knowledge of biological neural networks. The computation of neural networks model the information processing flow in neural systems and are commonly used for pattern recognition problem [GM88, KL90, BAH05, Cho97, Tay96]. The type of neural network used in this work is multilayer perceptrons [MP69].

3.3.1 Architecture

A neural network contains an array of processing elements (neurons) that linked by connections (synapses). Each synapse is assigned a weight to represent the strength of connection between the processing elements. The functionality of the neural network is encoded in the strengths of the connections.

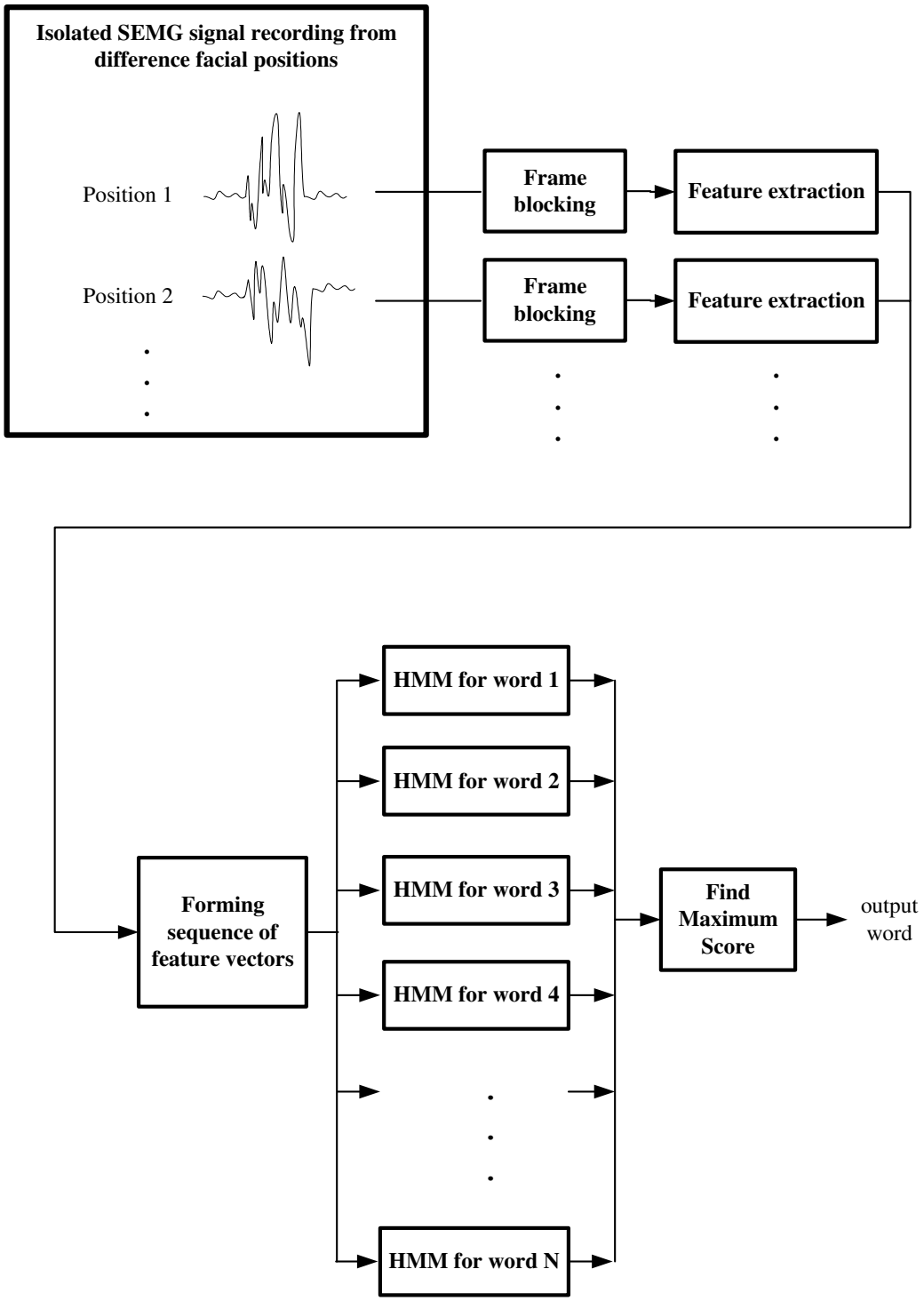


Figure 3.3: Hidden Markov model based SEMG word recognition.

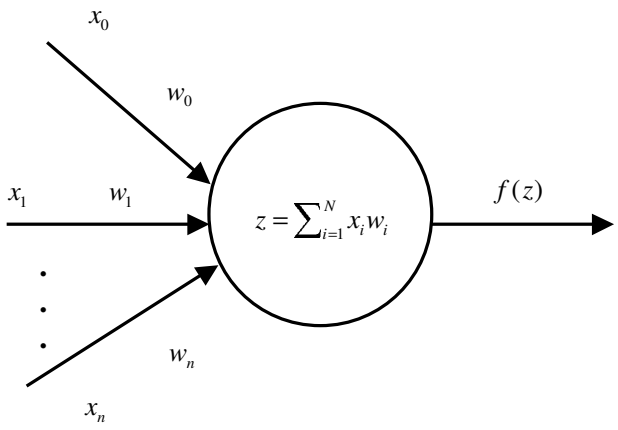


Figure 3.4: A neuron.

Figure 3.4 shows the basic processing element: a neuron. The arrows indicate the inputs and output, x denotes the output of previous neurons and w denotes the weighting of each synapse. A neuron computes the weighted sum of all its inputs to form an activation parameter z:

$$z = \sum_{i=1}^{N} x_i w_i, \tag{3.4}$$

where N is the number of inputs. The output of the neuron is defined by a transfer function $f(z)$. Some commonly used functions are:

Sign:

$$f(z) = \begin{cases} 1, & \text{if z} \geq 0 \\ -1, & \text{if z} < 0. \end{cases} \tag{3.5}$$

Sigmoidal:

$$f(z) = \frac{1.0}{1.0 + e^{-z}}. \tag{3.6}$$

Tan-Sigmoidal:

$$f(z) = tanh(z). \tag{3.7}$$

Log-Sigmoidal:

$$f(z) = log(z). \tag{3.8}$$

A three-layer neural network architecture is shown in Figure 3.5, in which neurons are arranged into three layers: input layer, hidden layer, and output layer. The network processes an input pattern in a strictly feed forward manner, i.e. an input pattern is presented to the input neurons and signal pass through the hidden layer and finally reach the output layer. Neurons in the hidden layer and output layer have multiple inputs and a single output. However, neurons in the input layer are different. They have a single input and single output with an identity transfer function $f(z) = z$. The outputs of the neurons in the hidden layer and output layer are:

$$z_j(l) = \sum_{i=1}^{N_{l-1}} x_i(l-1)w_{ij}(l) \tag{3.9}$$

$$x_j(l) = f(z_j(l)) \tag{3.10}$$

where l denotes the l-th layer ($2 \leq l \leq L$), L is the total number of layers ($L = 3$ in Figure 3.5), N_{l-1} is the number of neurons at the $(l-1)$-th layer, i denotes the i-th neurons in the $(l-1)$-th layer, j is the j-th neuron in the l-th layer, $w_{ij}(l)$ denotes the weight of the connection between the i-th neuron in the $(l-1)$-th layer to the j-th neuron in the l-th layer.

Because of the nonlinearities of neurons and the weights of connections, a large number of functions can be approximated by a multilayer neural network given sufficient number of hidden neurons.

3.3.2 Training

A neural network is trained by supplying a series of input patterns with corresponding responses (targets). The weights between neurons are adjusted according to the level of success in reproducing the targets, in other words, distances between the produced outputs and the targets.

The training of a neural network involves finding the optimal weights between neurons that minimize an error function:

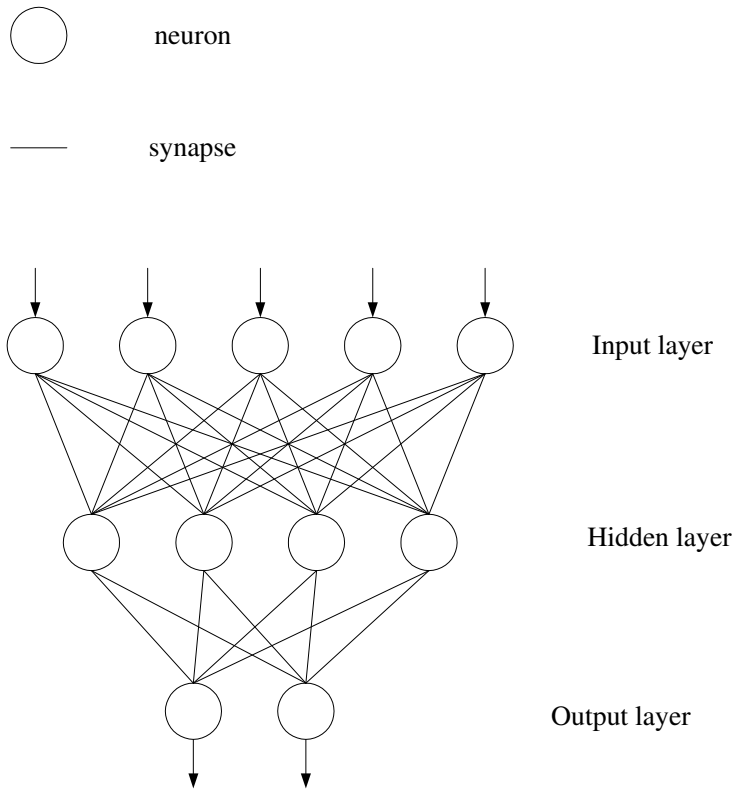


Figure 3.5: A three layer neural network architecture with five input nodes, four hidden nodes, and two output nodes.

$$e(w) = \frac{\sum_{k=1}^{P} \sum_{j=1}^{N} \left(t_j^{(k)} - {z_j}^{(k)}\right)^2}{N \times P}, \tag{3.11}$$

where N is is the number of neurons at the output layer, P is the number of training vectors, ${t_j}^{(k)}$ and ${z_j}^{(k)}$ are the target and neural network output at the j-th neuron for the k-th training vector respectively, w represents all the weights in the network.

A commonly used learning method is backpropagation algorithm [RHW86], which is a gradient descent based learning method, and the weights are updated according to the following formula:

$$\Delta w_{ij}(l) = -\eta \frac{\partial e(w)}{\partial w_{ij}(l)} \tag{3.12}$$

$$w_{ij}(l) = w_{ij}(l) + \Delta w_{ij}(l) \tag{3.13}$$

where η is a constant represents the learning rate.

3.3.3 SEMG-based speech recognition using neural networks

Figure 3.6 shows an SEMG based isolated word recognition system using a neural network. The number of output nodes equals to the number of words and each output node represents a word. During training, an output node is set to 1 when the target is the word represented by the node and, -1 or 0 otherwise. To recognize a word, isolated SEMG signals for a word are first recorded from different facial positions. Features are then extracted from the SEMG signals and input to the neural network in parallel. The values of the output nodes represent the scores of targets, in other words, the weighting that an input feature vector belongs to each target. By finding the maximum weighting, the most likely target word can be decided.

3.3.4 Features of neural networks

In this work, a three-layer (one hidden layer) neural network is used to map the features from SEMG domain to speech domain because its ability to model non-linear functions [HSW89] while making minimal assumptions about the statistical properties of the signals.

Noise tolerance

The recorded SEMG signals are noisy, e.g. the electrical noise picked up by human and transmission lines, noise in the filtering and amplification circuit board, and the noise due to human artifacts such as sudden movement of electrodes. Although the noise can be reduced by carefully design the experimental setup, it cannot be

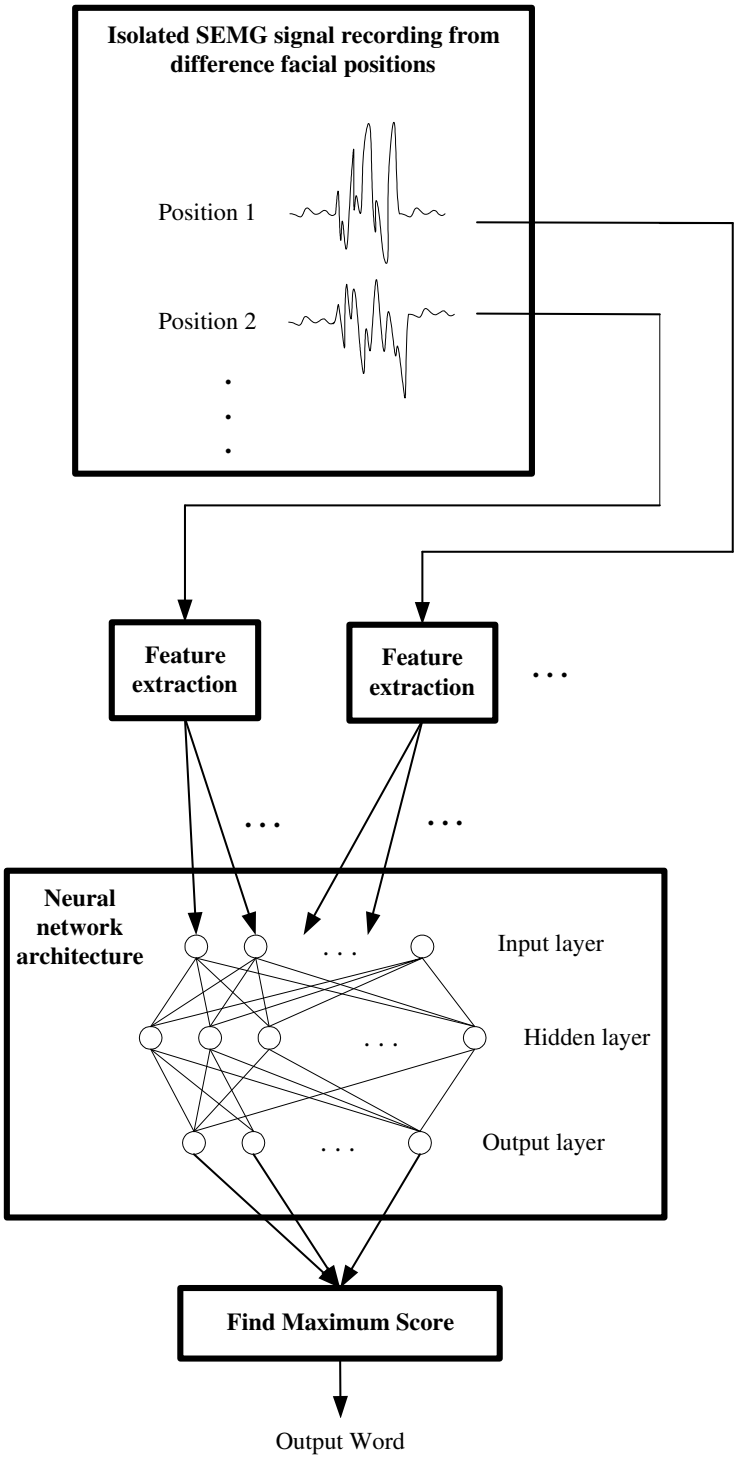


Figure 3.6: Neural network based SEMG word recognition.

removed totally. Thus, the noise is unavoidable and can be amplified together with the SEMG signal. Neural networks are capable of classifying noisy data [MT00], this capability making them a good choice for the analysis of SEMG signals.

Nonlinearity

As introduced in previous chapter, the SEMG signals are the superposition of action potentials of many muscle fibers and is by nature quasi-random and largely aperiodic [FC86]. This leads to the difficulties in feature selection during SEMG signal analysis, as the correlation between the selected features and physiological phenomena is unclear. The nonlinearity nature of neural networks thus makes it ideally suited for SEMG signal analysis. Gevins et. al. [GM88] show that neural networks are useful in analyzing signals with unknown characteristics and without prior assumptions about the statistical properties of the signals. Other than SEMG signals [HPS93], neural networks are also being used to analyze ECG (electrocardiogram) signal and identify cardiovascular diseases with very high accuracy [LJ91].

Neural network in speech processing

Besides the applications of neural networks in SEMG signal analysis, it is also being widely used in speech processing. Previous work has shown that neural networks yield high accuracy in conventional speech recognition [WC93] [CHS$^+$98] [KHJC04] and in fact the standard technique for this application.

3.4 Summary

In this chapter, brief introductions about neural networks and hidden Markov models and their applications to SEMG-based speech recognition were given. Neural networks are a computing architecture inspired by the interconnected neurons of the brain which can model nonlinear functions. In this work, a three-layer neural

network is chosen because of its powerful capability of nonlinear function approximation and outstanding features such as nonlinearity and noise tolerance. The functionality of the neural network used in this work is mapping the features from SEMG domain to speech domain, and the converted speech feature are then used to synthesis the speech waveform. In the next chapter, various speech synthesis techniques will be introduced.

Chapter 4

Speech synthesis

4.1 Introduction

Speech synthesis techniques, used to reconstruct speech waveforms in this work, are described in this chapter. Two major synthesis methods are introduced: the linear predictive coding (LPC) vocoder and the concatenative method. A brief introduction to the human speech production mechanism is given as it is the basis of LPC vocoder. As LPC coefficients are the widely used speech features, which is also used in this work, the method of computing LPC coefficients (called LPC analysis) is also described.

This chapter is organized as follows. A brief introduction to the human speech production mechanism is given in Section 4.2. The LPC vocoder and the computation of LPC coefficients are then described in Section 4.3. This is followed by an introduction to the concatenative synthesis method in Section 4.4. A summary is given in the last section.

4.2 Speech production

This section presents an overview of the human speech production mechanism, descriptions of speech production and LPC model are detailed in [RJ93]. Figure 4.1

illustrates a cross-section of the human speech production system. The gross components of the system are the lungs, vocal cords, nose and various parts of the mouth. Usually, the pharyngeal and oral cavities are called the vocal tract [HH01]. The speech production process involves the following processes:

- Air entering the lungs via normal breathing.
- Contraction of lungs to produce an air stream.
- Vibration of the air stream at the vocal cord.
- Resonance of the air stream at the vocal tract. By opening the velum, the air stream can also be resonated at the nasal cavity.

Various sounds are produced by different vibration frequencies of the vocal cords and resonance frequencies of the vocal tract, where the vibration frequency is controlled by the tension of vocal cords, and the resonance frequency is controlled by the shape of vocal tract, e.g. lip rounding and position of the tongue [Bre92].

4.2.1 Vibration of vocal cords

Figure 4.2 shows a superior view of the vocal cords. When the vocal folds are tense, the two vocal folds are held close together and the glottis is closed. The air stream from the lungs is obstructed and there is no air flow in the vocal tract. However, as the pressure keeps increasing and overcomes the resistance of the vocal folds, they are moved apart and the glottis opens. A rapid air stream then pass through the glottis and causes the pressure on the vocal folds to be decreased. The tension on the vocal folds makes them fall back into place rapidly and the glottis is closed again. This open-close process is repeated and the pitch of human sound is closely correlated to the open-close frequency of the glottis.

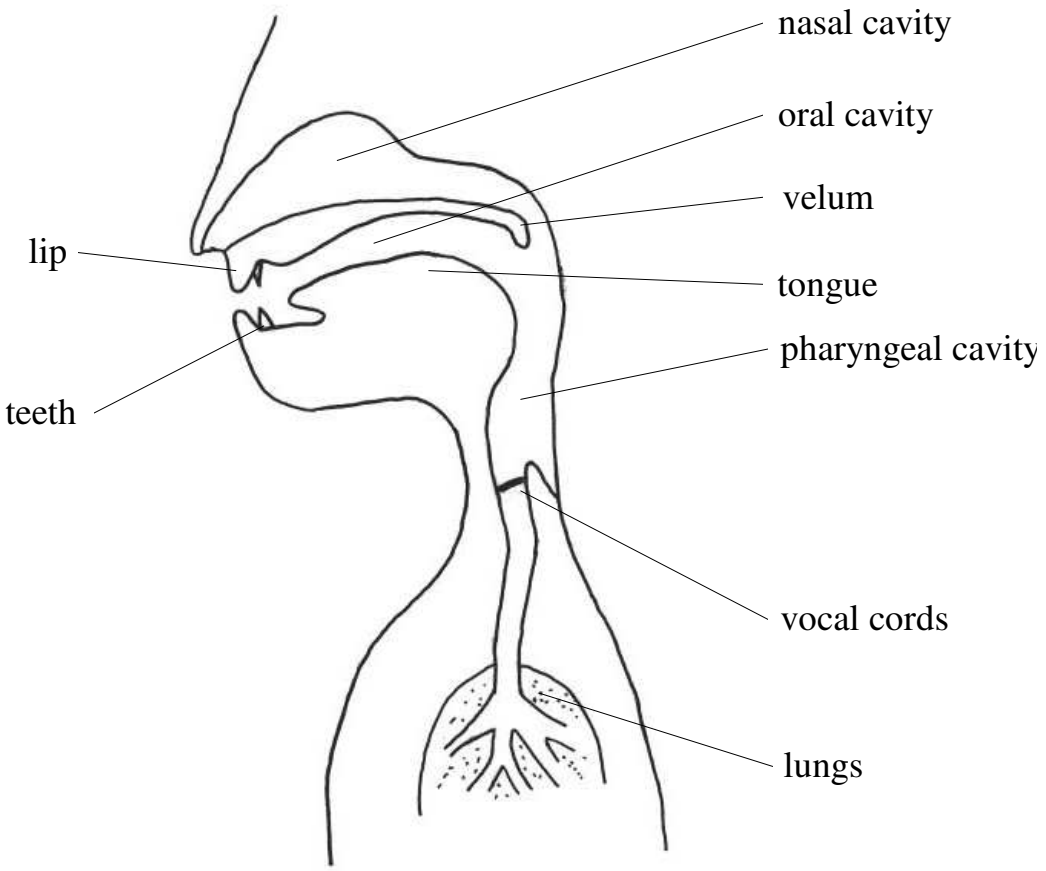


Figure 4.1: Schematic diagram of the human speech production system.

4.2.2 Voiced and unvoiced sound

The vibration of vocal cords plays an vital role in producing voiced and unvoiced sounds:

- *Voiced*: When the vocal folds are oscillating during a speech sound, the sound is said to be voiced, e.g. when pronouncing vowels.
- *Unvoiced*: When the vocal folds are too slack to oscillate during speech, the sound is said to be unvoiced, e.g. some consonants, *s*, *f*, etc.

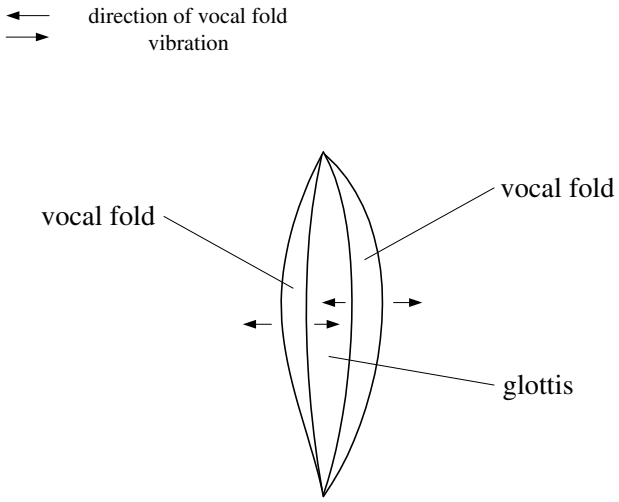


Figure 4.2: Superior view of the vocal cords.

4.3 LPC vocoder

Figure 4.3 shows an LPC vocoder for speech synthesis. The energy of the air stream expelled from lungs is modeled by a gain G. The vocal cords are modeled by two signal train generators which generate an excitation term $u(n)$. An impulse train generator models the vibration of vocal cords when it is tense and a white Gaussian noise generator models the slack state of the vocal cords. For a voiced sound, a periodic impulse train with unity amplitude from the impulse train generator is selected. A white Gaussian noise train is chosen for unvoiced sounds. A time varying digital filter is used for modeling the articulation tract, i.e. the vocal tract and nasal tract. The synthesis of speech can be described as follows [Dut97]:

$$S(z) = E(z)\frac{1}{A_p(z)} \tag{4.1}$$

$$s(n) = \sum_{i=1}^{p} a_i s(n-i) + e(n) \tag{4.2}$$

where $e(n) = Gu(n)$, p is the filter order and the a_i is the filter coefficients which

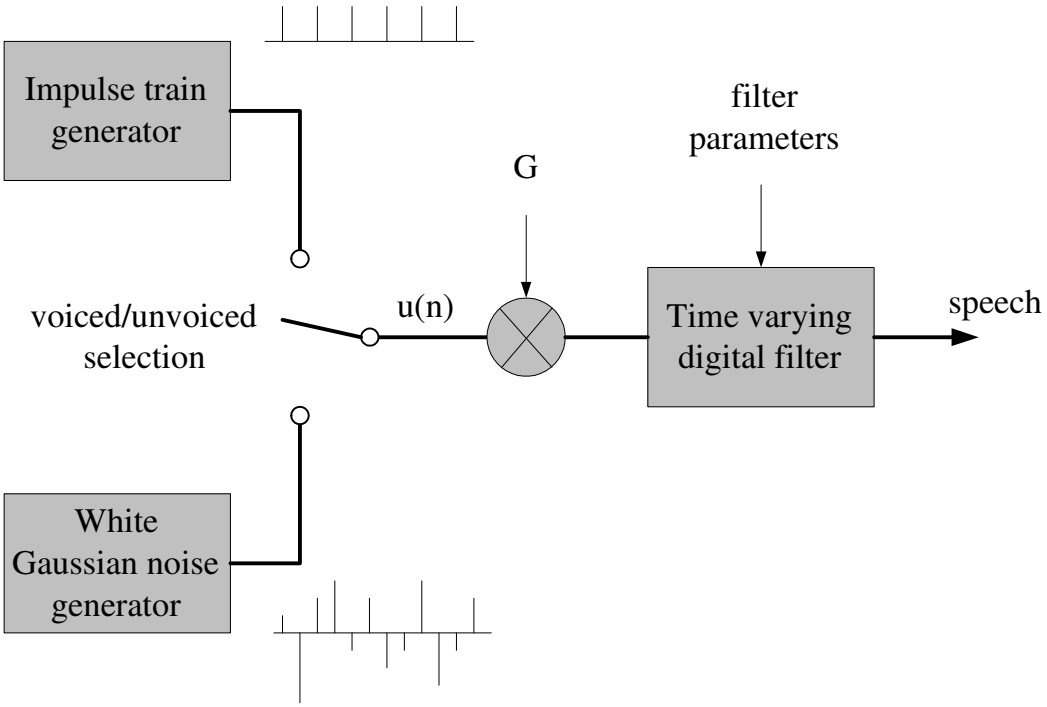


Figure 4.3: LPC-based speech synthesis.

define the resonance behaviors (frequency response) of the time varying digital filter. The computation of the filter coefficients is known as LPC analysis.

4.3.1 LPC analysis

From Equation 4.2, one can see that speech can be approximated as a linear combination of the previous p speech samples. LPC analysis thus computes the filter coefficients from the input speech signal, and minimizes the sum of the squared difference between the original speech and the approximated (synthetic) speech. The extracted filter coefficients are called LPC coefficients. This technique is widely applied in speech recognition and coding to represent speech features [Tre82]. In this work, it is also used as a speech feature.

Assume $s_{pd}(n)$ is the LPC approximated speech, and $s(n)$ is the original speech.

LPC analysis tries to find the filter coefficients a_i by minimizing the following error function:

$$\begin{aligned} E_e &= E[s(n) - s_{pd}(n)]^2 && (4.3) \\ &= E[s(n) - \sum_{k=1}^{P} a_k s(n-k)]^2 && (4.4) \end{aligned}$$

where P is the filter order. The minimum E_e can be found by taking the partial derivative with respect to each a_k and setting them to zero, giving:

$$E[s(n-i)s(n)] = \sum_{k=1}^{P} a_k E[s(n-i)s(n-k)]. \quad (4.5)$$

By defining:

$$\begin{aligned} \psi(i-k) &= E[s(n-i)s(n-k)], && (4.6) \\ \psi(i) &= E[s(n-i)s(n)] && (4.7) \end{aligned}$$

The filter coefficients a_k can be found by solving the following equations [RJ93]:

$$\begin{bmatrix} \psi(0) & \psi(1) & \psi(2) & \dots & \psi(P-1) \\ \psi(1) & \psi(2) & \psi(3) & \dots & \psi(P-2) \\ \psi(2) & \psi(3) & \psi(4) & \dots & \psi(P-3) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \psi(P-1) & \psi(P-2) & \psi(P-3) & \dots & \psi(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_P \end{bmatrix} = \begin{bmatrix} \psi(1) \\ \psi(2) \\ \psi(3) \\ \vdots \\ \psi(P) \end{bmatrix} \quad (4.8)$$

Two popular methods to solve the above equations are the autocorrelation and covariance methods. More details about the LPC analysis can be found in [RJ93].

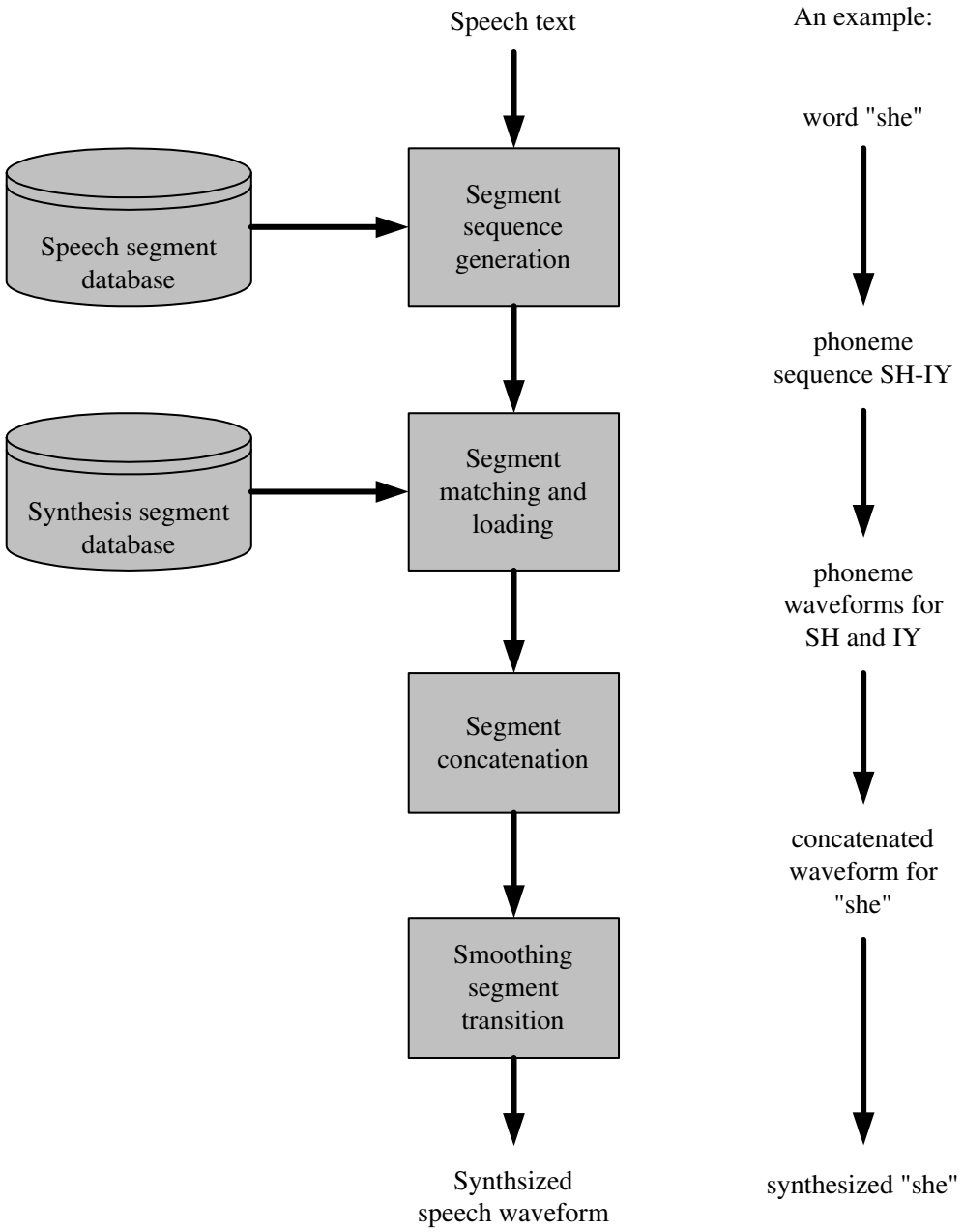


Figure 4.4: Speech synthesis by waveform concatenation.

4.4 Synthesis by waveform concatenation

The LPC-based speech synthesizer was developed from knowledge of the human speech production mechanism. Concatenative speech synthesis, however, uses limited information about the speech signals and synthesizes speech by concatenating pre-recorded speech segments [Bre92]. This method requires large memory to store the pre-recorded speech segments, but it is becoming more popular with the reduced cost of computer memory. This method is often yields superior speech quality [HAH01].

Figure 4.4 illustrates the synthesis flow. The speech segment database is formed from primitives from the given language, e.g. 42 phonemes for English. The synthesis segment database is the pre-recorded speech set for each element in the speech segment database. Each segment may be recorded several times under different conditions for better quality, e.g. different pitches to make the synthetic speech more natural. The speech is synthesized as follows:

- The speech text is split into segments based on the speech segment database and a segment sequence generated, e.g. if phonemes are chosen as the speech segment, the word *she* can be split into the phoneme sequence *SH-IY*.
- The segment matching and loading process sweeps the segment sequence and retrieves the corresponding segment waveform from the synthesis segment database.
- The retrieved segment waveforms are concatenated in order to form the segment sequence.
- The transitions between segment waveforms are smoothed to reduce the discontinuities and make the synthetic speech more natural.

In this work, phoneme frame is chosen as the basic speech segment. The synthesis segment database stores the enframed speech waveform for each phonemes. A

neural network is used to classify SEMG frames and decide which phoneme frame should be retrieved and concatenated. Details are described in Chapter 5. Using phonemes can reduce the complexity of the problem as there are only 42 phonemes. It is believed that the synthesis quality can be further improved by applying more sophisticated synthesis techniques. This work concentrates on generating correct phoneme sequences based on the corresponding SEMG signals.

4.4.1 Smoothing transition between speech segments

As mentioned above, the transition between speech segments should be smoothed to reduce the discontinuities. A popular method to reduce the discontinuity is called the overlap-and-add technique [HAH01]. Using this technique, each speech segment is multiplied with a tapered window, then the start of each segment is overlapped a certain duration with the end of its previous segment and added together. A commonly used windowing function is the Hanning window, defined as follows:

$$w[i] = 0.5 - 0.5\cos[\frac{2\pi i}{N-1}], \qquad 0 \leq i \leq N-1, \tag{4.9}$$

where N is the segment length. This window function is used in this work. Assume $p[n]$ and $q[n]$ are two speech segments to be concatenated. After multiplication with the window function:

$$p'[n] = w[n]p[n], \qquad 0 \leq n \leq d1 \tag{4.10}$$

$$q'[n] = w[n]q[n], \qquad 0 \leq n \leq d2 \tag{4.11}$$

where $d1$ and $d2$ are the length of $p[n]$ and $q[n]$ respectively, the resulting speech $s[n]$ can be formed as follows:

$$s[n] = \begin{cases} p'[n] & \text{if } 0 \leq n \leq d1 - R - 1 \\ p'[n] + q'[n - d1 + R] & \text{if } d1 - R \leq n \leq d1 - 1 \\ q'[n - d1 + R] & \text{if } d1 \leq n \leq d1 + d2 - R - 1 \end{cases} \tag{4.12}$$

where R is the length of the overlapping region.

4.5 Summary

This chapter began with an introduction to the human speech production mechanism, which involves three steps: the contraction of lungs, the vibration of vocal cords, and the resonance in the vocal tract. Two speech synthesis techniques were then introduced. The LPC vocoder is based on speech production mechanism and LPC coefficients are used as speech features in this work. The concatenative method synthesizes speech by simply concatenating pre-recorded speech segments. Phonemes are chosen as the speech segments in this work and the transition between phonemes can be smoothed using an overlap-and-add technique. In the next chapter, the design methodology will be presented, including the SEMG feature extraction and conversion to speech.

Chapter 5

An SEMG-based speech synthesis system

5.1 Introduction

In previous chapters, background on the nature of SEMG signals, classification techniques, and speech synthesis have been introduced. Previous SEMG-based speech recognition systems were also introduced where the features extracted from SEMG signal were classified into a set of words, however, various limitations exist in these kinds of systems. A frame-based approach will be introduced in this chapter, where the features are extracted from enframed SEMG signals and a speech waveform is synthesized on a frame-by-frame basis.

The chapter begins with a discussion of the limitations of previous work on SEMG-based speech recognition and the challenges of this work. The proposed methodology for SEMG-based speech synthesis is then presented. This is followed by a discussion of some design considerations and a summary is given in the last section.

5.2 Limitations of previous work

Previous work reviewed in Section 2.5 demonstrate the feasibility of recognizing speech based on SEMG signals. However, most of the previously proposed approaches focused on recognizing or classifying SEMG signals into a limited set of words. These approaches are similar to conventional isolated word recognition systems [RL81] in that there must be sufficient silence intervals before and after the speech signals, i.e., the words must be segmented and isolated from each other before recognition can be taken place. Although these approaches achieved satisfactory performance for SEMG signals, e.g. in [CEHL02b] and [KKAB04], the recognition accuracies were over 80%, they are not suitable for large vocabulary speech recognition and various limitations exist [HAH01] [RRWK83]. These include:

- *Untrained words*: Word recognition systems have difficulties in recognizing untrained words. Since the recognition model is built from words, in order to recognize a new word, the recognition models must be retrained.

- *Availability of training data*: When there are large numbers of words, it is difficult to collect a large amount of training data for each word while including includes all variabilities of the word.

To address the limitations of conventional isolated word recognition, instead of building whole-word recognition models, researchers proposed to recognize speech by building recognition models with smaller units [RRWK83]. This work proposes to synthesize speech from SEMG signals using a frame-based approach. Previous work on recognizing phonemes using a frame-based approach obtained poor accuracy, e.g. 64% in [ST85]. As pointed out by Morse and O'Brien [MO86], information for distinguishing SEMG signals for different words were observed throughout the duration of the whole word. Moreover, from their investigations on the correlation between data width and performance, they showed that using larger portions

of the whole word's SEMG signals to perform recognition could achieve better accuracy. These experimental results showed that recognizing SEMG frames (portion of the whole word's SEMG signals) is more difficult than recognizing SEMG signals of the whole words (isolated SEMG word recognition). Their experiment on distinguishing SEMG frames chopped from the whole words showed that, the recognition accuracy was only 40% for the same subject on a eight-word set. This work investigating the feasibility of synthesizing speech from SEMG signals using a frame-based approach is even more challenging.

5.3 The proposed methodology

To synthesize speech (words or sentences) using the proposed methodology, features are extracted from enframed SEMG signals and classified into a number of phonetic classes, the classification is done by a neural network which is trained using features extracted from parallel recorded SEMG and speech signals. The produced sequence of phonetic class number are mapped to acoustic signals by concatenating corresponding pre-recorded speech waveforms.

5.3.1 SEMG sensor positioning

Three channels of SEMG signals were collected and analyzed as shown in Figure 5.1. The first channel was collected from the cheek about 2.5cm from the nose, the second channel was collected from the chin and the third channel was collected from the lower lip. An additional electrode was attached to the forehead as a reference point. Speech was recorded using a microphone. The SEMG signal was amplified with a gain of 1000 using the circuit given in Appendix A. Both the amplified SEMG signal and speech were recorded concurrently using a National Instruments PCI6024E PCI data acquisition card [Nat] at a sampling rate of 8000Hz.

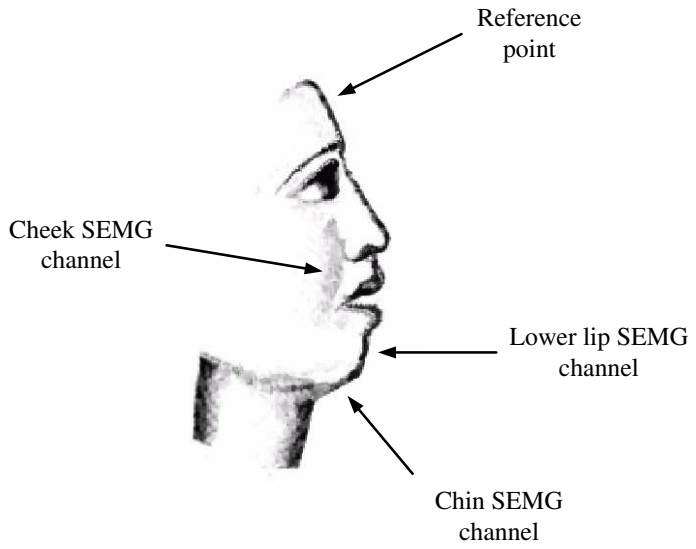


Figure 5.1: Electrode placement: SEMG signals were collected from the cheek, the chin, and lower lip, forehead was used as reference point.

5.3.2 Speech feature extraction

All speech signals (phonemes) in the training set are blocked into 22.5 ms frames, and there is no overlapping between frames. This scheme has been used in speech coding standard [Tre82]. For each speech frame, ten linear predictive (LP) coefficients, pitch and root mean square value are extracted. The pitch is the fundamental frequency of human speech, which is correlated to the vibration frequency of the vocal cords as described in Section 4.3, and the root mean square value (RMSV) is corresponding to the energy. The extracted LP coefficients, pitch, and root mean square value extracted from each speech frame are concatenated to form a speech feature vector (Figure 5.2).

Unsupervised clustering, based on the K-Means algorithm (see Appendix B), is then used to extract the representative feature vectors for the phonemes and silence. The extracted feature vectors form a speech-feature-vector codebook which can be used to label the SEMG signal during neural network training.

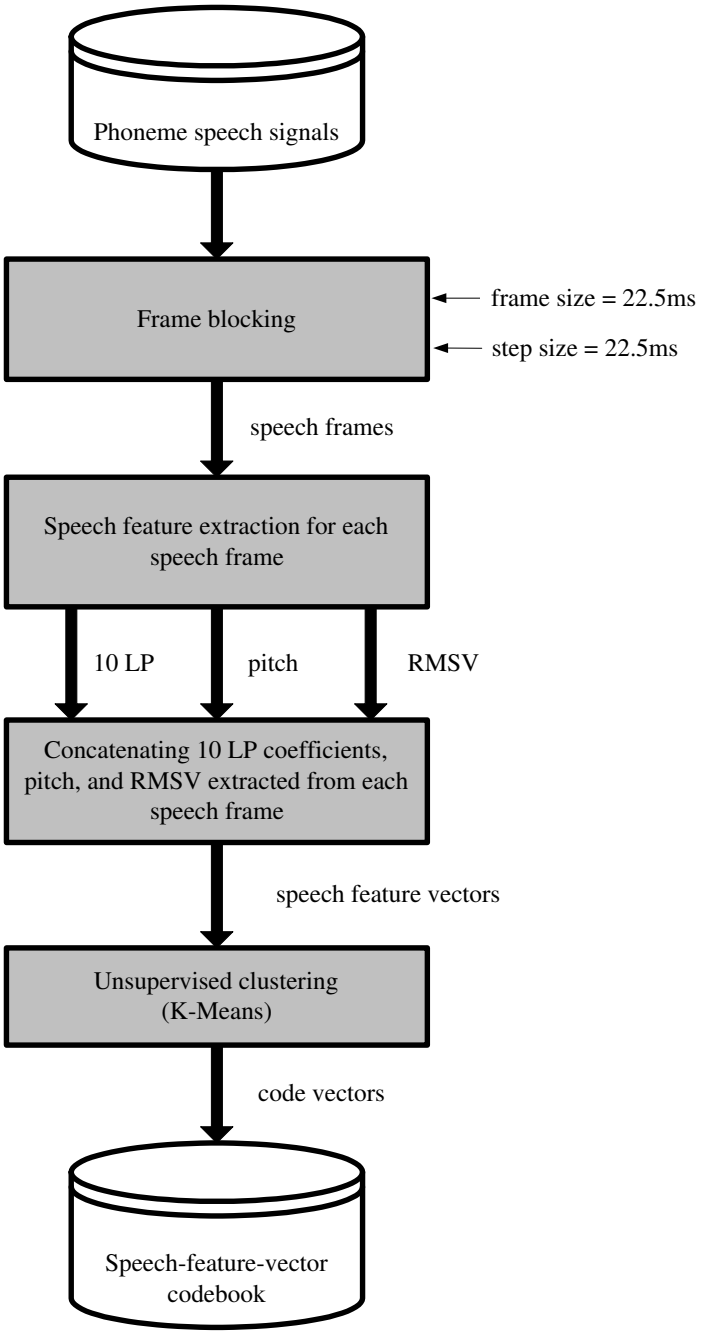


Figure 5.2: Speech feature extraction and forming speech-feature-vector codebook.

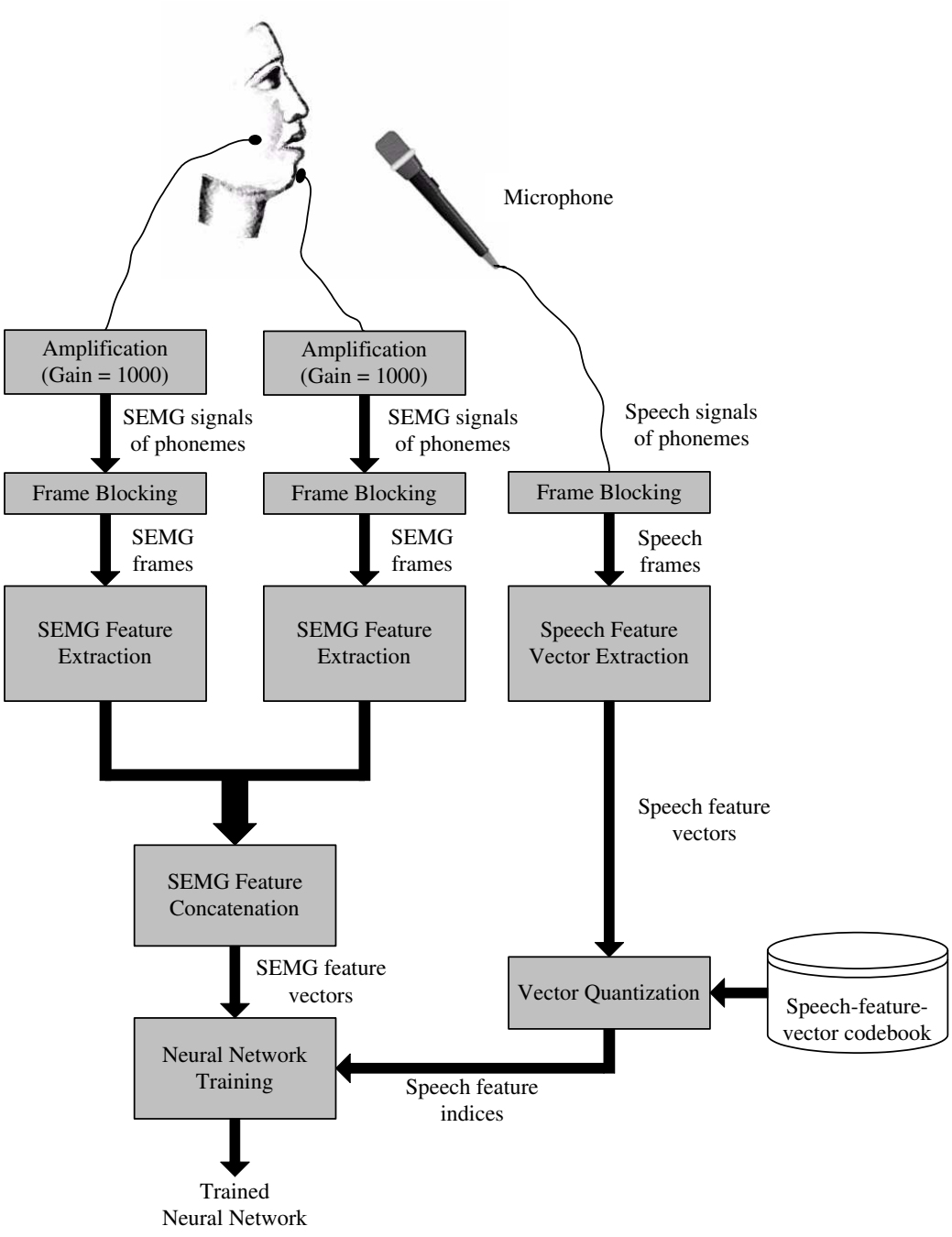


Figure 5.3: Frame-based feature extraction and neural network training.

5.3.3 Neural network training

The speech and SEMG signal are recorded concurrently, training pairs (input-target pairs) for the neural network being formed from the parallel recorded data as shown in Figure 5.3. This involves forming and labeling the SEMG feature vectors.

- *Forming SEMG feature vectors*: The SEMG signals of training phonemes from different SEMG channels (two channels in this figure) are blocked into frames, and features are extracted from each SEMG frame. SEMG features selection and SEMG channel positioning will be discussed in the next chapter. The extracted features from different channels are concatenated to form SEMG feature vectors.

- *Labeling the SEMG feature vectors*: The parallel recorded speech signals of the training phonemes are blocked into frames and the extracted speech feature vectors are quantized (see Appendix C) using the speech-feature-vector codebook. Thus, each speech frame is represented by a codebook index. Because the codebook is formed by the representative speech feature vectors, the speech feature index indicates to which phoneme a speech frame belongs. As the SEMG and speech signals are recorded in parallel, the speech feature index also indicates to which phoneme an SEMG frame belongs.

Each of the concatenated SEMG feature vectors is thus paired with the corresponding speech feature index to form an input-target training pair. The neural network, which takes an SEMG feature vector as input and produces speech feature indices as output, is trained using the input-target pairs. It is noted that only phonemes are involved in training.

In this work, a three-layer feed-forward backpropagation neural network is used. The number of input nodes is equal to the number of SEMG features. The number of output nodes is eight, as there are seven phonemes, one output node is allocated for each phoneme, and an additional one is used for silence.

5.3.4 Speech synthesis

After the neural network is trained, it can be applied to synthesize speech from input SEMG signals. In addition to synthesizing phonemes, the SEMG-based synthesis method proposed can also be applied to synthesize words as shown in Figure 5.4. To this end, SEMG signals recorded are blocked into frames, the features from different SEMG channels are concatenated to form SEMG feature vectors. Then the neural network is used to classify the concatenated SEMG feature vector into one of the seven phonemes or silence, which results in a sequence of speech feature indices for each word to be synthesized.

The error rate of the produced sequence of speech feature indices can be improved by using an phonetic smoothing technique, which is developed by assuming mid-term stationarity of speech signals. The details will be discussed in Section 5.4.

After smoothing the sequence of speech feature indices, a concatenative synthesis method is applied to reconstruct the target speech in a frame-by-frame basis. Based on the error corrected speech feature indices, target phoneme frames are loaded from the pre-recorded set and concatenated to form the complete speech. The transition between phonemes is smoothed using overlap and add method.

5.3.5 Potential Advantages

The training data set consists only of phonemes, but the proposed method is capable of recognizing any words whose phonetic transcription is formed from the training phoneme set. Although the number of recognized words increases exponentially with the number of phonemes involved in training, using this method, an unlimited vocabulary continuous speech synthesis is potentially realizable.

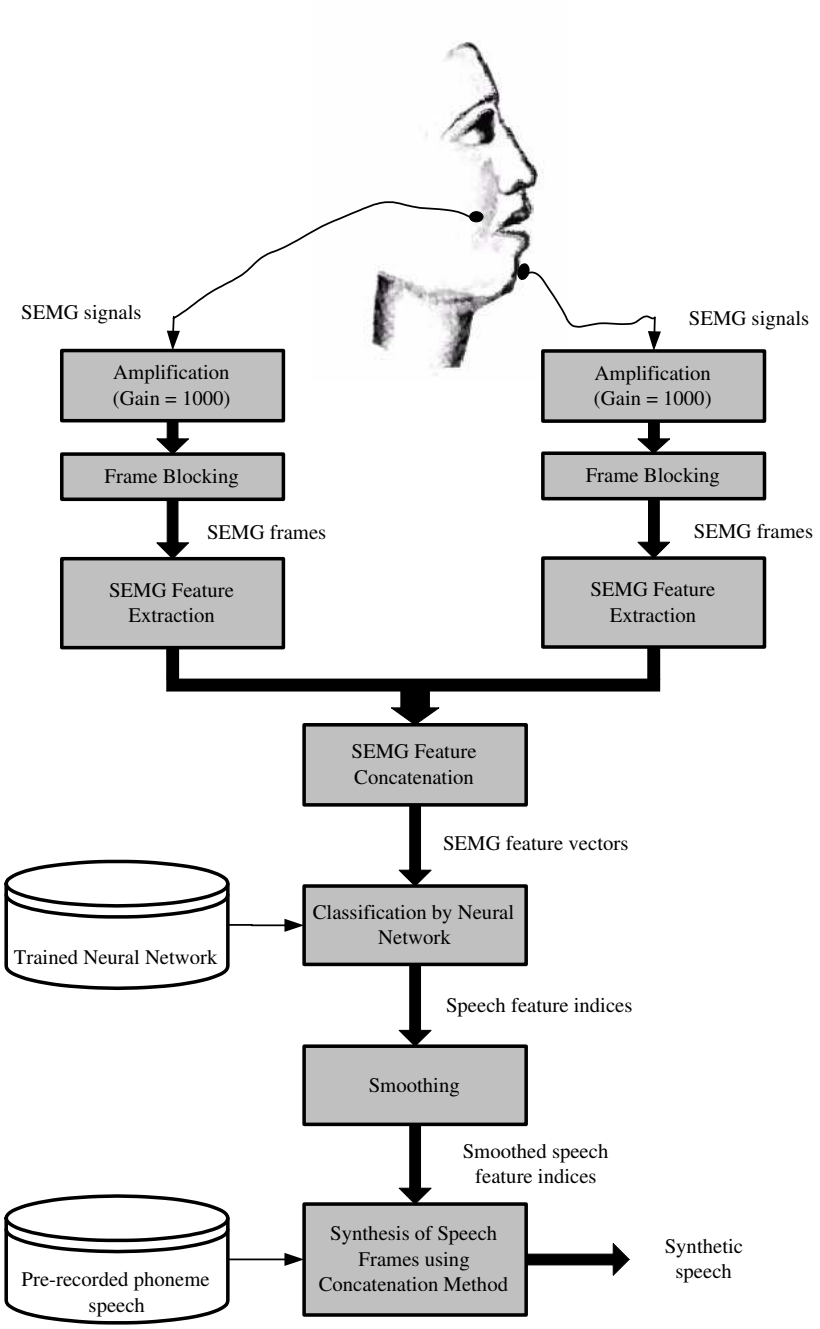


Figure 5.4: Speech synthesis from input SEMG signal.

5.4 Design considerations

The proposed methodology is a frame-based synthesis approach, several aspects need to be addressed.

5.4.1 SEMG feature extraction

The review in Section 2.5 showed that spectral features are useful and yield better performance than temporal features. The spectral features thus should be chosen carefully. This work provides an analytical analysis on the selection of the frequency band coefficients.

5.4.2 SEMG frame size

The SEMG frame size should be chosen carefully, because it affects the frequency resolution [KGA01]. If a small frame size is used, better time resolution can be obtained, but this results in poor frequency resolution. On the other hand, using larger frame size can improve the frequency resolution, but results in a loss of information between adjacent frames. In this work, correlation between frame size and performance is analyzed, finding optimal frame size that can balance the performance and maintain maximum time resolution is addressed.

5.4.3 Channel positioning

Previous proposed SEMG-based word recognition system using SEMG signal collected from different positions, such as a two-channel system [JLA03] from the chin, a three-channel system [MZ04] from cheek, chin, and upper lip, a five-channel system [CEHL01] from major facial muscles. The effect of different channels in distinguish SEMG frames for different speech has not been addressed. Analyzing the correlation between different sensor positions and performance is one of the major concern in this work.

5.4.4 Smoothing phonetic sequence

To synthesize speech, input SEMG signals are classified into an sequence of speech feature indices. Due to classification errors, some indices in the produced sequence are incorrect (Figure 5.5), and these misclassifications appear as fragments embedded in the sequence. Based on this observation, a smoothing technique was applied in an attempt to remove these fragments.

Majority-filter-based error correction

Based on the observation that voiced speech signals are fairly stationary over a short period of time and, in contrast, characteristics of the signal change over long periods of time, i.e. on the order of 200ms or more [RJ93]. A majority filter which attempts to remove glitches due to misclassification was studied. This correction technique involves scanning the produced sequence of speech feature indices over a window of 9 indices (i.e. 202.5ms) with step 1, the index id with the highest frequency f within the window is found, and a new index equal to id is produced if the frequency f exceed a threshold. From the example in Figure 5.5, one can see that, after applying majority filtering, error indices in phoneme *SH* are corrected.

Correction based on triggering

There are still some errors that cannot be corrected by employing a majority-filter-based error correction technique, especially when the error indices are close to each other. From the example in Figure 5.5, one can see that, there are still incorrect indices present in phoneme *IY* after applying this technique.

The trigger-based correction process sweeps the index sequence using a window of nine consecutive indices with step 1, and an index is generated based on the similarity of all indices in each window. If all the nine indices are the same, an index equal to the nine indices, is generated. The generated index remains unchanged if the indices in the next window are not the same, and the index changes again when

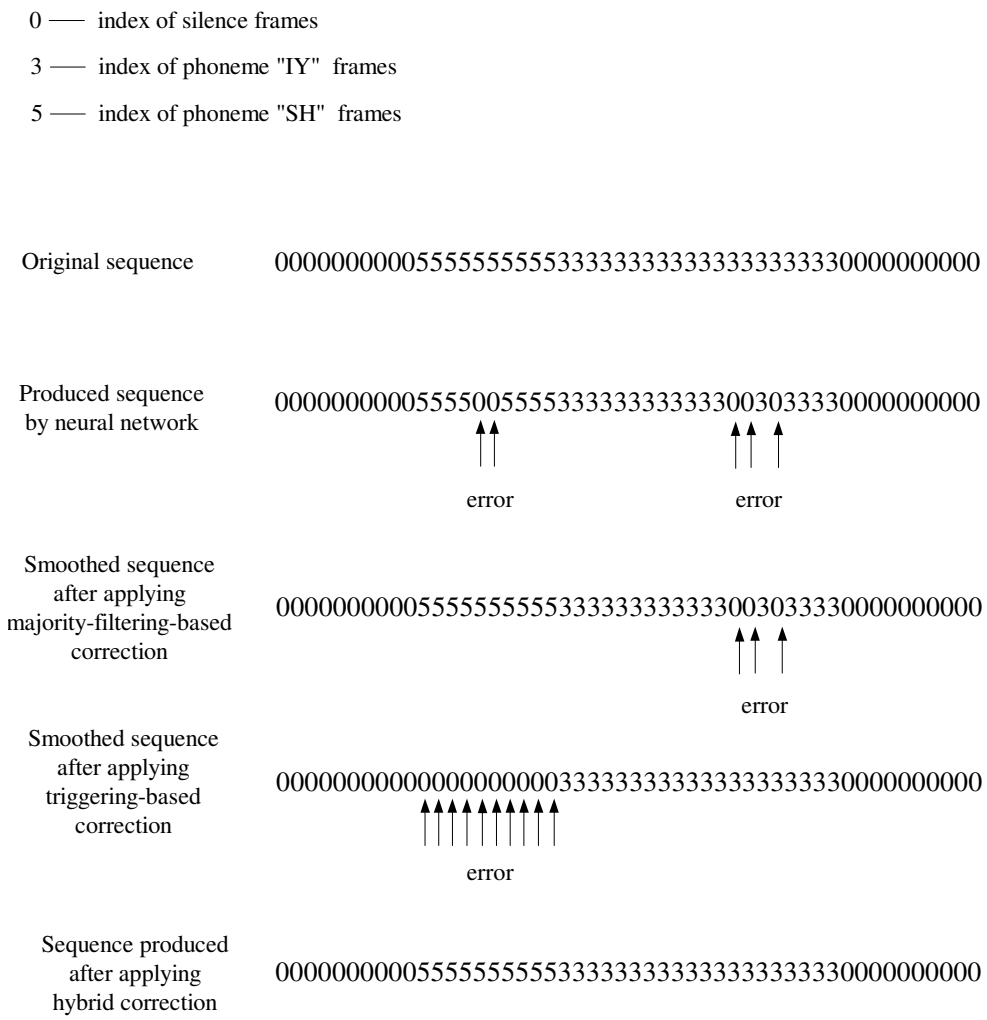


Figure 5.5: An example showing the classification error in the sequence of speech feature indices for word *she*. The arrows indicate the misclassification in the produced sequence.

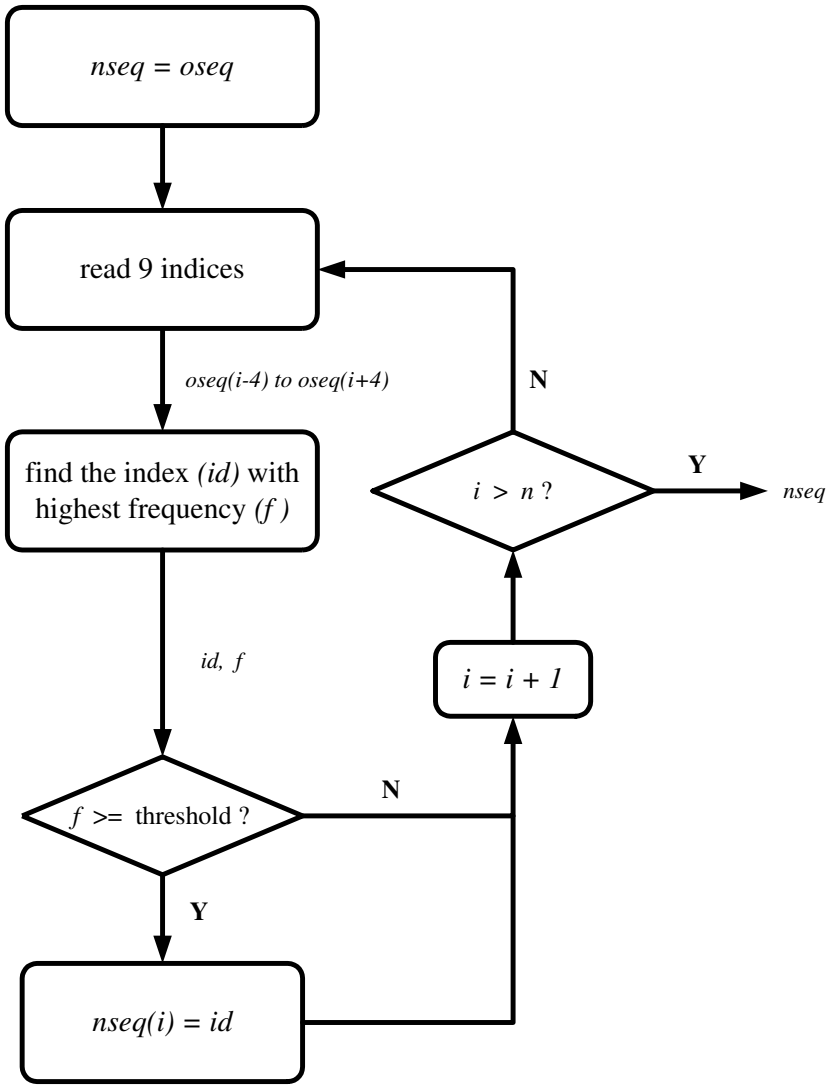


Figure 5.6: Majority filter based smoothing technique. Where *oseq* is the sequence produced by neural network, *nseq* is the smoothed sequence, *n* is the sequence length of *oseq*.

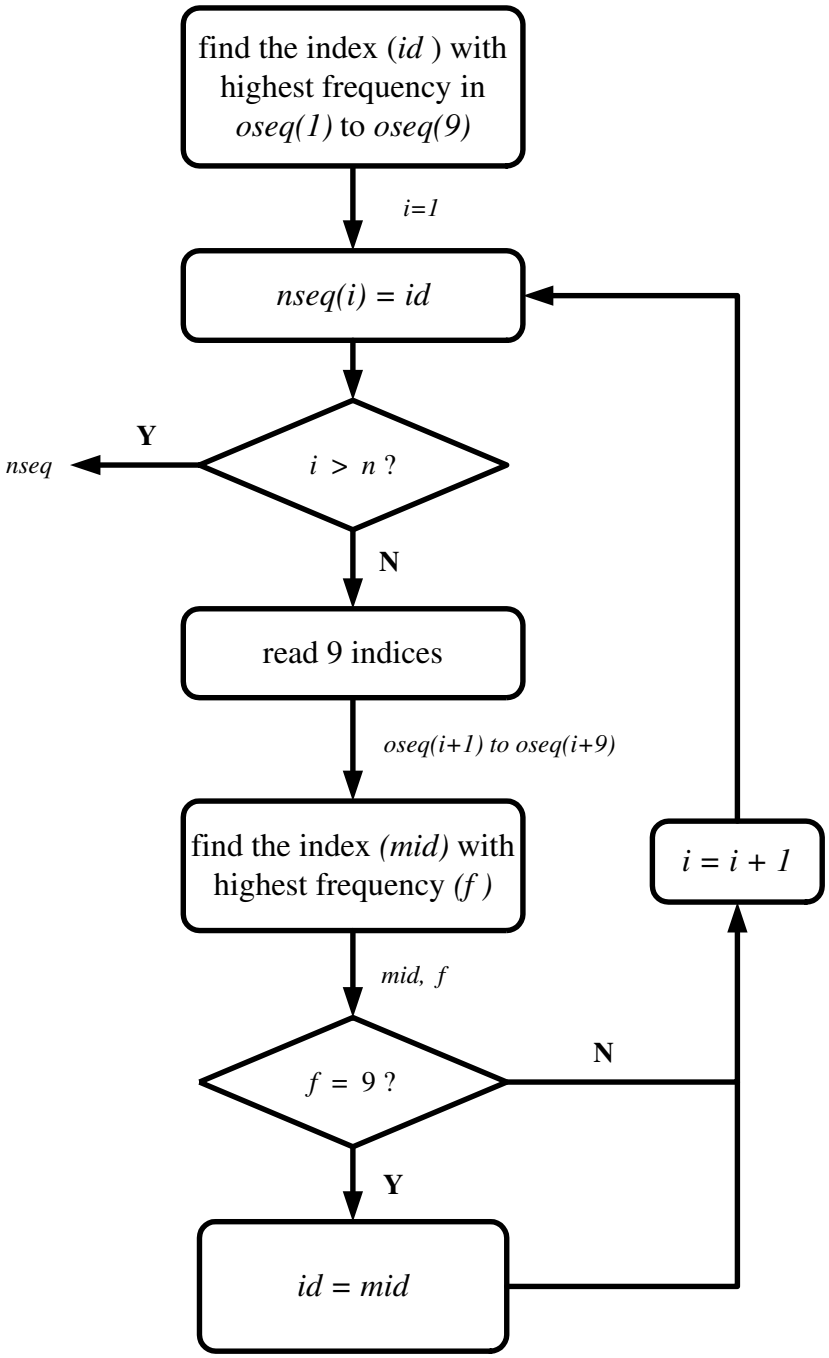


Figure 5.7: Trigger based smoothing technique. *oseq* is the sequence produced by neural network, *nseq* is the smoothed sequence, and *n* is the sequence length of *oseq*.

the next nine indices are the same. For the example in Figure 5.5, after sweeping the sequence produced by neural network, the errors in phoneme *IY*, cannot be corrected by majority-filter technique, is corrected. However, a problem arises, more errors are produced in the resulting sub-sequence of phoneme *SH*.

Hybrid approach

An hybrid approach that integrating majority-filter and trigger techniques is used to correct the classification error. The sequence generated by neural network is firstly pass through the correction process based on majority-filter, the smoothed sequence is then passed to trigger-based correction process for further smoothing. The majority-filter based process is removing the single index error, and the trigger-based process strengthen each sub-sequence. As shown by Figure 5.5, all error indices are corrected after using the hybrid smoothing technique.

5.4.5 Smoothing phoneme transition

The overlap-and-add technique is used to reduce the discontinuity between phoneme transition. Each phoneme is overlapped 25% of its length with its adjacent phonemes, the phoneme waveform is multiplied with a Hanning window before added together.

5.5 Summary

In this chapter, a methodology to synthesis speech from SEMG signal is presented. To synthesis speech, input SEMG signal is blocked into frames, features extracted are classified into sequence of phonetic labels, concatenative method is then used to reconstructed the original speech waveform based on the sequence of phonetic labels. Several aspects, such as SEMG feature selection and channel positioning, should be carefully addressed. A hybrid smoothing technique is proposed to correct the errors in the sequence of phonetic labels, which consists of majority-filter-based

and trigger-based correction technique.

Since the neural network is trained using features extracted from the enframed phoneme SEMG and speech signal, speech is then synthesized frame by frame, the proposed methodology is applicable for continuous speech synthesis with unlimited vocabulary. In the next chapter, some experimental results will be presented.

Chapter 6

Spectral feature assessment of SEMG signals

6.1 Introduction

In the previous chapter, a frame-based speech synthesis method was introduced, where the features were extracted from enframed SEMG signals and classified into phonetic classes. Frequency band coefficients are useful feature to analyze SEMG signals, however, parameters of filter band coefficient are often selected arbitrarily, e.g. in [PBYTI02]. It is believed that spectral feature selection plays a vital role in distinguish different speech and performance can be improved by carefully selecting spectral features. In this chapter, the spectral feature selection process conducted in this work will be discussed in detail.

This chapter is organized as follows. The divergence score, used to measure the quality of features, is introduced in Section 6.2. This is followed by a description of the data set used in Section 6.3. Two spectral feature extraction methods, non-overlapping frequency band and overlapping frequency band, are described and compared in Section 6.4 and 6.5. The spectral feature used in this work is presented in Section 6.6 and summary is given in the last section.

6.2 Separability measuring

To measure the efficacy of various features, one approach is to directly present the extracted feature vectors to a neural network and then evaluate the classification performance. However, as there are a large number of different feature extraction schemes, it would be prohibitively time consuming to test all combinations. In this work, divergence [TK03] is thus used to measure the utility of different features. The divergence is inferred from the Bayes classification rule and used as a separability measure of two distributions. In Bayes classification rule, given two classes ω_1 and ω_2, a feature vector $\mathbf{x}$ is classified into ω_1 if

$$P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x}).$$

Alternatively, the likelihood ratio between $P(\mathbf{x}|\omega_1)$ and $P(\mathbf{x}|\omega_2)$ thus represents the discriminatory capability between two classes ω_1 and ω_2, and the mean log likelihood ratio over class ω_1 is calculated as

$$\text{DIV}_1 = \int_{-\infty}^{+\infty} P(\mathbf{x}|\omega_1) \ln \frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} \tag{6.1}$$

and for ω_2

$$\text{DIV}_2 = \int_{-\infty}^{+\infty} P(\mathbf{x}|\omega_2) \ln \frac{P(\mathbf{x}|\omega_2)}{P(\mathbf{x}|\omega_1)} \tag{6.2}$$

The divergence between the two classes, ω_1 and ω_2, is calculated as the sum

$$\text{DIV}_{12} = \text{DIV}_1 + \text{DIV}_2. \tag{6.3}$$

The above equation can be transformed to

$$\begin{aligned} \text{DIV}_{12} = \frac{1}{2} trace(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 - 2I) + \\ \tfrac{1}{2}(\mu_1 - \mu_2)'(\Sigma_1^{-1} + \Sigma_2^{-1})(\mu_1 - \mu_2), \end{aligned} \tag{6.4}$$

where μ_1 and Σ_1 are the mean and covariance of class ω_1, and μ_2 and Σ_2 are the mean and covariance of class ω_2. The average divergence is calculated as follows:

$$\text{DIV\_AVG} = \frac{\sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{DIV}_{ij}}{\sum_{i=1}^{N-1} i}, \tag{6.5}$$

where N is the number of classes. In this work, N is equal to 8, which includes one silence class and 7 phoneme classes (Section 6.3).

Another term, ASF_DIV_AVG, is used to analyze the average separability capability of individual features within the feature vector, where the average divergences for individual features, DIV_AVGs, are calculated using Equation 6.5, and the average of these is called the ASF_DIV_AVG coefficient, i.e.

$$\text{ASF\_DIV\_AVG} = \frac{\sum_{k} \text{DIV\_AVG}}{P}, \tag{6.6}$$

where k is the kth dimension of the feature vector, and P is the total dimension number of the feature vector.

6.3 Analysis data set

Using the experimental setup described in Section 5.3.1, an analysis was done using a phoneme set consisting of: *ae, iy, ao, uw, sh, f* and *s*. Data sets were recorded in a twenty-second period, during which the speaker repeatedly pronounced a given phoneme. This process was repeated four times for each phoneme and the collected data was used for analysis. The concurrently recorded speech signals were used as a reference to label the SEMG signals.

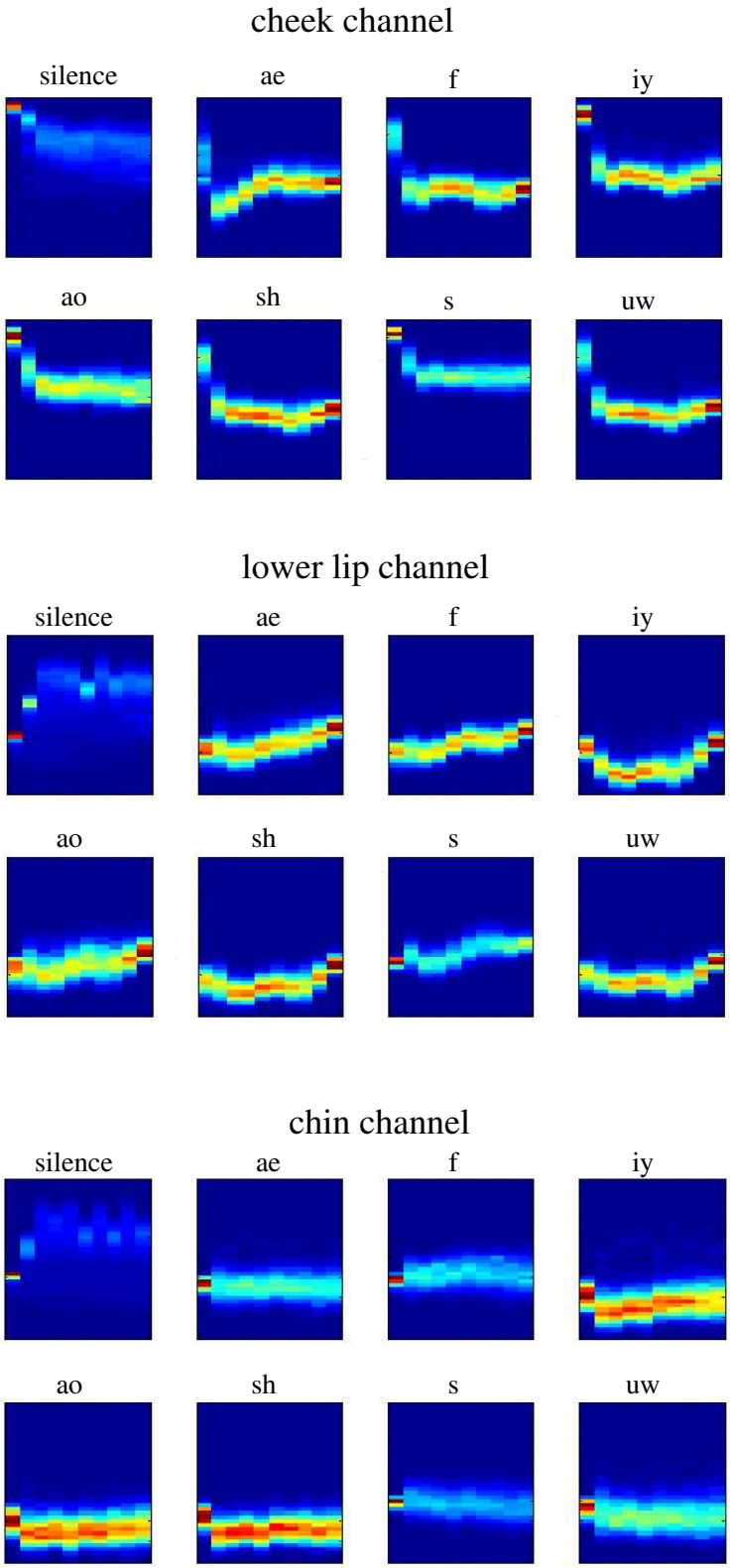


Figure 6.1: Distribution of NOFBC. The horizontal axis is the NOFBC number 1 - 10 from left to right, and the vertical axis is the amplitude of the NOFBC (lower corresponds to larger amplitude). The color represents the number of NOFBCs.

NOFBC number	Frequency region
NOFBC 1	0 Hz – 50 Hz
NOFBC 2	50 Hz – 100 Hz
NOFBC 3	100 Hz – 150 Hz
NOFBC 4	150 Hz – 200 Hz
NOFBC 5	200 Hz – 250 Hz
NOFBC 6	250 Hz – 300 Hz
NOFBC 7	300 Hz – 350 Hz
NOFBC 8	350 Hz – 400 Hz
NOFBC 9	400 Hz – 450 Hz
NOFBC 10	450 Hz – 500 Hz

Table 6.1: NOFBC number and corresponding frequency region for $N = 10$.

6.4 Non-overlapping frequency bands

The SEMG signals of the were blocked into 112.5 ms frames, the frequency spectra from 0 Hz to 500 Hz were calculated for each frame and divided into N equal non-overlapping frequency sections; the bandwidth of each section is $500/N$ Hz. The frequency components in each section were summed to give one coefficient (called the non-overlapping frequency band coefficient or NOFBC) corresponding to that section. This results in N NOFBCs and non-overlapping schemes are often used to analyze SEMG signals, e.g. [PBYTI02]. Table 6.1 shows the NOFBCs and their corresponding frequency regions for $N = 10$.

Figure 6.1 shows the distribution of the 10 NOFBCs for different phonemes and SEMG channels. This preliminary view clearly shows the variability between different phonemes and the similarities for the same phoneme. As we can see in this figure, the NOFBCs are compacted in a small variance for the same phoneme. On the other hand, variation can be found for different phonemes, particularly in the amplitude. For example, the amplitudes for silence are much smaller than other phonemes. Although the amplitude of some phoneme may be similar, variation can be observed by comparing their shape over the NOFBC bands. For example, the shape of phoneme *ae* and *f* in the cheek channel are different, but their amplitudes

	Silence	*ae*	*f*	*iy*	*ao*	*sh*	*s*	*uw*
Silence	0.00	33.98	41.70	13.67	6.45	48.67	8.67	26.27
ae	33.98	0.00	12.66	47.69	41.00	19.04	88.90	18.68
f	41.70	12.66	0.00	20.64	26.61	3.00	68.98	4.63
iy	13.67	47.69	20.64	0.00	5.08	21.77	15.77	12.55
ao	6.45	41.00	26.61	5.08	0.00	28.47	6.28	16.71
sh	48.67	19.04	3.00	21.77	28.47	0.00	74.46	1.15
s	8.67	88.90	68.98	15.77	6.28	74.46	0.00	49.04
uw	26.27	18.68	4.63	12.55	16.71	1.15	49.04	0.00
Mean score	25.63	37.42	25.46	19.00	18.65	28.08	44.59	18.43

Table 6.2: Divergence scores of different phonemes using 10 NOFBCs from cheek channel.

	Silence	*ae*	*f*	*iy*	*ao*	*sh*	*s*	*uw*
Silence	0.00	25.98	58.40	51.55	18.25	83.30	23.27	47.80
ae	25.98	0.00	3.09	10.07	2.12	10.73	9.78	7.32
f	58.40	3.09	0.00	19.38	2.54	13.62	15.17	7.98
iy	51.55	10.07	19.38	0.00	10.62	5.07	25.73	5.62
ao	18.25	2.12	2.54	10.62	0.00	10.25	9.13	5.01
sh	83.30	10.73	13.62	5.07	10.25	0.00	41.84	2.01
s	23.27	9.78	15.17	25.73	9.13	41.84	0.00	25.01
uw	47.80	7.32	7.98	5.62	5.01	2.01	25.01	0.00
Mean score	44.08	9.87	17.17	18.29	8.27	23.83	21.42	14.39

Table 6.3: Divergence scores of different phonemes using 10 NOFBCs from lower lip channel.

are quite similar.

Table 6.2, 6.3, 6.4 shows the divergence scores for different SEMG channel using 10 NOFBCs and the average scores (DIV_AVG) calculated using Equation 6.5 are shown in Table 6.5. These tables show that the cheek channel is best able to separate the SEMG feature vectors on average, the lower lip channel is the next, and the chin channel is the worst. However, the cheek channel is not always the best. For example, the lower lip channel is better than the cheek channel for separating silence from other phonemes, as the mean divergence score is 44.08, compared with 25.63 using the cheek channel.

The DIV_AVGs for different numbers of frequency bands are calculated and

	Silence	*ae*	*f*	*iy*	*ao*	*sh*	*s*	*uw*
Silence	0.00	11.57	10.15	17.44	28.59	32.95	11.30	12.55
ae	11.57	0.00	1.99	5.67	6.75	6.35	1.89	0.80
f	10.15	1.99	0.00	12.74	16.05	15.91	1.67	4.04
iy	17.44	5.67	12.74	0.00	1.04	1.25	14.15	3.82
ao	28.59	6.75	16.05	1.04	0.00	0.63	16.39	3.73
sh	32.95	6.35	15.91	1.25	0.63	0.00	14.21	3.78
s	11.30	1.89	1.67	14.15	16.39	14.21	0.00	2.73
uw	12.55	0.80	4.04	3.82	3.73	3.78	2.73	0.00
Mean score	17.79	5.00	8.93	8.02	10.45	10.72	8.91	4.49

Table 6.4: Divergence scores of different phonemes using 10 NOFBCs from chin channel.

Position	DIV_AVG
Cheek	27.2
Lower lip	19.7
Chin	9.3

Table 6.5: Comparison of DIV_AVG score using 10 NOFBC from different SEMG channel.

the results are shown in Figure 6.2. One can see that the DIV_AVG value is larger for a larger number of frequency bands. This shows that SEMG feature vectors are more separable for a larger number of frequency bands, and hence, using more frequency bands are better for capturing the variability of SEMG feature vectors between different phonemes.

The ASF_DIV_AVGs for different numbers of frequency bands, are calculated and the results shown in Figure 6.3. The separability of individual features can be seen to be increasing with bandwidth. This is reasonable as each frequency band is capable of capturing more SEMG features with larger bandwidth and is thus more representative of the SEMG characteristics for different phonemes.

Figure 6.2 shows that DIV_AVG increases with the number of frequency bands, however, it is almost saturates after the number of frequency bands is larger than 5. One of the reasons is that the bandwidth is smaller for more frequency bands and the separability of each frequency band becomes lower. A problem of balancing the

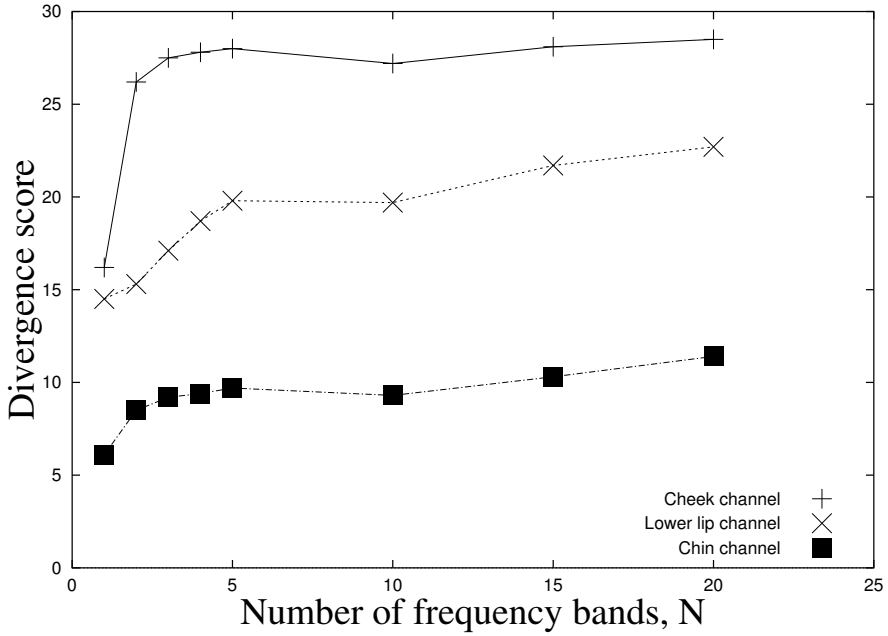


Figure 6.2: DIV_AVG scores for different numbers of frequency bands using the NOFBC feature.

trade-off between the number of frequency band and bandwidth is thus confronted. In the next section, overlapping bands are used to address this problem.

6.5 Overlapping frequency bands

The overlapping method partitions the full frequency range into several bands, adjacent bands being overlapped over a certain interval. Define N to be the number of frequency bands of bandwidth ω. The frequency range FR in each frequency band is:

$$FR_i = [(i-1)\eta, (i-1)\eta + \omega], \quad \text{for} \quad i = 1, 2, 3, ..., N$$

where

$$\eta = (500 - \omega)/(N - 1).$$

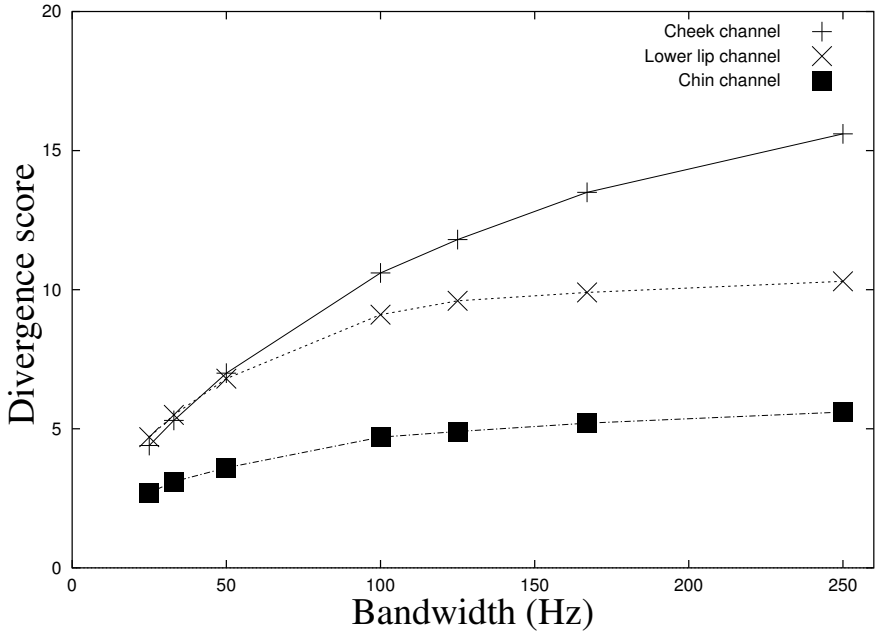


Figure 6.3: ASF_DIV_AVG scores for different bandwidths using NOFBC feature.

The SEMG signals for the analysis data set were also blocked into 112.5 ms frames and the frequency spectra from 0 Hz to 500 Hz were calculated for each frame. Frequency responses in each band were summed to give one coefficient (called the overlapping frequency band coefficient or OFBC) corresponding to that band, which results in N OFBCs.

Figure 6.3 shows the separability of individual frequency band increases with the bandwidth. However, especially for the lower lip and chin channels, the separability tends to saturate for bandwidths larger than 140 Hz. As a result, OFBC features with $N = 10$ and $\omega = 140$ Hz were selected in this work. Table 6.6 shows OFBCs and their corresponding frequency regions,

Table 6.7, 6.8, 6.9 shows the divergence scores for different SEMG channels using 10 OFBCs with a bandwidth of 140 Hz. Average scores (DIV_AVG) calculated are shown in Table 6.10. It is interesting to note that there are similarities compared with the results obtained using NOFBCs. In particular, the cheek channel is best for separating the SEMG feature vectors on average, the next is the lower lip

OFBC number	Frequency region
OFBC 1	0Hz – 140Hz
OFBC 2	40Hz – 180Hz
OFBC 3	80Hz – 220Hz
OFBC 4	120Hz – 260Hz
OFBC 5	160Hz – 300Hz
OFBC 6	200Hz – 340Hz
OFBC 7	240Hz – 380Hz
OFBC 8	280Hz – 420Hz
OFBC 9	320Hz – 460Hz
OFBC 10	360Hz – 500Hz

Table 6.6: OFBC number and corresponding frequency region for $N = 10$, $\omega = 140$ Hz.

	Silence	*ae*	*f*	*iy*	*ao*	*sh*	*s*	*uw*
Silence	0.00	51.19	79.57	17.51	9.09	64.69	13.44	33.91
ae	51.19	0.00	13.80	66.96	80.78	21.72	216.07	21.26
f	79.57	13.80	0.00	34.29	54.22	3.49	145.57	5.62
iy	17.51	66.96	34.29	0.00	5.19	30.45	17.37	19.70
ao	9.09	80.78	54.22	5.19	0.00	44.17	6.51	27.86
sh	64.69	21.72	3.49	30.45	44.17	0.00	119.63	1.17
s	13.44	216.07	145.57	17.37	6.51	119.63	0.00	80.11
uw	33.91	21.26	5.62	19.70	27.86	1.17	80.11	0.00
Mean score	38.49	67.40	48.08	27.35	32.55	40.76	85.53	27.09

Table 6.7: Divergence scores of different phonemes using 10 OFBCs from cheek channel, where $\omega = 140$ Hz.

channel, and the lower lip channel is better than the cheek channel for separating silence from other phonemes. On average, using 10 OFBCs is better than using 10 NOFBCs for distinguishing different phonemes (see Figure 6.4).

From Table 6.7, 6.8, 6.9, one can see that different channels are better for distinguishing different phonemes, i.e. while a channel may be bad for separating a particular phoneme, another channel may be good. For example, the cheek channel is better than the chin channel for distinguishing between *s* and *f*. The lower lip channel is better than the cheek for distinguishing between *iy* and *s*. However, the chin channel is almost always poor for all phoneme pairs. It only performs slightly

	Silence	*ae*	*f*	*iy*	*ao*	*sh*	*s*	*uw*
Silence	0.00	59.74	93.69	104.00	38.26	147.28	49.25	70.97
ae	59.74	0.00	3.47	9.08	2.34	11.90	12.97	7.04
f	93.69	3.47	0.00	19.66	2.65	14.63	17.12	8.69
iy	104.00	9.08	19.66	0.00	11.09	3.45	52.01	5.14
ao	38.26	2.34	2.65	11.09	0.00	12.34	10.80	5.23
sh	147.28	11.90	14.63	3.45	12.34	0.00	65.66	2.76
s	49.25	12.97	17.12	52.01	10.80	65.66	0.00	31.80
uw	70.97	7.04	8.69	5.14	5.23	2.76	31.80	0.00
Mean score	80.46	15.22	22.84	29.20	11.81	36.86	34.23	18.80

Table 6.8: Divergence scores of different phonemes using 10 OFBCs from lower lip channel, where $\omega = 140$ Hz.

	Silence	*ae*	*f*	*iy*	*ao*	*sh*	*s*	*uw*
Silence	0.00	19.34	16.65	24.78	44.95	50.30	17.81	22.21
ae	19.34	0.00	2.52	9.29	9.90	9.79	1.95	0.89
f	16.65	2.52	0.00	26.10	28.24	27.26	1.73	5.42
iy	24.78	9.29	26.10	0.00	1.41	1.61	22.89	5.33
ao	44.95	9.90	28.24	1.41	0.00	0.83	23.51	4.93
sh	50.30	9.79	27.26	1.61	0.83	0.00	22.49	5.58
s	17.81	1.95	1.73	22.89	23.51	22.49	0.00	3.53
uw	22.21	0.89	5.42	5.33	4.93	5.58	3.53	0.00
Mean score	28.01	7.67	15.42	13.06	16.25	16.84	13.41	6.84

Table 6.9: Divergence scores of different phonemes using 10 OFBCs from chin channel, where $\omega = 140$ Hz.

better in separating a limited number of phoneme pairs, such as *sh* and *f*.

The results obtained in Table 6.7, 6.8, 6.9 also show that characteristics of SEMG and speech signals are very different. All channels are bad for separating phoneme *sh* from *uw*, but the speech signals for these two phonemes are actually totally different, where phoneme *uw* is a voiced sound with vocal cord vibration and *sh* is a unvoiced sound. These two phonemes can be easily distinguished by human or acoustic speech recognition systems. The divergence score shows that SEMG characteristics of these two phonemes are similar as the degree of lip-rounding is quite similar resulting in similar muscle activities.

Position	DIV_AVG
Cheek	45.9
Lower lip	31.2
Chin	14.7

Table 6.10: Comparison of DIV_AVG using 10 OFBCs from different SEMG channel, where $\omega = 140$ Hz.

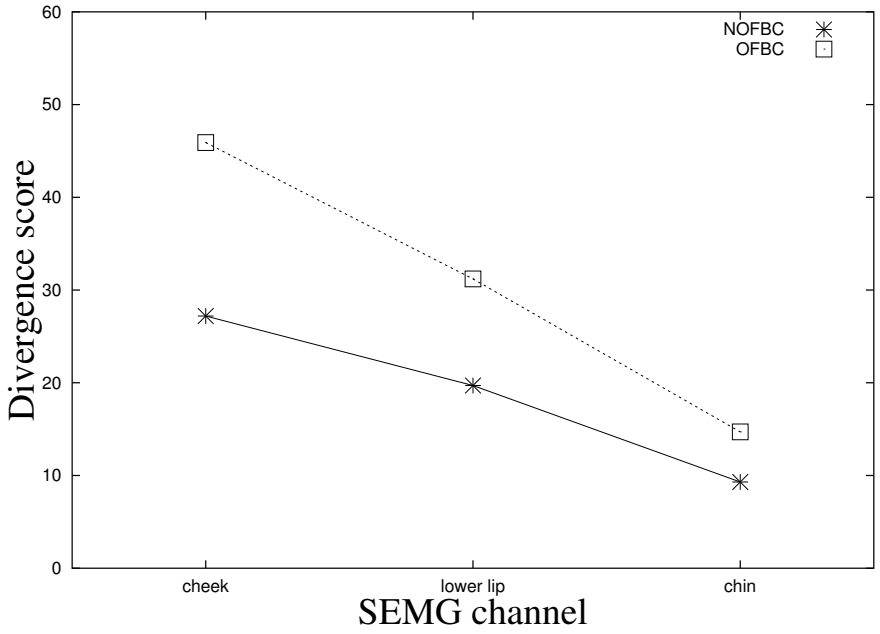


Figure 6.4: Comparison of DIV_AVG for 10 NOFBCs and 10 OFBCs.

6.6 Feature selection

The above analysis shows that the cheek and lower lip channels are the best two channels for distinguishing phonemes. These two channels were chosen for this reason in this work. Using a bandwidth of 140 Hz, 10 OFBCs extracted from each channel are concatenated and results in a total of 20 OFBCs. This configuration is used for further experiments. The divergence scores calculated using 20 OFBCs are shown in Table 6.11. One can see that, all divergences are improved compared with the divergences obtained using a single channel. The DIV_AVG calculated using 20

	Silence	*ae*	*f*	*iy*	*ao*	*sh*	*s*	*uw*
Silence	0.00	105.04	154.55	170.11	47.79	174.31	64.55	90.61
ae	105.04	0.00	18.54	148.93	95.37	35.31	239.37	33.36
f	154.55	18.54	0.00	116.10	66.38	25.73	170.23	20.92
iy	170.11	148.93	116.10	0.00	19.78	54.47	59.68	55.81
ao	47.79	95.37	66.38	19.78	0.00	48.52	15.56	32.65
sh	174.31	35.31	25.73	54.47	48.52	0.00	157.20	4.12
s	64.55	239.37	170.23	59.68	15.56	157.20	0.00	98.54
uw	90.61	33.36	20.92	55.81	32.65	4.12	98.54	0.00
Mean score	115.28	96.56	81.78	89.27	46.58	71.38	115.02	48.00

Table 6.11: Divergence scores of different phonemes using 20 OFBCs from the cheek and lower lip channels, where $\omega = 140$ Hz.

OFBCs becomes 83.0.

6.7 Summary

In this chapter, the spectral feature selection process conducted in this work was presented. The quality of features was measured using a divergence score. Two spectral feature extraction schemes, non-overlapping and overlapping frequency band, were compared. The results show that the overlapping frequency band is better than the non-overlapping frequency band to distinguish phonemes. The cheek channel yields the best average divergence score and this is followed by the lower lip channel. The chin channel is the worst. A configuration, using 20 OFBCs extracted from the cheek and lower lip channels with a bandwidth of 140 Hz, was found to yield better performance than a single channel and thus chosen for the rest of this work. In the next chapter, the speech synthesis results using this configuration will be presented.

Chapter 7

Results

7.1 Introduction

The previous chapter showed that using overlapping frequency band coefficients (OFBCs) had advantages over non-overlapping frequency band coefficients (NOFBCs). In this work, two additional features were employed, the root mean square amplitude (RMSA) and zero-crossing rate (ZCR). Moreover, the analysis in Chapter 6 showed that the cheek and lower lip channels can achieve a better average divergence score than the chin channel. In this chapter, speech synthesis results, obtained using OFBC, RMSA, and ZCR as features and two SEMG channels including the cheek and lower lip, will be presented.

This chapter begins by describing the experimental data set used. This is followed by experimental results obtained to find the most suitable SEMG frame size. The classification performance of the neural network is analyzed. The performance of an error correction technique, used to correct the classification error of the neural network, is then presented. Synthesis results for words are then described and a summary is given in the last section.

7.2 Experimental data sets

Data sets were recorded in twenty-second duration, during which a speaker repeatedly pronounced a given phoneme or word. Each data set contains two twenty-second SEMG signals (recorded from the cheek and lower lip) and one twenty-second speech signal.

7.2.1 Training phoneme set

The phoneme set described in Section 6.3 was chosen, four data sets for each phoneme were used for the training of neural network and results in twenty-eight data sets in total. SEMG signals from each channel were blocked into frames every 22.5 ms. As a result, the number of SEMG frames for each channel is:

$$28 \text{ data sets} \times \lfloor \frac{20 \text{ sec}}{22.5 \text{ ms}} \rfloor = 24864 \text{ frames}.$$

For each SEMG frame, 10 OFBCs, one RMSA and one ZCR were extracted, thus for both channels, each SEMG feature vector contains contains 20 OFBCs, 2 RMSAs and 2 ZCRs. As a result, there were a total of 24864 vectors used to train the neural network.

7.2.2 Testing phoneme set

The testing phoneme sets contains one data set for each phoneme, and results in seven data sets. SEMG signals from each channel were blocked into frames every 22.5 ms. The number of SEMG frames for each channel is:

$$7 \text{ data sets} \times \lfloor \frac{20 \text{ sec}}{22.5 \text{ ms}} \rfloor = 6216 \text{ frames}.$$

SEMG feature vectors containing 20 OFBCs, 2 RMSAs and 2 ZCRs were extracted in the same way as the training phoneme set. As a result, there were 6216 SEMG

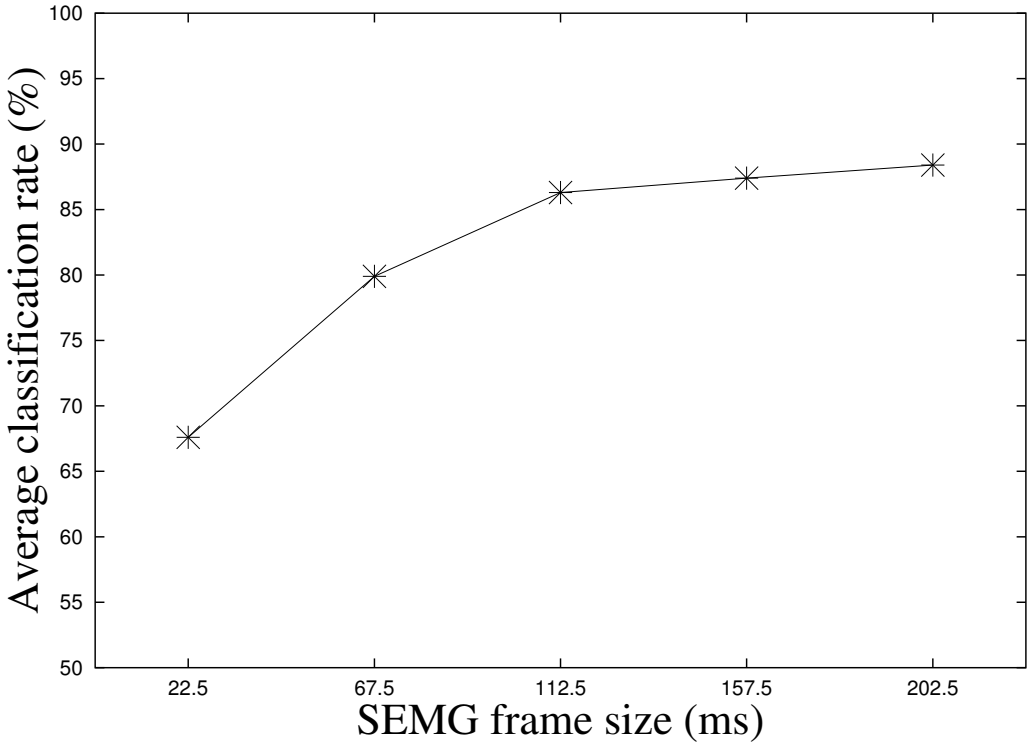


Figure 7.1: Average classification rate for different SEMG frame sizes using 20 OFBCs, 2 RMSAs, 2 ZCRs.

feature vectors in total. These vectors were used to evaluate the classification performance of the neural network. Speech signals are recorded concurrently and used as a reference to label the SEMG signals for performance evaluation.

7.2.3 Testing word set

The word used for testing are *shaw, she, ash, shoe, see, saw, fee* and *off*. The phonetic transcriptions are formed by concatenating the training phonemes. The testing word set contains one data set for each word, and results in eight data sets. SEMG feature vectors containing 20 OFBCs, 2 RMSAs and 2 ZCRs were extracted in the same way as the training phoneme set. The testing word set was used to evaluate the speech synthesis performance.

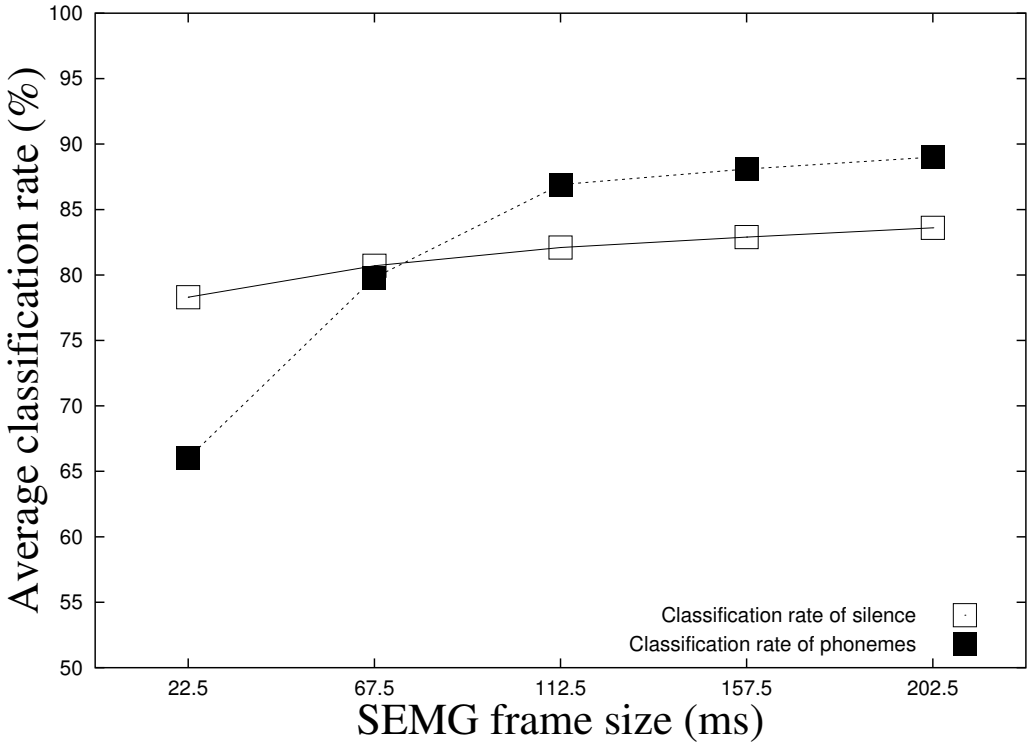


Figure 7.2: Average classification rate of silence and phonemes for different SEMG frame sizes using 20 OFBCs, 2 RMSAs, 2 ZCRs.

7.3 SEMG frame size

To find the SEMG frame size that balances the trade-off between the time and frequency resolution, SEMG features vectors were extracted for different frame sizes. The neural network was trained under different frame sizes and classification performance was evaluated using the testing phoneme set. In this work, SEMG frame sizes from 22.5 ms to 202.5 ms with a step of 45 ms were analyzed. As the feature vectors contains 20 OFBCs, 2 RMSAs, and 2 ZCRs, a three-layer neural network with 24 input nodes, 24 hidden nodes and 8 output nodes were chosen. One third of the training data was used as a cross-validation set.

The average classification rates for SEMG frame sizes from 22.5ms to 202.5ms are shown in Figure 7.1. A clear trend can be seen in this figure: the classification

rate is higher for larger SEMG frame sizes and becomes saturated for frame sizes larger than 112.5 ms. Because smaller frame size gives better time resolution, a frame size of 112.5 ms is chosen for further experiments despite larger frame size giving a slightly higher classification rate.

An interesting result is shown in Figure 7.2, which shows the classification rate of silence and phonemes separately for different SEMG frame sizes. One can see that the classification rates for silence are almost the same for all frame sizes, however, the classification rate for phonemes increases with the frame size and becomes saturated for frame sizes larger than 112.5 ms. SEMG signals for silence have similar characteristics, e.g. low amplitude, low response of frequency, these characteristics can be classified using either small or large frame sizes. This may explain why the classification rate curve of silence is flat. On the other hand, the classification rate curve for phonemes suggests that larger frame size can achieve better classification rate since a larger SEMG frame contains more information related to the muscle contraction when speaking. Thus larger frame sizes are able to capture variabilities among different phonemes.

7.4 Neural network classification

7.4.1 Number of hidden nodes

To analyze the effects of varying the number of hidden nodes, using 20 OFBCs, 2 RMSAs, and 2 ZCRs as features, the classification performance of the neural network was analyzed using the testing phoneme set. Figure 7.3 shows the average classification rate for different numbers of hidden nodes. It can be seen that the number of hidden nodes has little impact on the result. Hidden nodes ranging from 8 to 24 give almost the same classification rate, and 24 gives a slightly better rate. As a result, the number of hidden nodes was chosen as the number of input nodes in this work.

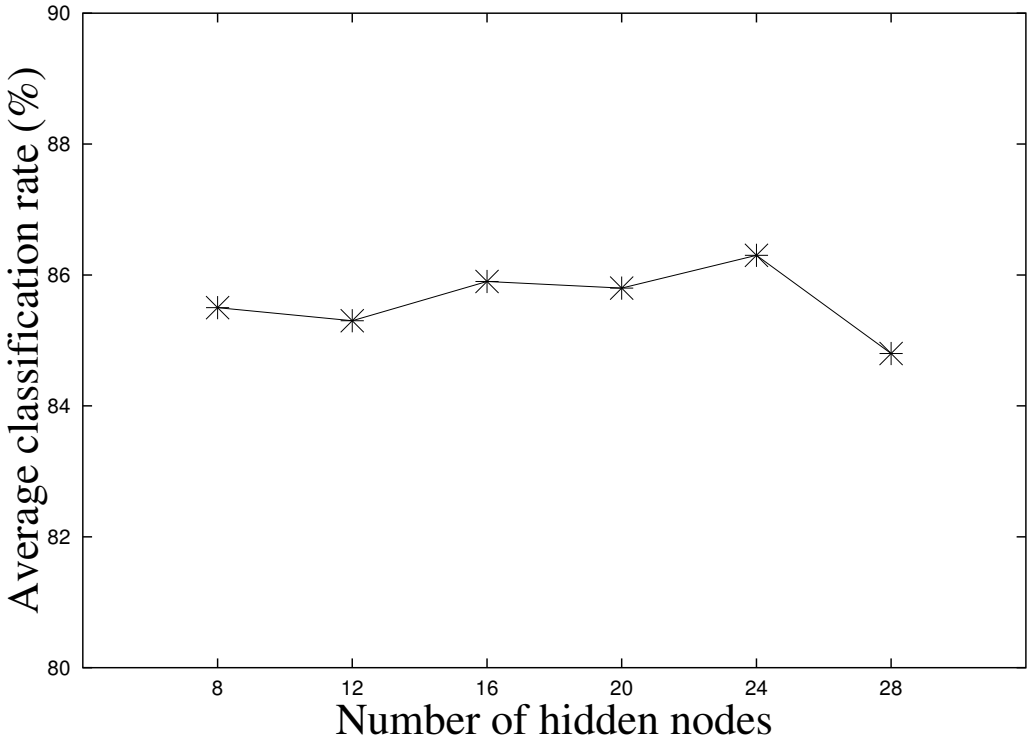


Figure 7.3: Average classification rate for different number of hidden nodes using 20 OFBCs, 2 RMSAs, 2 ZCRs.

7.4.2 Single channel

To further analyze the correlation between sensor position and performance, instead of using 20 OFBCs, 2 RMSAs, and 2 ZCRs extracted from both SEMG channels, the neural network was trained and tested using SEMG features extracted from a single SEMG channel, i.e. 10 OFBCs, 1 RMSA, and 1 ZCR. An SEMG frame size of 112.5 ms was used.

Table 7.1 shows the confusion matrix for classification using 10 OFBCs, 1 RMSA, and 1 ZCR extracted from the cheek, and the results obtained for lower lip channel are shown in Table 7.2. In these tables, the columns show the neural network classified labels and the rows represent the true labels. The average classification rates for the cheek and lower lip channels are 74.3% and 60.4% respectively.

Classified phoneme label	True phoneme label							
	Silence	*ae*	*f*	*iy*	*ao*	*sh*	*s*	*uw*
Silence	76.6%	0.7%	0.0%	1.9%	3.3%	0.0%	5.3%	0.5%
ae	1.6%	91.3%	9.6%	0.2%	0.0%	3.2%	0.4%	0.9%
f	1.7%	2.8%	77.9%	0.2%	0.5%	13.5%	0.0%	6.1%
iy	2.3%	3.0%	0.0%	74.2%	11.2%	1.7%	0.7%	2.6%
ao	5.3%	1.1%	0.2%	16.0%	71.6%	0.7%	5.3%	0.7%
sh	2.1%	0.4%	9.0%	0.0%	0.7%	51.7%	0.0%	26.6%
s	9.0%	0.7%	0.0%	6.0%	12.2%	0.4%	88.3%	0.2%
uw	1.4%	0.0%	3.3%	1.5%	0.5%	28.8%	0.0%	62.4%

Table 7.1: Confusion matrix showing the classification performance using 10 OFBCs, 1 RMSA, and 1 ZCR extracted from the cheek, the SEMG frame size is 112.5 ms. The average classification rate is 74.3%.

Classified phoneme label	True phoneme label							
	Silence	*ae*	*f*	*iy*	*ao*	*sh*	*s*	*uw*
Silence	79.7%	1.1%	0.0%	0.4%	0.7%	0.0%	1.6%	0.0%
ae	2.7%	50.2%	16.2%	8.5%	23.9%	16.1%	8.7%	2.4%
f	1.3%	15.7%	63.7%	0.4%	22.2%	1.3%	1.2%	0.9%
iy	0.9%	0.4%	0.0%	62.6%	0.0%	11.8%	0.0%	5.9%
ao	4.0%	9.1%	15.6%	1.0%	30.0%	1.9%	3.0%	1.6%
sh	2.2%	0.0%	0.0%	15.4%	0.0%	53.2%	0.0%	30.1%
s	8.0%	23.5%	4.3%	2.7%	13.4%	0.6%	85.5%	0.9%
uw	1.2%	0.0%	0.2%	9.0%	9.8%	15.1%	0.0%	58.2%

Table 7.2: Confusion matrix showing the classification performance using 10 OFBCs, 1 RMSA, and 1 ZCR extracted from the lower lip, the SEMG frame size is 112.5 ms. The average classification rate is 60.4%.

One can observe that the cheek channel provides more discriminative information for SEMG frame classification for the phoneme set used in this work.

From these two tables, the results are consistent with the results obtained in the spectral feature assessment of SEMG signals in Chapter 6. For example, on average the cheek channel is better than the lower lip channel and the lower lip channel is better than the cheek for separating silence from other phonemes. The cheek is better than the lower lip channel for distinguishing phonemes except *sh*. Comparing Table 7.1 and 7.2, it can be seen that that confusion between phoneme *sh* and *uw* exists in both cases. An average misclassification rate of 27.7% was

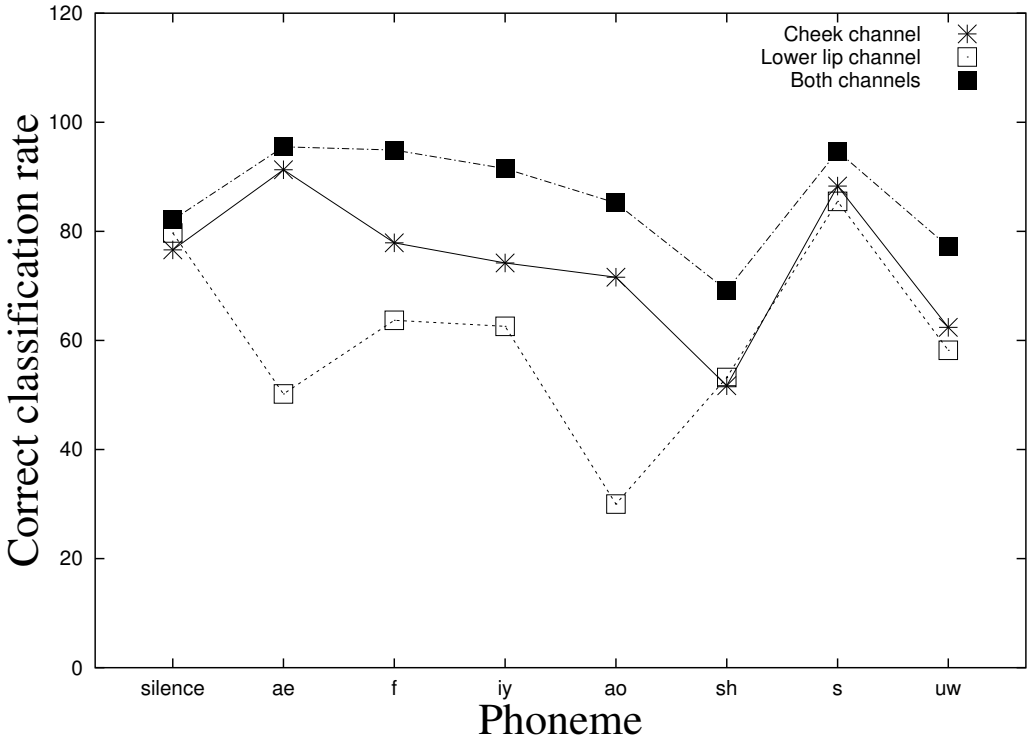


Figure 7.4: Correct classification rate of silence and each phoneme using the cheek, the lower lip, and both channels respectively.

found using the cheek channel, and 22.6% misclassification rate was found using the lower lip channel.

7.4.3 Classification using both channels

Figure 7.4 shows the correct classification rates for silence and each phoneme using the cheek, the lower lip, and both channels respectively. This figure shows that, compared with using a single channel, correct classification rates of all phonemes and silence are improved when both channels are used.

Table 7.3 shows the classification results using 20 OFBCs, 2 RMSAs, and 2 ZCRs. The average classification rate is 86.3%. From this table and Figure 7.4,

Classified phoneme label	True phoneme label							
	Silence	*ae*	*f*	*iy*	*ao*	*sh*	*s*	*uw*
Silence	82.1%	1.8%	0.0%	1.0%	1.7%	0.2%	1.4%	0.2%
ae	2.1%	95.5%	5.1%	0.0%	0.0%	1.9%	0.2%	0.5%
f	1.5%	1.7%	94.9%	0.0%	1.2%	1.3%	0.0%	0.2%
iy	1.2%	0.0%	0.0%	91.5%	0.7%	0.0%	1.2%	0.0%
ao	4.0%	0.6%	0.0%	4.4%	85.2%	0.6%	2.3%	1.2%
sh	2.1%	0.0%	0.0%	0.0%	0.2%	69.2%	0.0%	20.0%
s	5.3%	0.4%	0.0%	2.9%	9.3%	0.4%	94.7%	0.7%
uw	1.7%	0.0%	0.0%	0.2%	1.7%	26.4%	0.2%	77.2%

Table 7.3: Confusion matrix showing the classification performance using 20 OFBCs, 2 RMSAs, and 2 ZCRs extracted from the cheek and lower lip channels, the SEMG frame size is 112.5 ms. The average classification rate is 86.3%.

one can see that, although the correct classification rates of all phonemes and silence are improved when using both channels, there is no improvement in separating phoneme *sh* from *uw*. The average misclassification rate between these two phoneme is 23.2% when both channels are used, however, a misclassification rate of 22.6% is obtained when the lower lip channel is used, this rate is even slightly lower than for two channels. Moreover, from Figure 7.4, it is found that the correct classification rates of phoneme *sh* and *uw* are the worst among the seven phonemes. These results indicate that the SEMG signals collected from the cheek and lower lip may be inadequate for separating phoneme *sh* from *uw*, as the degree of lip-rounding for pronouncing these two phoneme are similar.

Table 7.3 shows that there are still 17.9% silence SEMG frames being misclassified to other phonemes when using both SEMG channels. This error is introduced by the fact that SEMG activities exist prior or posterior to acoustic signals (see Figure 7.5). In the training phoneme set, some SEMG frames prior (region A2 in Figure 7.5) or posterior (region A3 in Figure 7.5) to acoustic signals are labeled as silence, and as these kinds of SEMG frames are actually associated with rich muscle activities, the classifier may be confused, as SEMG frames in regions A1 and A4 in Figure 7.5, which are associated with less muscle activities, are also labeled as silence and thus form a one-to-many situation.

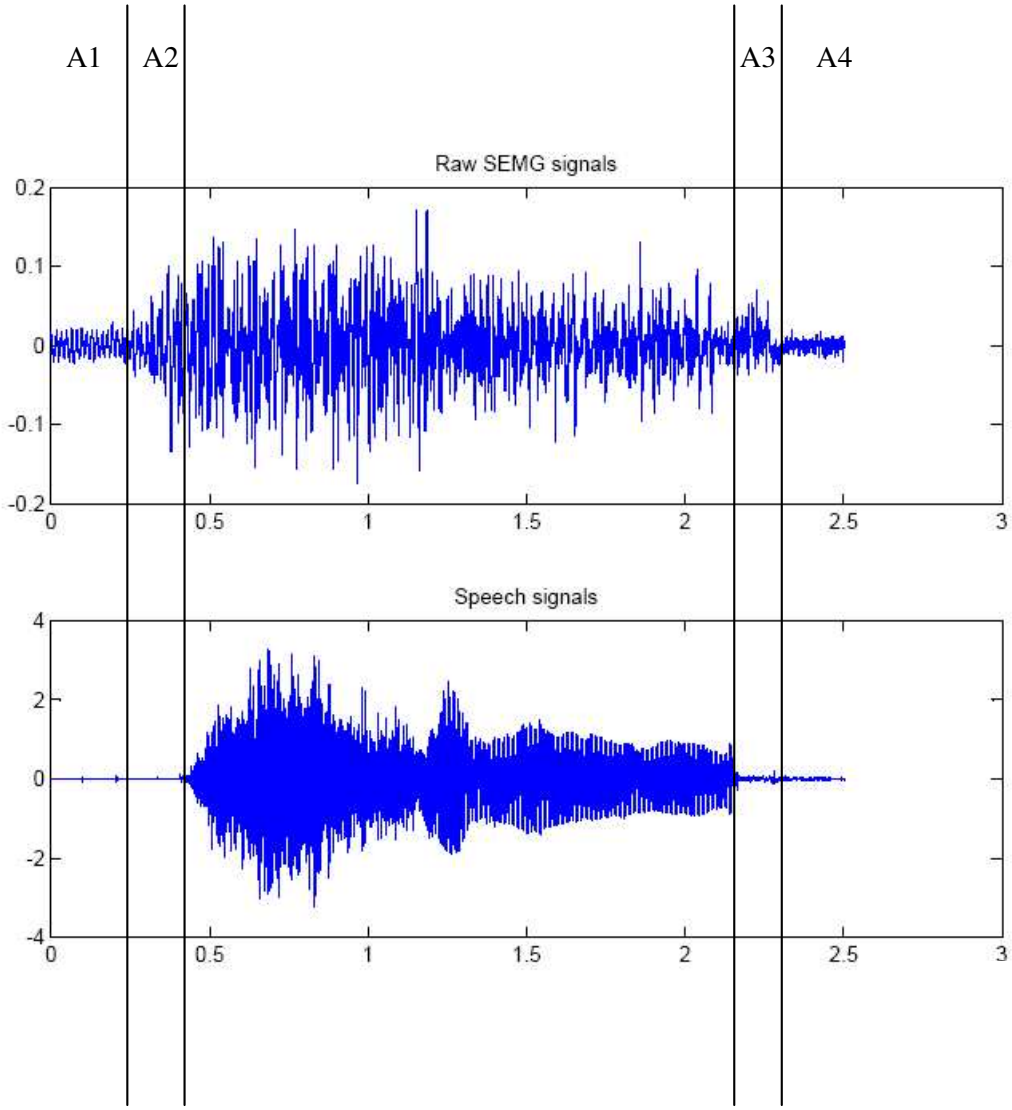


Figure 7.5: SEMG activities prior and posterior to speech.

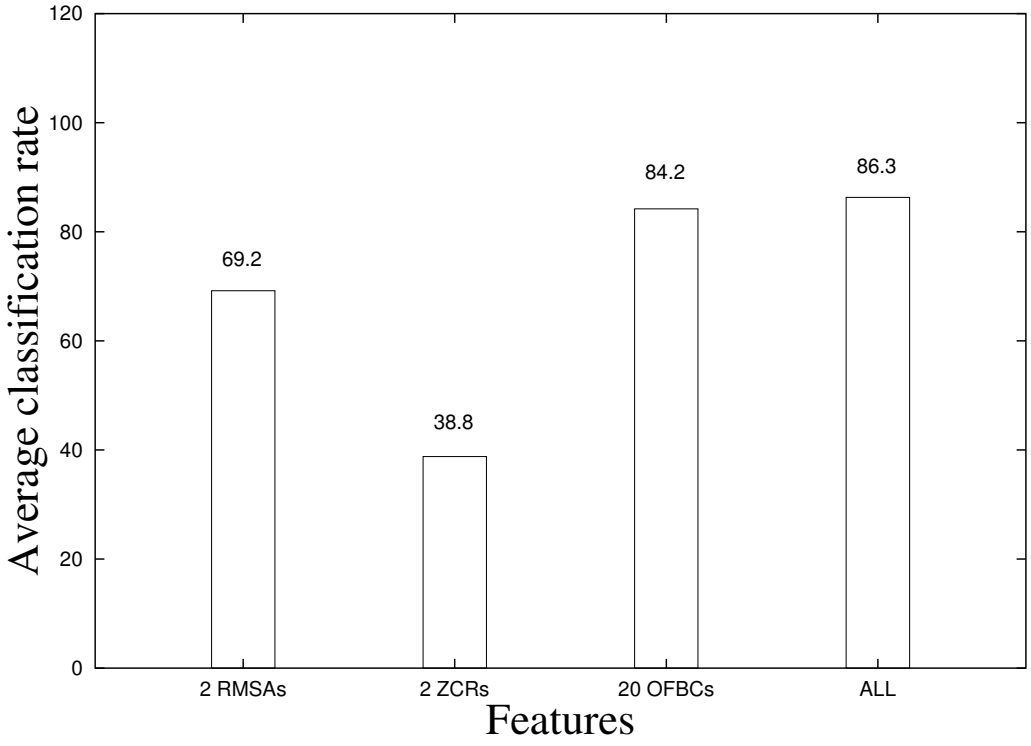


Figure 7.6: Average classification rates using different features extracted from both SEMG channels. The feature labeled "ALL" means using 20 OFBCs, 2 RMSAs, and 2 ZCRs.

Classified phoneme label	True phoneme label							
	Silence	*ae*	*f*	*iy*	*ao*	*sh*	*s*	*uw*
Silence	87.0%	0.4%	0.0%	0.4%	1.0%	0.0%	2.3%	3.3%
ae	1.8%	99.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
f	2.1%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%
iy	1.6%	0.0%	0.0%	94.8%	0.0%	0.0%	0.0%	0.0%
ao	1.6%	0.0%	0.0%	3.8%	99.0%	0.0%	0.0%	0.0%
sh	3.2%	0.0%	0.0%	0.0%	0.0%	96.3%	0.0%	0.0%
s	1.4%	0.0%	0.0%	1.0%	0.0%	0.0%	97.7%	0.0%
uw	1.3%	0.0%	0.0%	0.0%	0.0%	3.7%	0.0%	96.7%

Table 7.4: Confusion matrix after applying error correction to the produced sequence of speech feature indices. The average classification rate is 96.4%.

Figure 7.6 shows the classification results for different features using both SEMG channels. Using 2 ZCRs and 2 RMSAs achieve the worst accuracy and this result is consistent with previously obtained results [ST85, MO86]. In these studies, the authors tried to classify SEMG frames using similar features and also achieved poor results. In their work, the number of crossing threshold [ST85] and average amplitude [MO86] were used as features. Figure 7.6 also shows that OFBC is the best feature among these three kinds of features, using 20 OFBCs can achieve almost the same classification accuracy as using all features.

7.5 Phonetic sequence smoothing

7.5.1 Data set

Using 20 OFBCs, 2 RMSAs, and 2 ZCRs as features, SEMG signals in the testing phoneme set were classified into sequences of speech feature indices. Since there were seven phonemes, seven sequences were produced, each sequence containing 888 indices. As indicated in Table 7.3, misclassifications exist in the classified sequences of speech feature indices. A hybrid smoothing technique, described in Section 5.4.4, was applied to correct these classification errors.

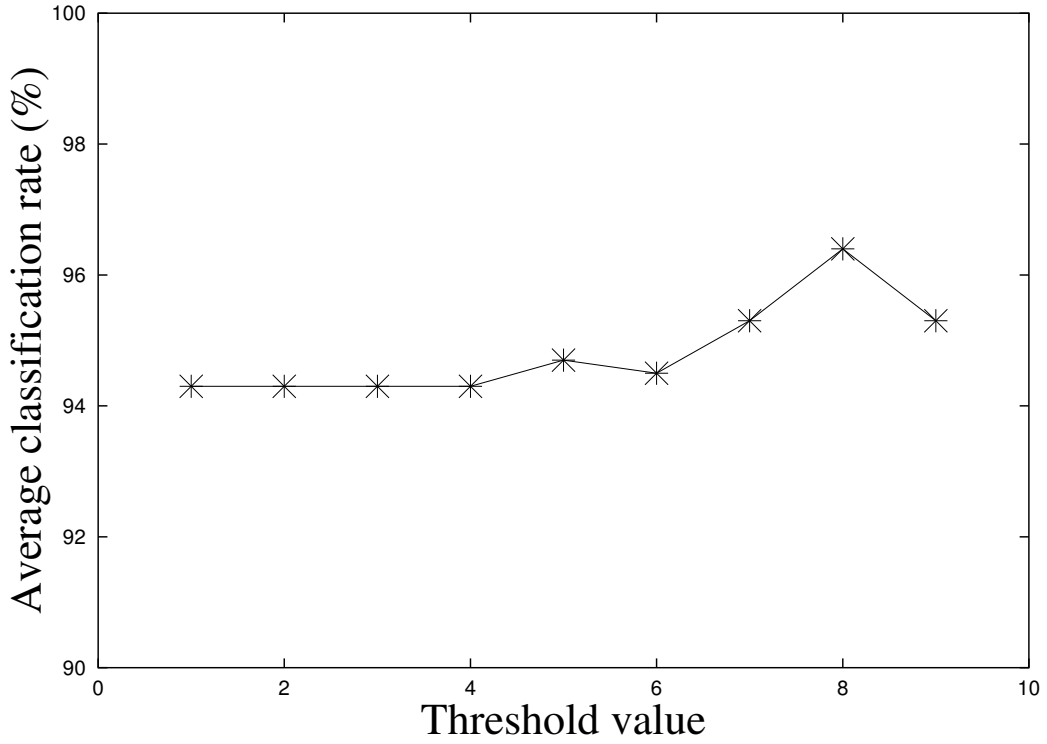


Figure 7.7: Average classification rates after smoothing for different threshold values in the majority filtering process.

7.5.2 Smoothing of classifications

Figure 7.7 shows the average classification rates after smoothing for different threshold values in the majority filter process. The classification rate is similar for different threshold values, and a value of 8 achieves the best classification rate. Thus this value was chosen for further experiments.

Table 7.4 shows the results obtained after applying the hybrid smoothing using a threshold of 8 for the majority filter process. An average classification rate of 96.4% was achieved. A comparison of classification rates before and after smoothing is shown in Figure 7.8. This figure shows that the classification rates for each phoneme and silence are improved with the smoothing technique. In particular, phonemes *ao*, *sh* and *uw* exhibit the greatest improvement.

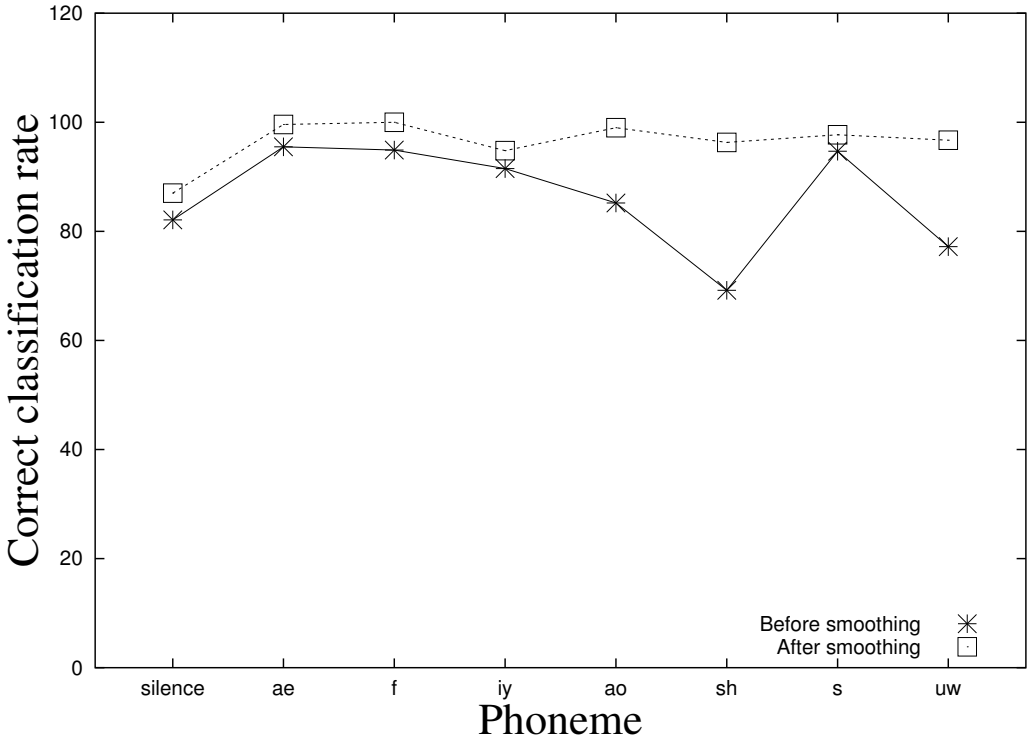


Figure 7.8: Comparison of correct classification rate before and after smoothing.

7.6 Speech synthesis

The testing word set was used for evaluating the performance of a speech synthesis system. Using 20 OFBCs, 2 RMSAs, and 2 ZCRs as features, the SEMG signals for each data sample were classified into a sequence of speech feature indices. After performing error correction, speech waveforms for each data set were then synthesized using the concatenative method (see Section 5.3.4).

Table 7.5 shows the synthesis results. 92.9% of the words are synthesized correctly. A word is regarded as synthesized correctly if the phonetic transcriptions of the synthesized word is correct, e.g. a synthesized word *off* is regarded as synthesized correctly if it's phonetic transcriptions is a phoneme *ao* followed by a phoneme *f*.

Figure 7.9 and 7.10 show the sub-sequences of speech feature indices before

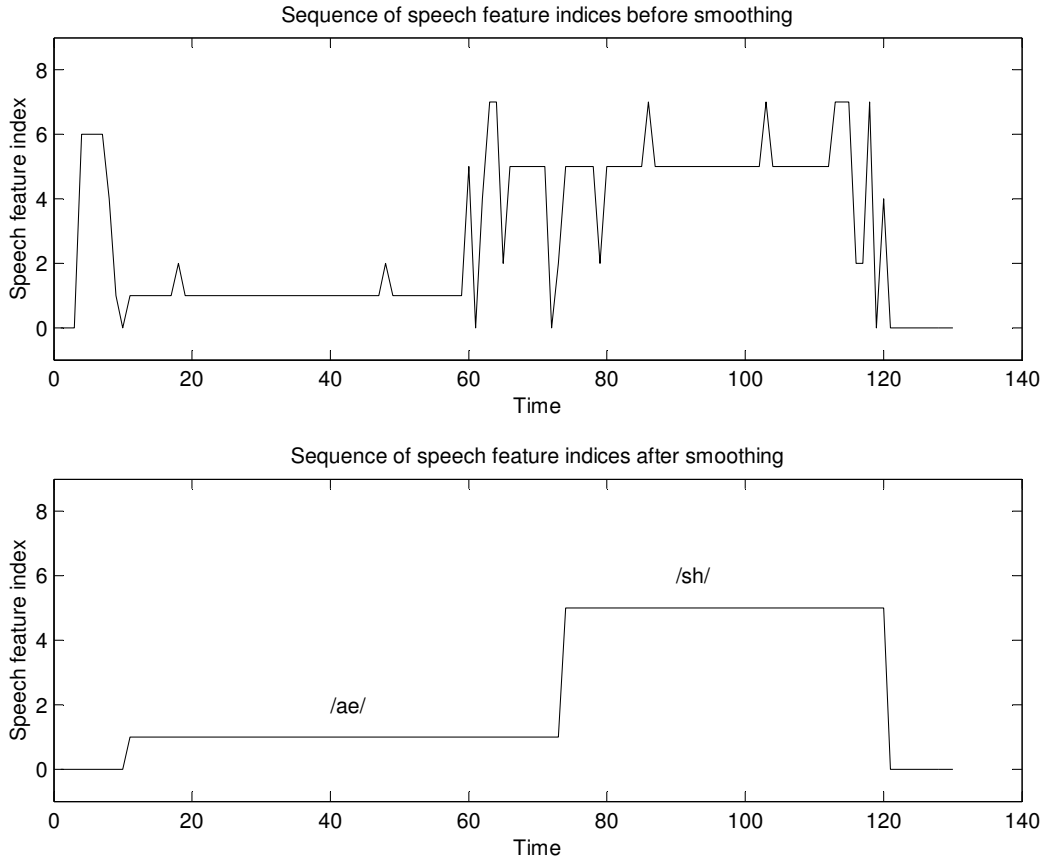


Figure 7.9: A sub-sequence of speech feature indices before and after smoothing for word *ash*.

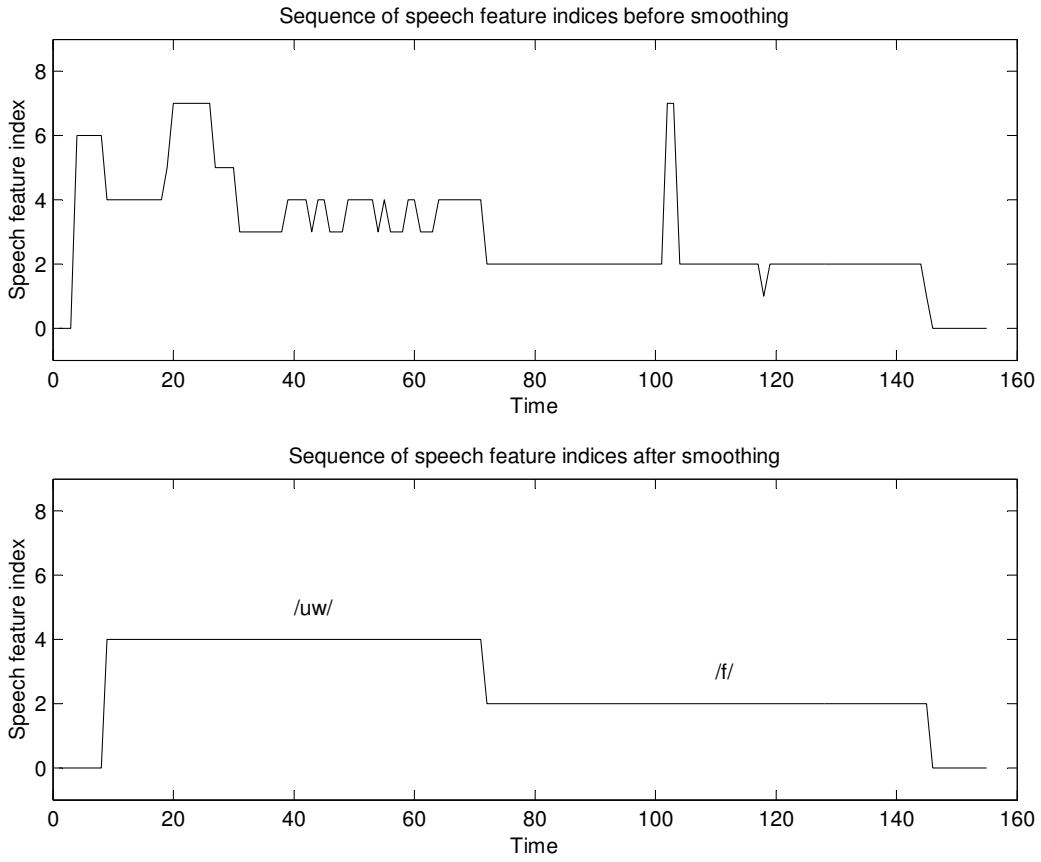


Figure 7.10: A sub-sequence of speech feature indices before and after smoothing for word *off*.

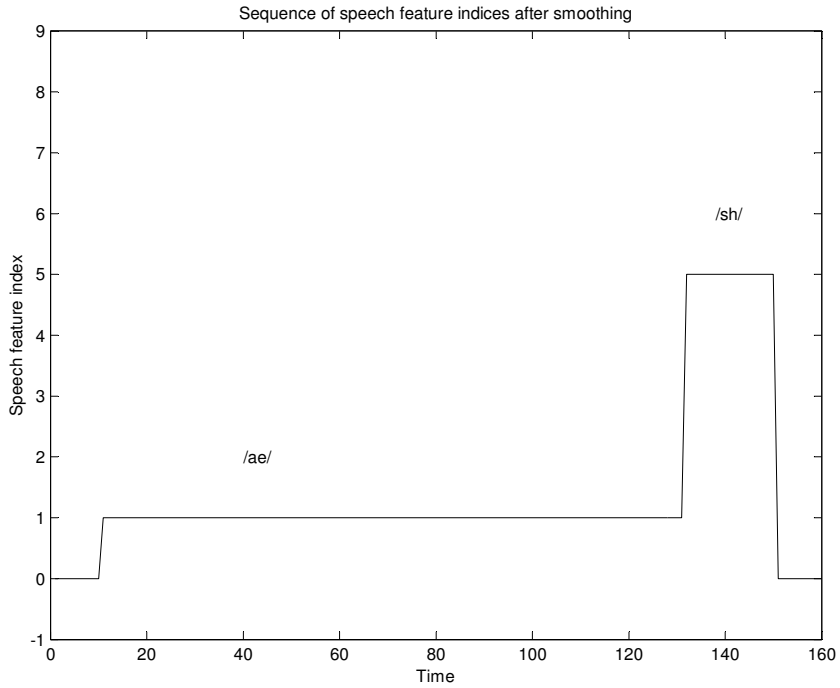


Figure 7.11: An example showing the sub-sequence of speech feature indices after smoothing for word *ash*.

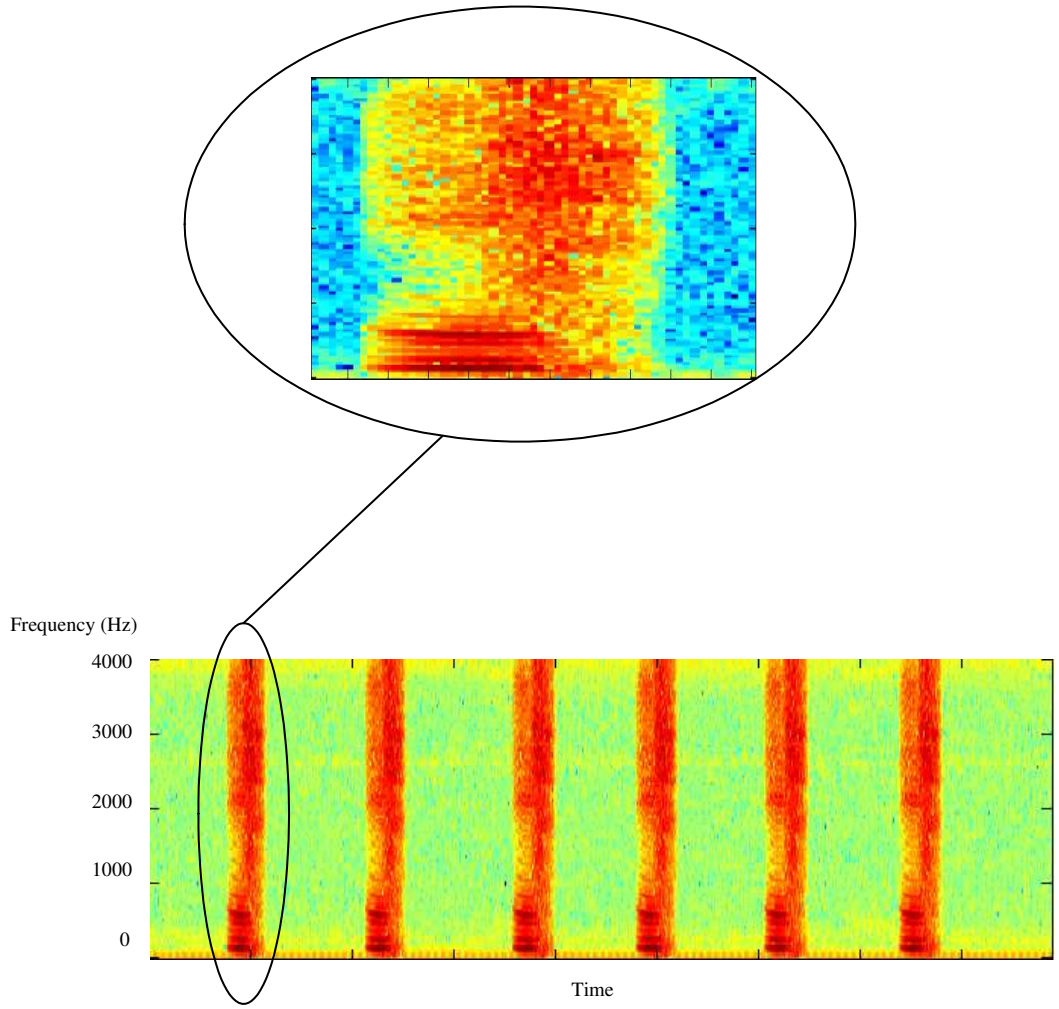


Figure 7.12: Spectrogram of the synthesized speech of six repetition of the word *ash*.

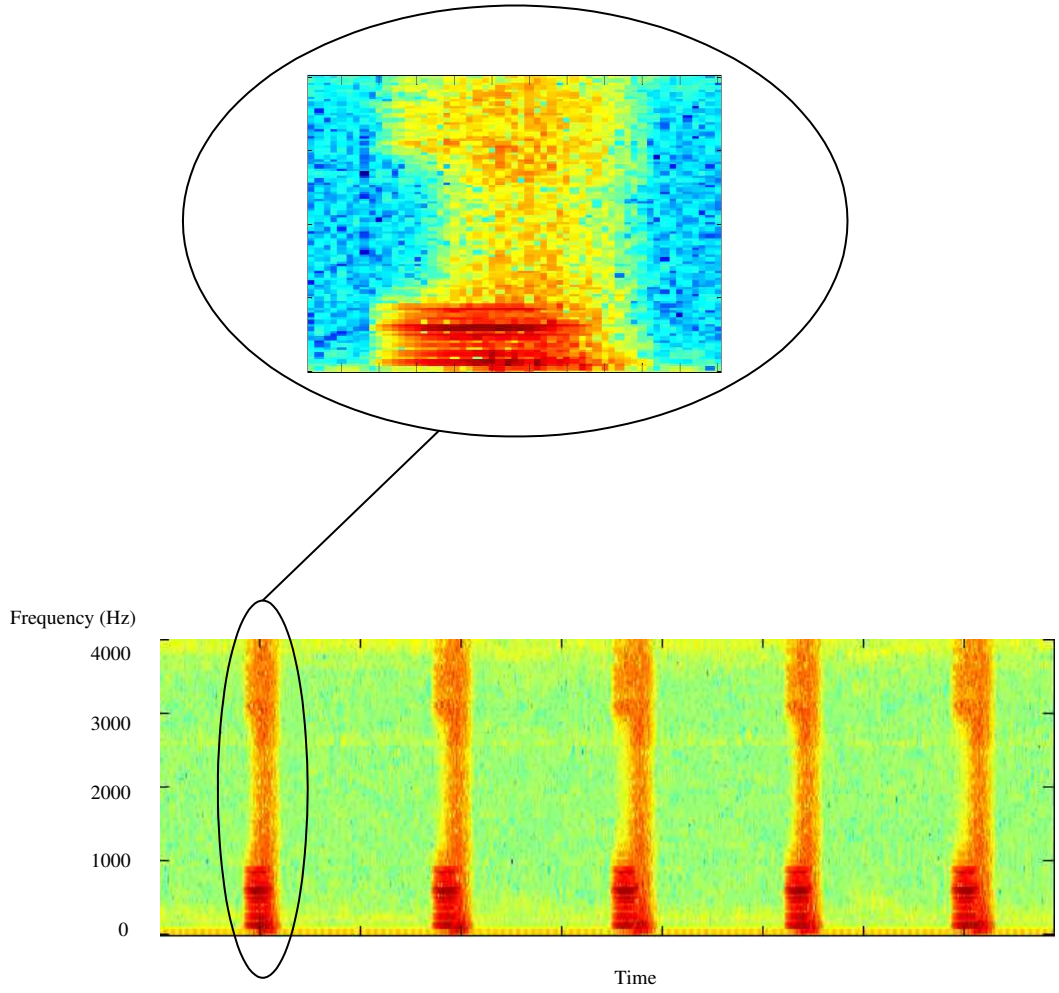


Figure 7.13: Spectrogram of the synthesized speech of five repetition of the word *off*.

Words	Number of words synthesized	Number of words synthesized correctly
she	5	5
ash	6	6
shaw	6	6
see	4	4
saw	6	3
shoe	4	4
fee	6	6
off	5	5
Total	42	39

Table 7.5: Synthesis results for words

and after smoothing for two synthesized instances of word *ash* and *off* respectively. Note that there are many errors in the sequences produced by the neural network (i.e. before smoothing). After applying the smoothing technique, clear phonetic transcriptions for both words are obtained. Thus, both words are regarded as synthesized correctly, since they can produce the correct phonetic transcriptions. This work focuses on generating correct phonetic transcriptions for input SEMG signals. As this is directly correlated to the intelligibility of the synthesized speech. Once a correct phonetic transcription is obtained, more sophisticated speech synthesis techniques can be used to generate the speech waveform with better quality and intelligibility.

Although the duration of phonemes in the synthesized words using the presented technique may be incorrect, the intelligibility of the synthesized words is not affected. For example, Figure 7.11 shows a smoothed sub-sequence of speech feature indices for word *ash* with longer *ae*, the intelligibility of the synthesized speech will not be affected, as a correct phonetic transcription is obtained, i.e. a phonetic transcription of an *ae* followed by a *sh* can be heard. Moreover, the duration of each phoneme can be adjusted by the speech synthesizer. In the above case, a speech synthesizer could potentially shorten the duration of phoneme *ae* and extend the length of phoneme *sh*. If the duration of both phonemes are too long, the synthesizer can

also shorten the durations of both phonemes. The spectrograms of the synthesized instances of *ash* and *off* are shown in Figures 7.12 and 7.13.

7.7 Summary

Experimental results were presented in this chapter. The results showed that an SEMG frame size of 112.5 ms achieves a good balance between time and frequency resolution. Using two SEMG channels and features as 20 OFBCs, 2 RMSAs, and 2 ZCRs, an average classification of 86.3% was obtained. This result can be improved to 96.4% after applying the hybrid smoothing technique. Experimental evaluations based on the synthesis of eight words showed that on average, 92.9% of the words could be synthesized correctly.

Chapter 8

Conclusion

The main objective of this work was addressing the feasibility of unlimited vocabulary continuous speech synthesis from SEMG signals. Several subproblems and original contributions made to address them were studied.

Spectral feature selection

Two kinds of spectral features, NOFBCs and OFBCs, were compared in this work. Using a divergence metric, the assessment showed that overlapping frequency band coefficients achieve a higher divergence score, and the separability of each frequency band increases with the bandwidth. The classification results showed that spectral features alone can achieve nearly the same performance as all features combined. This showed that spectral features were important despite temporal features being widely used in previously proposed systems, and that the results can be improved by carefully analyzing and selecting the appropriate spectral features. This work showed that OFBCs are excellent features for the analysis of SEMG signals.

SEMG frame size

The effects of the SEMG frame size were investigated. Classification performances of the neural network for different SEMG frame sizes were analyzed. Experimental results showed that classification accuracy increases with the SEMG frame size and tends to saturate for frame size larger than 112.5 ms. This frame size was chosen in

this work as it can balance the requirement of frequency and time resolution. The experimental results showed a strong relationship exist between SEMG frame size and accuracy of phoneme frame classification, suggesting that it should be chosen carefully. The effects of the frame size should be considered when new phonemes are appended to the phoneme set.

SEMG sensor positioning

SEMG sensor positioning is a critical element that affect the classification performance. The divergence test on spectral features showed that, the cheek channel can achieves the best divergence score, then the lower lip channel and the chin channel has the lowest score. The experimental results from a neural network classification also showed that the cheek channel can achieve higher accuracy, and the performance is much better when more SEMG channels are used. However, phonemes having similar degrees of lip-rounding are difficult to distinguish using SEMG signals, despite the fact that their acoustic signals may be totally different. For example, using SEMG signals from the cheek and lower lip is hard to distinguish phoneme *sh* from *uw*. This result suggests that more SEMG channels from various positions should be used.

Error correction

Generating correct phonetic transcriptions, i.e. sequences of speech feature indices, was one of the major problems studied in this work. As the accuracy of the sequence of speech feature indices is directly correlated to the intelligibility of the synthesized speech, a hybrid smoothing technique, which was developed that assumes mid-term stationary of the speech signals. Experimental results showed that it is capable of enhancing the accuracy of the produced sequences of speech feature indices to some degree, as shown in Chapter 7, the accuracy is improved from 86.3% to 96.4%, a 10.1% improvement was achieved.

Approach	Work	Number of subjects	Word set	Accuracy	SEMG channels	Error correction	Training unit	Recognized unit	Recognize untrained unit
WS	this thesis	1	8 Wds + 7 Phs	92.9%	2	Y	Phoneme frame	Word	Y
IWR	[MO86]	1	6 Wds	$< 70\%$	4	N	Word	Word	N
IWR	[CEHL02b]	2	10 Wds	83%	5	N	Word	Word	N
IWR	[MZ04]	10	10 Wds	63.7%	3	N	Word	Word	N
IWR	[JLA03]	3	6 Wds	92%	2	N	Word	Word	N
IPR	[KKAB04]	3	5 Vws	88%	3	N	Phoneme	Phoneme	N
IPR	[MHS03]	3	5 Vws	94.7%	3	N	Phoneme	Phoneme	N
IPR	[JB05]	2	41 Vws	33%	2	N	Phoneme	Phoneme	N
PS	[ST85]	3	5 Vds	64%	3	N	Phoneme frame	Phoneme	N

Table 8.1: A comparison between this work and previous work. WS - word synthesis, IWR - isolated word recognition, IPR - isolated phoneme recognition, PFR - phoneme frame recognition, PS - phoneme synthesis, Wds - words, Vws - vowels. Previous work was reviewed in Section 2.5.

SEMG word synthesis

This work presented an SEMG-based speech synthesizer for a small phoneme and word set. A comparison between this work and previous work (reviewed in Section 2.5) is shown in Table 8.1. The major differences are that this work addressed the problem of speech synthesis from SEMG signals, the recognition model was built from phoneme frames, several sub-problems such as error correction and transition smoothing between speech segments were addressed. Although the training process involved only phonemes, it was demonstrated that words can be synthesized by using a frame-based feature extraction and conversion approach. Using a three-layer neural network with 24 hidden nodes, the experimental results showed that the neural network can classify the SEMG frames at an accuracy of 86.3% and this was improved to 96.4% by applying the error correction process, and 92.9%

of words can be synthesized correctly and demonstrated the feasibility of unlimited vocabulary speech synthesis from SEMG signals.

8.1 Future work

This work introduced a new approach to speech recognition from SEMG signals. There are many potential directions for future research directions.

8.1.1 Speech recognition techniques

The applicability of techniques from conventional speech recognition can be investigated. For example, a hybrid recognizer that integrates neural networks and HMMs [BMFK92], the neural network being used to produce observation probabilities for the HMM, could be applied. This approach is commonly used in conventional speech recognition and yields good performance.

8.1.2 SEMG sensor positioning

As shown in the experimental results, the positioning of the SEMG sensor plays a vital role in performance. In this work, three positions (the cheek, the lower lip, and the chin) were investigated and only two channels were used to perform classification. The contributions of SEMG channels to the discrimination of phonemes can be analyzed for other facial positions, and it is believed that performance can be further improved by using additional SEMG channels.

8.1.3 Large phoneme/word set and multiple subjects

A small phoneme and word set from a single subject was used in this work. Larger phoneme and word sets from multiple subjects can be used to further explore the feasibility of this approach and the variabilities of SEMG features for different subjects can be analyzed.

8.1.4 Potential applications

Previously proposed SEMG-based applications, such as a human-computer interface [JB05] using a limited word set was inconvenient to a user. Once a larger phoneme/word set is addressed, more SEMG-based practical systems can be implemented, such as practical speech prosthetic devices, human-computer interfaces, underwater communications and silent communication devices.

8.2 Concluding remarks

The feasibility of speech synthesis from SEMG signals was addressed in this work. Encouraging results were obtained for a small phoneme and word set using only two SEMG channels. Continuous speech synthesis from SEMG signals is a challenging and difficult task. It is hoped that the approach described in this work inspires further advances in this area and large phoneme and word sets become practical.

Appendix A

Schematic circuit diagram

The following circuit was used to obtain all of the SEMG data used in this work. The output of thie circuit was sent to a National Instruments PCI6024E PCI data acquisition card [Nat] for digitization (see Section 5.3.1).

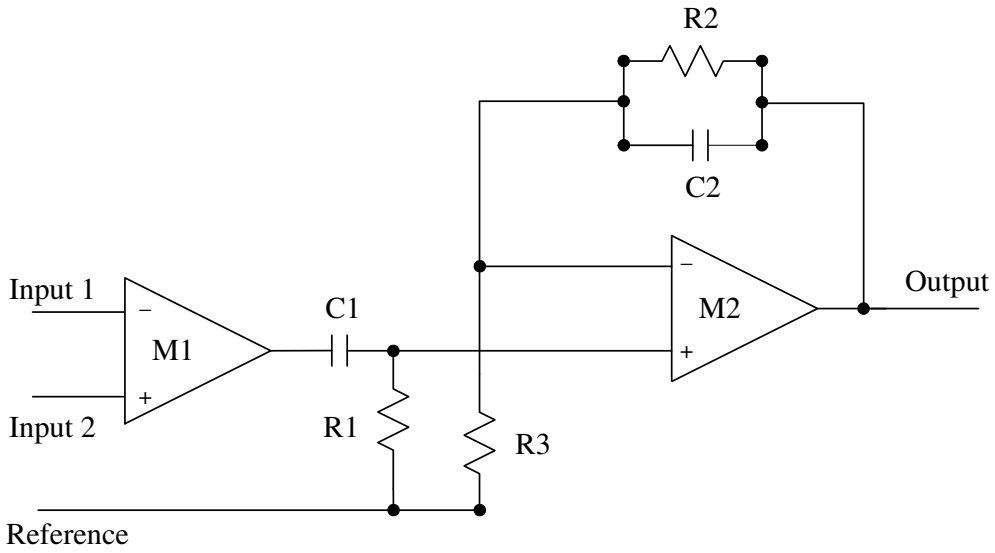


Figure A.1: Schematic circuit diagram of SEMG signal amplification. R1 = 10kΩ, R2 = 100kΩ, R3 = 1kΩ, C1 = 1μF, C2 = 3.2nF, M1: Analog Devices AD625 amplifier, M2: Analog Devices AD210 amplifier.

Appendix B

K-Means clustering algorithm

The following pseudo code describes the K-Means algorithm used in this work (see Section 5.3.2)

```
Initialize guesses for the means m(1), m(2), ..., m(N)
Set the counts c(1), c(2), ..., c(N) to zero
Set iteration to zero
When iteration < threshold
  For all datas
    Retrieve a new data x
    If distance between m(k) and x is minimum
      c(k) = c(k) + 1
      delta_m = (1/c(k))(x-m(k))
      m(k) = m(k) + delta_m
    End if
  End for
  iteration = iteration + 1
End when
```

Appendix C

Vector quantization

The vector quantization (VQ) process finds a codebook index specifying the codebook vector that best represents a given vector. The codebook vectors can be obtained by clustering a set of training vectors, and the K-Means clustering algorithm was used in this work.

Figure C.1 shows the VQ processing, for an input vector sequence $V\{v(1), v(2), v(3), ..., v(N)\}$, the VQ process calculates the vector distance between each vector in the codebook $C\{c(1), c(2), c(3), ..., c(P)\}$ and each input vector $v(n)$. The codebook index with minimum distance will be chosen as output. After vector quantization, a sequence of codebook indices $I\{i(1), i(2), i(3), ..., i(N)\}$ is produced.

In this work, the vector distance between an input vector $v(n)$ and each vector in codebook was calculated using:

$$d(v(i), c(j)) = \sum_{k=1}^{K} [v(i)(k) - c(j)(k)]^2, \tag{C.1}$$

where $v(i)(k)$ is the k-th element of the i-th input vector $v(i)$, $c(i)(k)$ is the k-th element of the j-th codebook vector $c(i)$, K is the vector length. The following pseudo code describes the vector quantization algorithm used in this work (see Section 5.3.3).

```
for p from 1 to codebook_size {
```

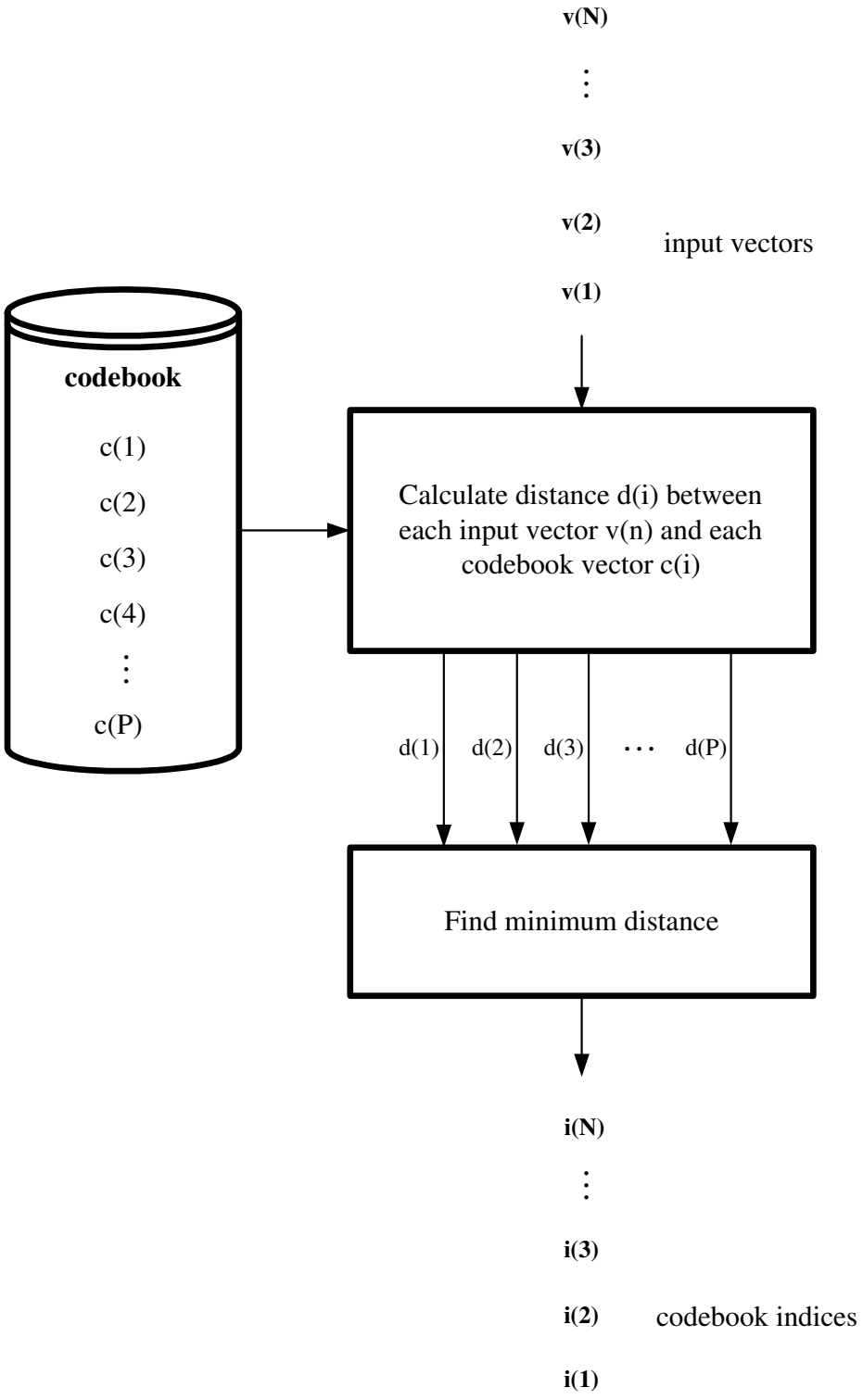


Figure C.1: Vector quantization process

```
    distance(p) = 0;
    for k from 1 to vector_length {
      temp = (v(n)(k) - c(p)(k))*(v(n)(k) - c(p)(k));
      distance(p) = distance(p) + temp;
    }
  }
  i(n) = arg min_p (distance(p));
```

In the above pseudo code, $i(n)$ is the n-th element of the output codebook indices sequence as shown in Figure C.1. Similar input vectors are clustered together in vector quantization.

Bibliography

[AHK+87] J. Allen, M.S. Hunnicutt, D. Klatt, R.C. Armstrong, and D. Pisoni. *From Text to Speech: the MITalk System*. Cambridge University Press, 1987.

[BAH05] Y. Barniv, M. Aguilar, and E. Hasanbelliu. Using EMG to Anticipate Head Motion for Virtual-Environment Application. *IEEE Transactions on Biomedical Engineering*, 52(6):1078–1093, June 2005.

[BMFK92] Y. Bengio, R.D. Mori, G. Flammia, and R. Kompe. Global Optimization of a Neural Network-Hidden Markov Model Hybrid. *IEEE Transactions on Neural Networks*, 3(2):252–259, March 1992.

[Bre92] A. Breen. Speech Synthesis Models: A Review. *Electronics & Communication Engineering Journal*, 4:19–31, February 1992.

[CEHL01] A.D.C. Chan, K. Englehart, B. Hudgins, and D.F. Lovely. Hidden Markov Model Classification of Myoelectric Signals in Speech. In *Proceedings of the 23rd Annual International Engineering in Medicine and Biology Society*, volume 2, pages 1727–1739, 2001.

[CEHL02a] A.D.C. Chan, K. Englehart, B. Hudgins, and D.F. Lovely. A Multi-Expert Speech Recognition System using Acoustic and Myoelectric Signals. In *Proceedings of the Second Joint EMBS/BMES Conference*, pages 72–73, 2002.

[CEHL02b] A.D.C. Chan, K. Englehart, B. Hudgins, and D.F. Lovely. Hidden Markov Model Classification of Myoelectric Signals in Speech. *IEEE Engineering in Medicine and Biology*, 21(5):143–146, Sept-Oct 2002.

[CH97] G.I. Chiou and J.N. Hwang. Lipreading from Color Video. *IEEE Transactions on Image Processing*, 6(8):1192–1195, 1997.

[Cho97] S.B. Cho. Neural-Network Classifiers for Recognizing Totally Unconstrained Handwritten Numerals. *IEEE Transactions on Neural Networks*, 8(1):43–53, January 1997.

[CHS+98] P. Cosi, J.P. Hosom, J. Shalkwyk, S. Sutton, and R.A. Cole. Connected Digit Recognition Experiments with the OGI Toolkit's Neural Network and HMM-based Recognizers. In *Proceedings of IEEE 4th Workshop on Interactive Voice Technology for Telecommunications Applications*, pages 135–140, 1998.

[DN93] E. Dorken and S.H. Nawab. Time-Frequency Analysis of Non-stationary Harmonic Sounds. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 3, pages 249–252, 1993.

[Dut97] T. Dutoit. *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers, 1997.

[FC86] A.J. Fridlund and J.T. Cacioppo. Guidelines for Human Electromyographic Research. *Psychophysiology*, 23(5):567–589, September 1986.

[GHK+04] E.A. Goldstein, J.T. Heaton, J.B. Kobler, G.B. Stanley, and R.E. Hillman. Design and Implementation of a Hands-Free Electrolarynx Device Controlled by Neck Strap Muscle Electromyographic Activity. *IEEE Transactions on Biomedical Engineering*, 51(2):325–332, 2004.

[GM88] A.S. Gevins and N.H. Morgan. Applications of Neural-Network (NN) Signal Processing in Brain Research. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(7):1152–1161, July 1988.

[HAH01] X.D. Huang, A. Acero, and H.W. Hon. *Spoken Language Processing*. Prentice Hall PTR, 2001.

[HH01] J. Holmes and W. Holmes. *Speech Synthesis and Recognition*. Taylor & Francis, 2nd edition, 2001.

[HPS93] B. Hudgins, P. Parker, and R.N. Scott. A New Strategy for Multifunction Myoelectric Control. *IEEE Transactions on Biomedical Engineering*, 40(1):82–94, January 1993.

[HSW89] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[JB05] C. Jorgensen and K. Binsted. Web Browser Control Using EMG Based Sub Vocal Speech Recognition. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 294c–294c, 2005.

[JLA03] C. Jorgensen, D.D. Lee, and S. Agabon. Sub Auditory Speech Recognition Based on EMG Signals. In *Proceedings of the International Joint Conference on Neural Networks*, volume 4, pages 3128–3133, 2003.

[KGA01] J.S. Karlsson, B. Gerdle, and M. Akay. Analyzing Surface Myoelectric Signals Recorded During Isokinetic Contractions. *IEEE Engineering in Medicine and Biology Magazine*, 20(6):97–105, Nov-Dec 2001.

[KHJC04] S.S. Kim, M. Hasegawa-Johnson, and K. Chen. Automatic Recognition of Pitch Movements Using Multilayer Perceptron and Time-Delay Recursive Neural Network. *IEEE Signal Processing Letters*, 11(7):645–648, July 2004.

[KKAB04] S. Kumar, D.K. Kumar, M. Alemu, and M. Burry. EMG Based Voice Recognition. In *Proceedings of the Intelligent Sensors, Sensor Networks and Information Processing Conference*, pages 593–597, 2004.

[KL90] A. Khotanzad and J.H. Lu. Classification of Invariant Image Representations Using a Neural Network. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 38(6):1028–1038, June 1990.

[KM96] S. Kumar and A. Mital. *Electromyography in Ergonomics*. Taylor & Francis Ltd., 1996.

[Lip67] O.C.J. Lippold. Electromyography. In P.H. Venables & I. Martin, editor, *A Manual of Psychophysiological Methods*, pages 245–298. North-Holland Publishing Company - Amsterdam & London, 1967.

[LJ91] P.H.W. Leong and M.A. Jabri. Arrhythmia Classification Using Two Intracardiac Leads. In *Proceedings of Computers in Cardiology*, pages 189–192, 1991.

[MHS03] H. Manabe, A. Hiraiwa, and T. Sugimura. Unvoiced Speech Recognition using EMG - Mime Speech Recognition. In *Proceedings of the Conference on Human Factors in Computing Systems*, pages 794–795, 2003.

[Mit01] S.K. Mitra. *Digital Signal Processing: A Computer-Based Approach*. McGraw-Hill, 2001.

[MO86] M.S. Morse and E.M. O'Brien. Research Summary of a Scheme to Ascertain the Availability of Speech Information in the Myoelectric

Signals of Neck and Head Muscles Using Surface Electrodes. *Computers in Biology and Medicine*, 16(6):399–410, 1986.

[MP69] M. Minsky and S. Papert, editors. *Perceptrons: An Introduction to Computational Geometry*. Cambridge. MA: The MIT Press, 1969.

[MT00] E. Micheli-Tzanakou. Neural Networks in Biomedical Signal Processing. In J.D. Bronzino, editor, *Biomedical Engineering Handbook (2nd Edition) Volume I*, pages 58–1. CRC Press & IEEE Press, 2000.

[MZ04] H. Manabe and Z. Zhang. Multi-stream HMM for EMG-Based Speech Recognition. In *Proceedings of the 26rd Annual International Engineering in Medicine and Biology Society*, volume 2, pages 4389–4392, 2004.

[Nat] National Instruments Inc. *6023E/6024E/6025E User Manual*. December 2000 Edition.

[NQL83] S.H. Nawab, T.F. Quatieri, and J.S. Lim. Signal Reconstruction from Short-time Fourier Transform Magnitude. *IEEE Transactions on Acoustics, Speech, and Processing*, 31(4):986–988, 1983.

[PBYTI02] D. Peleg, E. Braiman, E. Yom-Tov, and G.F. Inbar. Classification of Finger Activation for Use in a Robotic Prosthesis Arm. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 10(4):290–293, Dec 2002.

[PWS96] W.J. Pielemeier, G.H. Wakefield, and M.H. Simoni. Time-Frequency Analysis of Musical Signals. *Proceedings of the IEEE*, 84(9):1216–1230, 1996.

[Rab89] L. R. Rabiner. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77:257–286, February 1989.

[RHW86] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning Internal Representations by Error Propagation. In D.E. Rumelhart, J.L. McClelland, and the PDP Research Group, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1, pages 318–362. Cambridge, MA, USA: MIT Press, 1986.

[RJ86] L. R. Rabiner and B. H. Juang. An Introduction to Hidden Markov Models. *IEEE ASSP Magazine*, 3(1):4–16, January 1986.

[RJ93] L. Rabiner and B.H. Juang. *Fundamentals of Speech Recognition*. Prentice Hall PTR, 1993.

[RL81] L.R. Rabiner and S.E. Levinson. Isolated and Connected Word Recognition - Theory and Selected Applications. *IEEE Transactions on Communications*, 29(5):621–659, May 1981.

[RRWK83] A.E. Rosenberg, L.R. Rabiner, J.G. Wilpon, and K. Kahn. Demisyllable-Based Isolated Word Recognition System. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 31(3):713–726, June 1983.

[ST85] N. Sugie and K. Tsunoda. A Speech Prosthesis Employing a Speech Synthesizer - Vowel Discrimination from Perioral Muscle Activities and Vowel Production. *IEEE Transactions on Biomedical Engineering*, 32(7):485–490, July 1985.

[Tay96] J.G. Taylor, editor. *Neural Networks and Their Applications*. John Wiley & Sons Inc., 1996.

[TK03] S. Theodoridis and K. Koutroumbas. *Pattern Recognition, 2nd Edition*. Academic Press, 2003.

[Tre82] T.E. Tremain. The Government Standard Linear Predictive Coding Algorithm: LPC-10. *Speech Technology*, pages 40–49, April 1982.

[WC93] J.X. Wu and C. Chan. Isolated Word Recognition by Neural Network Models with Cross-Correlation Coefficients for Speech Dynamics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(11):1174–1185, November 1993.

[Wil98] W.J. Williams. Recent Advances in Time-Frequency Representations: Some Theoretical Foundations. In M. Akay, editor, *Time Frequency and Wavelets in Biomedical Signal Processing*. IEEE press, 1998.

[WJJ96] R.H. Westgaard, T. Jansen, and C. Jensen. EMG of Neck and Shoulder Muscles: the Relationship Between Muscle Activity and Muscle Pain in Occupational Settings. In S. Kumar and A. Mital, editors, *Electromyography in Ergonomics*, pages 227–258. Taylor & Francis Ltd., 1996.

Publications

Journal Papers

- Y.M. Lam, P.H.W. Leong, M.W. Mak: Frame-based Speech Synthesis using Surface Electromyogram Signals, *IEEE Transactions on Biomedical Engineering*, (*in preparation*).

- Y.M. Lam, P.H.W. Leong, M.W. Mak: Word-based Speech Synthesis using Surface Electromyogram Signals, *Electronics Letters*, (*under review*).

Full Length Conference Papers

- Y.M. Lam, P.H.W. Leong, M.W. Mak: Frame-Based SEMG-to-Speech Conversion, *Proceedings of the IEEE International Midwest Symposium on Circuits and Systems*, San Juan 2006.

- Y.M. Lam, M.W. Mak and P.H.W. Leong: Speech Synthesis from Surface Electromyogram Signal, *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology*, Athens 2005.