NAIST-IS-DD0961011

Doctoral Dissertation

Surface Electromyography Derived with Electrode Grid from Submental Region and its Application to Vowel Recognition

Takatomi Kubo

February 2, 2012

Department of Bioinformatics and Genomics
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Takatomi Kubo

Thesis Committee:

Professor Kazushi Ikeda	(Supervisor)
Professor Kiyohiro Shikano	(Co-supervisor)
Associate Professor Tomoki Toda	(Co-supervisor)
Professor Masaki Yoshida	(Osaka Electro-Communication University)

Surface Electromyography Derived with Electrode Grid from Submental Region and its Application to Vowel Recognition[*]

Takatomi Kubo

Abstract

Speech of dysarthric patients becomes slurred and makes communication difficult. However, communication assistance methods for dysarthric patients have not been established enough yet. There are great expectations to develop a novel assistance method. In fact, there are various researches done on the communication assistance. One of those is a study by Deng et al., which showed the effectiveness of the speech recognition based on surface electromyography (sEMG). However, electrode locations used in those previous studies are still controversial. This is because disc electrodes or parallel bar electrodes were used in those studies. Although such electrodes are commonly-used, they cannot avoid deterioration of signal-to-noise ratio caused by the influence of innervation zones and crosstalks, which must be taken into account for determining the electrode location.

sEMG measurement using an electrode grid which has multichannel is used as an effective method to cope with the problems caused by innervation zones and crosstalks in the electro-physiological researches. In this dissertation, we introduce the use of electrode grid based measurement to speech recognition based on sEMG and investigate whether this measurement method is effective or not. Producing five vowels and submental region are employed as the experimental task and measurement site, respectively. The reason why we choose the submental region is that electrical activity of muscles which control the movement of tongue can be measured partly from it.

[*] Doctoral Dissertation, Department of Bioinformatics and Genomics, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0961011, February 2, 2012.

In this dissertation, first, we illustrate that the positions of innervation zones of superficial muscles can be estimated by using an electrode grid. Second, we investigate the feasibility of vowel recognition based on sEMG derived with electrode grid and show that vowel recognition can be realized to some extent from sEMG signals of submental region. And lastly, we describe the results of applying sparse discriminant analysis to the sEMG signals which can contain redundancy. Thus, it is shown that redundant channel can be removed by the proposed method.

Keywords:

surface electromyography, electrode grid, submental region, speech recognition, dysarthria

格子状電極を用いたオトガイ下部からの表面筋電図計測および母音認識へのその応用*

久保　孝富

内容梗概

構音障害を来した患者では、発話が不明瞭となり、コミュニケーションに支障を来してしまう。しかし、構音障害者へのコミュニケーション支援の方法は、十分に確立されているとは言い難く、新たな支援方法の開発に期待が寄せられている。実際にコミュニケーション支援の目的で様々な研究が行われており、そのような研究の一つに表面筋電図信号に基づいた音声認識が有効であることを示したDengらの報告がある。一方で、表面筋電図信号に基づいた音声認識の先行研究では、電極の配置の妥当性に議論の余地が残っている。これら先行研究では、一般的に用いられる皿状電極やパラレルバー電極を用いて表面筋電図計測を行っているが、神経支配帯やクロストーク等の影響により信号対ノイズ比の低下を生じ得るため、電極の配置の際にそれらの要因に対して十分な考慮が必要となる。

神経支配帯やクロストークの影響に対処する上で、多計測点を有する格子状電極を用いた表面筋電図計測が、電気生理学の研究分野では有効な方法だとされている。そのため、本研究では、表面筋電図信号に基づいた音声認識に対して、格子状電極を用いた計測方法の導入を試み、その有用性の検証を行う。実験課題には、音声においての重要性を考慮して5母音を採用することとし、そして計測部位はオトガイ下部とする。オトガイ下部を対象とする理由は、母音生成において重要と考えられる器官である舌の運動に関与する筋活動の計測を行えるからである。

本論文では、まず格子状電極を用いることで表層にある筋肉の神経支配帯の位置が特定され得ることを示す。次に、格子状電極によって導出された表面筋電図信号を用いて母音認識の実現が可能か検証し、実際にオトガイ下部からのみで

[*] 奈良先端科学技術大学院大学 情報科学研究科 情報生命科学専攻 博士論文, NAIST-IS-DD0961011, 2012年2月2日.

あっても一定の精度で母音認識を実現できることを示す。最後に、格子状電極を用いることで計測信号に冗長性が生じてしまうと考えられるが、sparse discriminant analysisを用いることで、冗長なチャンネルを削減できる可能性があることを示す。

キーワード

表面筋電図, 格子状電極, オトガイ下部, 音声認識, 構音障害

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1. Background

Speech is a unique, complex, and dynamic motor activity through which individuals express thoughts and emotions. It is one of the most powerful tools of the human species, and it contributes greatly to the quality of life. However, dysarthria deprives people of such an invaluable tool. According to Duffy [17], dysarthria is defined as "a collective name for a group of neurologic speech disorders resulting from abnormalities in the strength, speed, range, steadiness, tone, or accuracy of movements required for control of the respiratory, phonatory, resonatory, articulatory, and prosodic aspects of speech production. The responsible pathophysiologic disturbances are due to central or peripheral nervous system abnormalities and most often reflect weakness; spasticity; incoordination; involuntary movements; or excessive, reduced, or variable muscle tone." This definition implies that dysarthria can be categorized into types, each types characterized by different underlying neuropathophysiological findings. **Table 1.1** summarizes the categorization scheme developed by Darley, Aronson, and Brown [11] (unilateral upper motor neuron dysarthria is added [17]). Dysarthria arises from neuromuscular disease, such as cerebrovascular disease, parkinson disease, amyotrophic lateral sclerosis (ALS), cerebral palsy, etc. The number of dysarthric patients in Japan were estimated to be approximately from 650,000 to 700,000 [62]. Furthermore, because aging of population will worsen the situation, communication problem caused by dysarthria will gain in importance.

Since severe dysarthric speech is considerably difficult even for caregivers and fam-

Table 1.1. Types of dysarthria and their responsible lesion and neuromotor bases

Type	Responsible lesion	Neuromotor basis
Flaccid	Lower motor neuron	Weakness
Spastic	Bilateral upper motor neuron	Spasticity
Ataxic	Cerebellum	Incoordination
Hypokinetic	Extrapyramidal circuit	Rigidity, or reduced range of movement
Hyperkinetic	Extrapyramidal circuit	Abnormal movements
UUMN*	UUMN*	Weakness, incoordination, or spasticity
Mixed	More than one	More than one basis

*UUMN:Unilateral upper motor neuron

ilies to understand, some of dysarthric patients have to use other communication tools called augmentative and alternative communications (AAC), either temporarily or permanently [17,63]. AACs are heterogeneous, and include gesture, facial expression, eye gaze, writing, boards with alphabets and/or pictures, PC-based system with/without special user interface (**Fig. 1.1**), and etc. Even if limb movements are insufficient, these special interfaces can be used by any part of body which remains under volitional control. However, communication tools described above are significantly less efficient than speech despite requiring residual function. It is obviously needed to develop more efficient AAC devices.

2. Related Works

In fact, there have been various researches on the communication assistance. In this section, some of such researches are introduced.

2.1 Speech Recognition for Dysarthric Patients

If the content of what dysarthric patients want to say can be estimated, the method will be a more efficient AAC device. In fact, there are some researches where speech recognition was applied for dysarthric patients to estimate what they intended to say

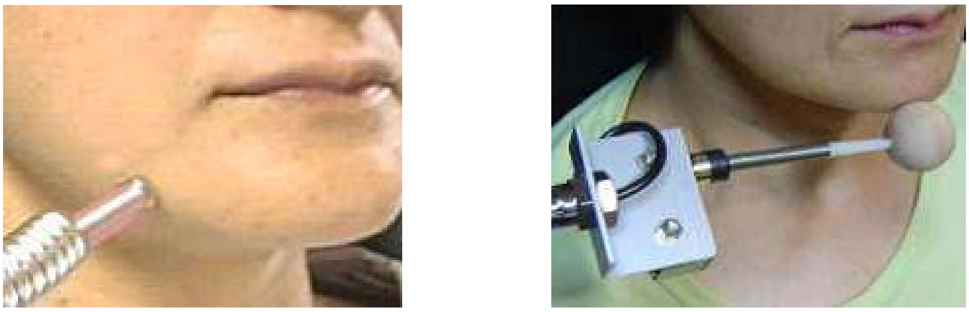

Figure 1.1. Example of user interfaces of augmentative and alternative communications. Users can input touching or pushing these interface. This image is reprinted from [10].

[56], since speech recognition technology has advanced to the point of being utilized in our daily lives. However, users' speech impairment have caused low recognition accuracy. Attempts to apply speech recognition with standard, commercially available technology have been largely unsuccessful, or successful for only a limited vocabulary of words. Blaney et al. [3], Thomas-Stonell et al. [51], and Raghavendra et al. [43] showed that speech recognition accuracy were significantly lower for individuals with moderate to severe dysarthria compared to individuals without dysarthria. It is also suggested that greater speech variability often correlates with increasing severity of dysarthria. Therefore, it seems to be difficult to achieve high recognition accuracy with severe dysarthric patients. Extensive efforts have been underway to develop speech recognition technology based on models of dysarthric speech [18, 21, 22, 34, 44].

2.2 sEMG-based Speech Recognition

To improve the recognition accuracy of dysarthric patients, Deng et al. [15] proposed a speech recognition system based on surface electromyography (sEMG), with and without acoustic signal. In the former case, they showed that high word recognition accuracy (over 95%) could be achieved for dysarthric patients suffering from stroke and cerebral palsy under speaker-dependent isolated-word recognition condition. sEMG is a procedure that measures muscle electrical activity associated with muscle fiber contraction by using electrodes attached on the skin. More detail about sEMG is provided in the chapter 2.

Not only in cases when a user makes usual voiced speech, but also when voiceless mouthed speech is made, sEMG-based speech recognition can support communication. Therefore, sEMG-based speech recognition may also enable tracheostomized or ventilator-dependent individuals who have difficulty in producing voice. Actually, Fukuda et al. [19] proposed an sEMG-based Japanese speech synthesizer system, where six Japanese phonemes (five vowels, i.e. /a/, /i/, /u/, /e/, /o/, and one syllabic nasal /n/) were recognized from the patterns of sEMG signals using a probabilistic neural network, and then words were recognized from series of phonemes using hidden Markov models (HMMs). Although the recognition accuracy of continuous speech production was lower than that of syllable-wise speech production, effective phoneme recognition with a laryngectomee was achieved by Fukuda's system. The results of their study indicate that sEMG-based speech recognition has the potential to be a novel type of speech prosthesis.

sEMG-based speech recognition has been investigated as an augmentative or alternative information source not only for dysarthric patients, but also for healthy people [14].The first study of the sEMG-based speech recognition dates back to the mid-1980s. In 1985, Sugie et al. used three channels of sEMG and a finite automaton to discriminate five Japanese vowels [48]. The target muscles were the digastricus, the zygomaticus major, and the orbicularis oris. An average recognition accuracy of 64% was achieved, and they also demonstrated a pilot real-time system. Morse et al. investigated the availability of speech information in four channels of sEMG from neck and head [37]. Recognition accuracies of 97% and 35% was observed for the vocabularies of 2-word and 17-word, respectively.

Over the last decade, there has been significant progress in the researches on sEMG-based speech recognition. Encouraging performance was first reported by Chan et al. [6], who achieved an average word accuracy of 93% on a vocabulary of the English ten digits. Linear discriminant analysis (LDA) was used as classifier in this study. Chan also demonstrated that an HMM is applicable in sEMG-based speech recognition in other studies [4, 5]. Application scenario in these studies was to use sEMG-based speech recognition in a noisy environment. Jorgensen et al. proved the applicability of sEMG signals for non-audible speech recognition [24]. They chose a scaled conjugate gradient net as a classifier, and reported 92% word accuracy on a set of six words with dual tree wavelet feature. In the later study, they extended the vocabulary to six words used in previous study and ten English digits and reported a recognition accuracy of 73% with support vector machine using radial basis function [23].

Since the early studies described above were performed using isolated word recognition, as a result, it was difficult to undertake vocabulary expansion. Therefore, new word addition required training a new classifier. The training of reliable acoustic models for a larger vocabulary requires breaking words into sequences of sub-word units, such as phonemes, syllables, or etc. Jou et al. has shown that larger vocabularies of 108 words can be recognized with a word accuracy of around 70% in a single speaker setup by using the phoneme-based acoustic models [26]. They explored various feature extraction methods that represented the sEMG signals for continuous speech recognition better. Scheme et al. also incorporated an approach based on a phoneme model [45]. In order to process the data, they used an HMM, each of which represented a phoneme instead of a word. An 18-phoneme vocabulary, which contained words from “zero” to “nine”, was applied. The overall word accuracy was over 94.7%. Walliczek et al. researched sEMG-based speech recognition based on sub-word units: phoneme or syllable [54]. HMMs with Gaussian mixture models are used as classifiers. With a 32-word vocabulary in continuously spoken speech, phoneme model outperformed syllable model slightly with the accuracy of 79.8% and 79.3%. In the experiment with the vocabulary which was not included in training data, phoneme model outperformed syllable model with the accuracy of 62.4% and 55.1%. They also developed a time domain feature extraction method that gains significant improvement for words and sub-word units.

Wand et al. presented an experiments on speaker independent and speaker adaptive sEMG-based speech recognition, based on an sEMG data recorded from 14 speakers reading sentences in audible speaking mode, in a collaboration between Carnegie Mellon University, University of Pittsburgh, and Chatham University [55]. Schultz et al. described the training of context dependent phonetic feature bundles, which further improved recognition performance on the 101-word vocabulary, with up to 90% word accuracy in a speaker dependent setup [46]. This result was based on a large collection of sEMG data recorded from 78 speakers. Zhou et al. showed dramatically improved word recognition performance increasing recognition by an average of 20% over the approach of Schema et al., and achieved an average word classification accuracy of 98.5% [59]. The sEMG data were processed by class-specific PCA prior to feature extraction, and Mel-frequency cepstral coefficients (MFCCs) were used for feature extraction. Then, an uncorrelated linear discriminant analysis was used for dimensionality reduction. The resulting data were classified through an HMM classifier to obtain the phonemic log likelihoods of the phonemes, which are mapped to corresponding words using a word classifier (GMM).

Thus, previous studies have indicated the potential effectiveness of sEMG-based speech recognition for both healthy people and dysarthric patients.

3. Research Purpose

Although feasibility of sEMG-based speech recognition has been shown, in order to achieve a high recognition accuracy in sEMG-based speech recognition, it is necessary to decide the appropriate location of the electrodes. However, in previous studies, conventional disc electrodes or parallel bar electrodes were used and located empirically according to anatomical knowledge, as shown in **Fig. 1.2**. Because there exist relatively small muscles in proximity to each other in the face or neck region, it is difficult to avoid the influence of cross talks and innervation zones when conventional measurement methods are applied. It is required to take these factors into account carefully for deciding the electrode location.

To cope with the influence of innervation zones and crosstalks, sEMG measurement using electrode grid which has multichannel is used as an effective method in the electro-physiological research area [36]. Lapatki et al. [27–29] proposed a high density multichannel sEMG system using electrode grid to improve signal-to-noise ratio of sEMG signals recorded from the lower facial muscles.

To avoid missing out information about speech in the sEMG measurement step, we introduce the use of an electrode grid which consists of densely-spaced multielectrodes in this dissertation. Submental region is focused on as measurement site of sEMG signals during the production of 5 vowel sounds. There are multiple muscles in the submental region that play important roles in controlling the movements of the mandible and tongue [2]. However, the submental region was not given much emphasis in previous studies, and only one or at most two channels were used in these experiments. Their function should receive considerably more attention for speech recognition, especially for vowel recognition.

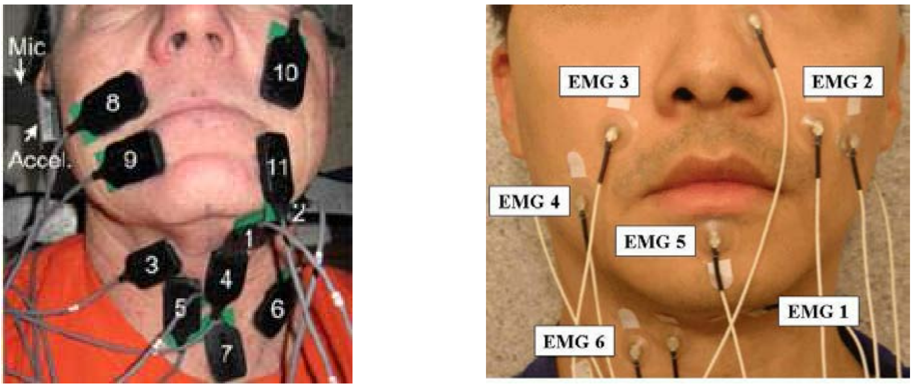

Figure 1.2. (*left*) Electrode location used by Deng et al. This image is reprinted from [15] (also used in [35]).
(*right*) Electrode location proposed by Jou et al. This image is reprinted from [26] (also used in [25, 31, 46])

.

In this dissertation, we verify three subjects. First, we examine whether the positions of innervation zones of superficial muscles can be estimated by using electrode grid. Second, we investigate feasibility of vowel recognition based on sEMG derived with electrode grid. Finally, we deal with an issue caused by redundant signals derived by electrode grid.

4. Organization of Dissertation

The remaining parts of this dissertation is organized as follows:

- Chapter 2 provides fundamental information about speech production and sEMG measurement. These are considerably important and necessary to understand this dissertation.

- In the chapter 3, we present sEMG system which we developed, and describe sEMG recording procedure used in our experiment, and illustrate that the positions of innervation zones of superficial muscles can be estimated by using electrode grid.

- Chapter 4 shows feasibility of vowel recognition based on sEMG derived with electrode grid.
- In the chapter 5, we deal with an issue caused by redundant signals of electrode grid.
- We end this dissertation with the chapter 6 which concludes our work and provides suggestions for future directions.

Chapter 2

Anatomical and Physiological Bases Related to sEMG-based Speech Recognition

1. Speech Production Mechanism

Human speech is produced by vocal organs presented in **Fig. 2.1** [30]. In “source-filter model”, speech production system is conceptualized as a combination of two parts: a sound source and an acoustic filter. In this model, it is assumed that speech sounds are produced by the action of a filter, the vocal tract, on a sound source.

The main energy source to provide the airflow is the lungs with the diaphragm and breathing muscles. The diaphragm muscle and breathing muscles act in compressing and decompressing the lungs. When speaking, the air flow is forced through the glottis between the vocal cords and the larynx to the three main cavities of the vocal tract, the pharynx and the oral and nasal cavities. From the oral and nasal cavities the air flow exits through the nose and mouth, respectively. Opening between the vocal cords, called the glottis, is the most important sound source in the vocal system. The most important function of the vocal cord is to modulate the airflow by rapidly opening and closing, causing sound source from which vowels and voiced consonants are produced. [30, 58].

The pharynx connects the larynx to the oral cavity. Its length can be changed slightly by raising or lowering the larynx and the soft palate. The soft palate also

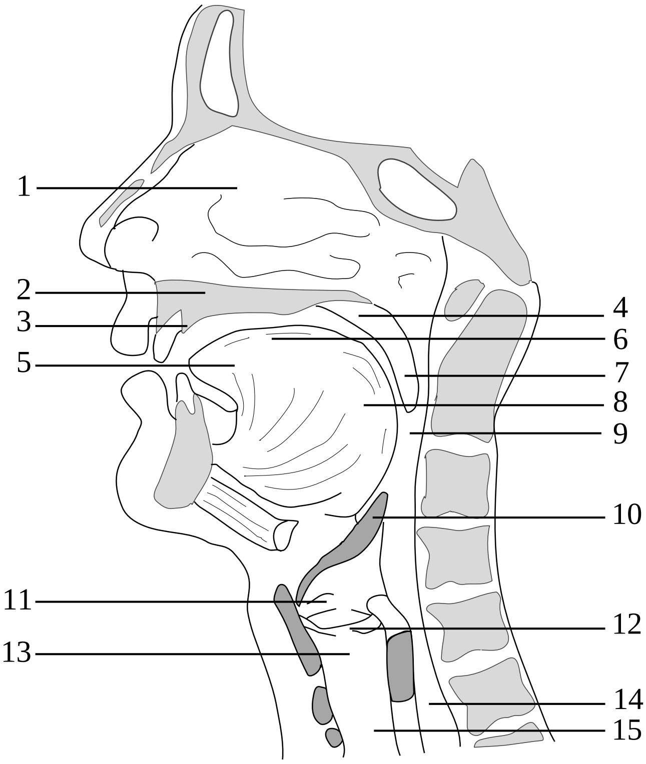

Figure 2.1. Vocal organs. This image is reprinted from [30]. (1) Nasal cavity, (2) Hard palate, (3) Alveoral ridge, (4) Soft palate (Velum), (5) Tip of the tongue (Apex), (6) Dorsum, (7) Uvula, (8) Radix, (9) Pharynx, (10) Epiglottis, (11) False vocal cords, (12) Vocal cords, (13) Larynx, (14) Esophagus, and (15) Trachea.

isolates or connects the route from the nasal cavity to the pharynx [30, 58]. The oral cavity is one of the important parts of the vocal tract. Its size, shape and acoustics can be varied by the movements of the palate, the tongue, the lips, the cheeks and the teeth. In particular, the tongue moves very flexibly. The tip and the edges of tongue can be moved independently. The entire tongue can move forward, backward, upward and downward. Therefore they allow constrictions to occur at various positions along the vocal tract. The lip controls the size and shape of the mouth opening through which speech sound is radiated. Unlike the oral cavity, the nasal cavity has fixed dimensions and shape. The airflow to the nasal cavity is controlled by the soft palate [30, 58].

All of the muscles related to speech is controlled by the motor cortex of the brain.

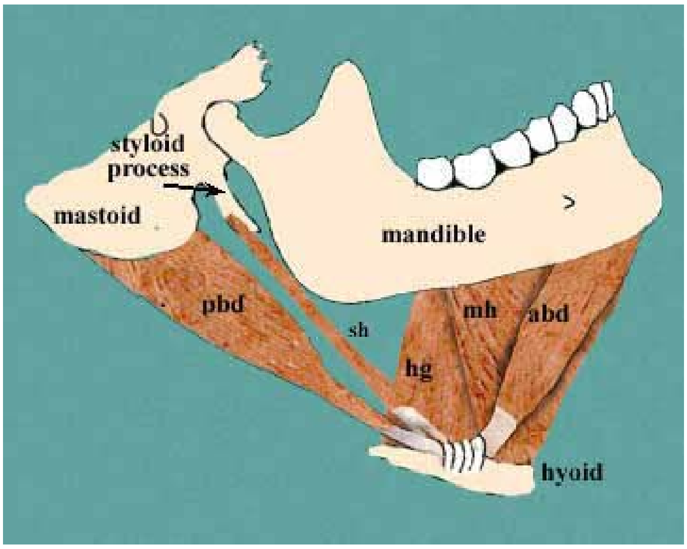

Figure 2.2. Muscles related to the tongue movement. This image is reprinted from [39]. “abd”: anterior belly of digastric, “hg”: hyoglossus, “mh”: mylohyoideus, “pbd”: posterior belly of digastric,“sh”: stylohyoideus

Motor signals produced by the brain for the movement of the face and tongue are transmitted through some specialized cranial nerves. During speech production, vocal organs are driven by coordinated muscle activations to manipulate the vocal tract shape and provide proper sound source. Tongue is the most complex and important speech organ that forms vocal tract shapes for producing most of the vowels and consonants. The tongue is driven by activating a set of associate muscles when producing speech [2, 49].(**Fig. 2.2** [39]). The anterior belly of the digastric, the mylohyoideus, and the geniohyoideus act to pull the hyoid bone upward and forward, or to depress the mandible. The anterior genioglossus is responsible for pulling the dorsum forward and downward. The posterior genioglossus has the function of pulling the tongue root forward and raising the dorsum. Thus, these muscles manipulate tongue shape, which is relevant with the vowel production. Despite their importance in human speech communication, the physiological mechanisms of the tongue muscles are poorly understood, or only assumed by information based on gross anatomy and a small number of muscle electrical-physiology [2].

2. Physiological Basis of sEMG

This chapter aims to provide fundamental information to understand the recording of the electrical activity of muscle using surface electromyography.

2.1 Anatomy of Motor Unit

The structural unit of skeletal muscle is the muscle fiber. A muscle fiber is a thin structure ranging from 10 to 100 microns in diameter [47]. The contraction of skeletal muscle is controlled by the motor neurons, as shown in **Fig. 2.3** [66]. Each muscle fiber can be activated by one α-motor neuron in spinal cord. On the other hand, one α-motor neuron can branch in up to multiple branches, each one terminating in a different muscle fiber. A functional unit which consists of α-motor neuron and all fibers innervated by it is called “motor unit” (MU). The term “endplate” refers to the junction between a muscle fiber and the terminal of the α-motor neuron [65]. Endplates tend to be localized near the central zone of muscles which is called “innervation zones”. The membrane current induced in the α-motor neuron by the synaptic innervation sites determines the firing pattern of the MU.

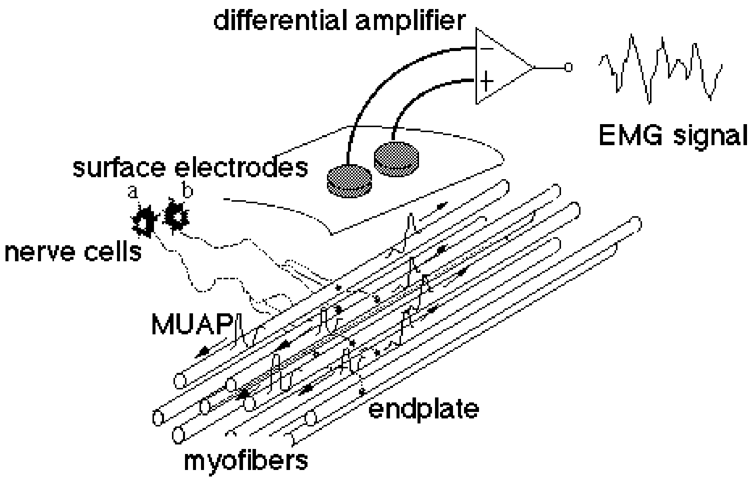

Figure 2.3. sEMG measurment. (This image is reprinted from “EMG Website” [66].)

2.2 Motor Unit Recruitment and Firing Frequency

In voluntary contractions, force is modulated by a combination of MU recruitment and changes in MU activation frequency [36]. MUs are recruited in order of increasing size of the α-motor neuron ("size principle"). It is well documented that motor unit recruitment and firing frequency (rate coding) depend primarily on the level of force and the speed of contraction. When low-threshold MUs are recruited, this results in a muscle contraction characterized by low force generating capabilities and high fatigue resistance. With recruitments for greater force and/or faster contraction, high threshold fatigable MUs are recruited. In the intrinsic muscles of human hands, motor units appears to be complete at relatively lower force. On the contrary, recruitments in the biceps, brachialis, and deltoid muscles may continue until greater force is modulated. It is demonstrated that firing rates of active MUs increase monotonically with increasing force output.

2.3 Muscle Electrical Activity

The membrane of muscle fiber is the seat of the bioelectric phenomena which result in sEMG signals. Key factor is dynamic voltage-dependent behavior of the membrane permeability to the main ions. The semi-permeable membrane of muscle fiber is composed of a lipid bilayer. It forms a physical barrier between intracellular and extracellular fluids, over which an ionic equilibrium is maintained. The composition of the extracellular fluid and intracellular fluid are different, as shown in **Table 2.1**. These ionic equilibrium forms a resting potential at the muscle fiber membrane, typically -80 to -90 mV. Ion pumps passively and actively control the flow of ions through the cell membrane.

When muscle fibers become innervated, the diffusion characteristics on the muscle fiber membrane are modified, and Na^+ flows into muscle fiber. When a certain threshold level is exceeded by the influx of Na^+, a depolarization of the cellular membrane, an action potential is developed. It is characterized by a quick change from -80 mV to +30 mV. Beginning from the endplates, the action potential spreads across the muscle fibers in both directions at a propagation speed of 2-6 m/s. The action potential leads to a release of Ca^{2+} ions in the intracellular fluid. They produces a chemical response resulting in a shortening of the contractile elements of the muscle cells [47]. This monopolar electrical burst is restored in the repolarization phase and is followed by a hyperpolarization phase.

Table 2.1. Intracellular and Extracellular Ion Concentration for Mammalian Muscle (mEq/L)

Ion	Intracellular fluid	Extracellular fluid
K^+	140	4
Na^+	14	142
Cl^-	4	125
HCO_3^-	8	28

2.4 sEMG Measurement Procedure

sEMG is a non-invasive technique to measure muscle electrical activity relating to muscular contractions with electrodes attached on the skin overlying a muscle or group of muscles. The typical equipment consists of electrodes, (preamplifier,) amplifier, analog-to-digital converter, and computer.

The skin can be considered as the boundary between two media: body tissue that contains source of electric field, and insulating space (air). The sources of electric field generate two-dimensional potential distribution on the skin. sEMG comprises the sum of the electrical contributions made by the active MUs as detected by the surface electrodes, as shown in **Fig. 2.3**. When surface electrodes are applied, the distance between the current source and the detection point is significant, and the spatial low-pass filtering effect of the volume conductor becomes relevant. The surface signals are usually detected as a differential of the signals recorded at different electrodes. Once the signal has been amplified by the preamplifiers if they exist, it is amplified further by the main amplifiers. After that, the signal is filtered and may be conditioned or processed. For example, processing may consist of rectifying, averaging, or integrating the signal. Only the raw signal may be recorded and interpreted by itself. Most EMG data, however, usually are subjected to some type of processing. More detailed explanations can be found in [12, 47, 65].

sEMG measurement based on linear electrode array and electrode grid has proposed to estimate muscle fiber conduction velocity (**Fig. 2.4**) and position of innervation zones and to decompose sEMG signals into motor unit action potentials [16, 33, 50, 65]. The linear electrode array and electrode grid are comprised of point electrodes from which single or double differential signals could be extracted. **Fig. 2.4** shows the propagation of the action potentials, from which muscle fiber conduction velocity can be estimated. The information obtained from multichannel signals was shown to be important and useful for research and clinical application.

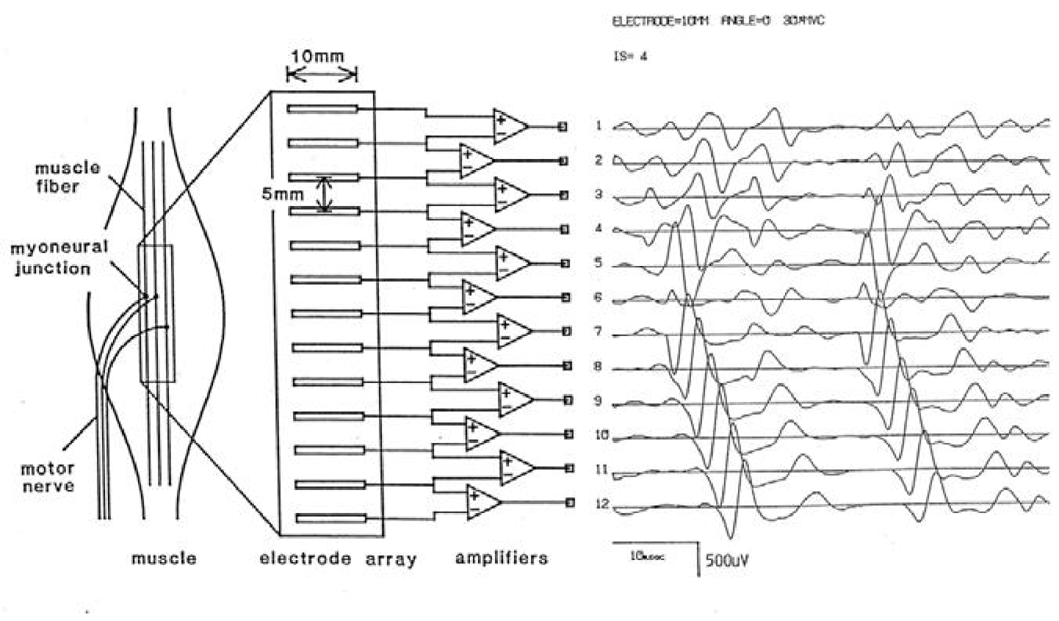

Figure 2.4. Patterns of sEMG signals derived with linear electrode array. (This image is reprinted from “EMG Website” [66].)

2.5 Feature Extraction for sEMG Signals

sEMG signals are expected to be used as effective system input not only for the speech recognition but also for prosthetic hand. Features commonly used for sEMG signals are introduced in this section [36, 40, 41, 57].

Time domain features

Features in the time domain are generally calculated quickly, and have been widely used in research and in clinical practice.

- Integrated EMG

  Integrated EMG (IEMG) is calculated as the summation of the absolute values of the sEMG signal amplitude. Generally, IEMG is used as an onset index to detect the muscle activity.It can be expressed as

  $$IEMG = \sum_{n=1}^{N} |x_n|$$

  where N denotes the length of the segment and x_n represents the n-th sample in the segment.

- Average rectified value

  Average rectified value (ARV) can be calculated using the moving average of full-wave rectified sEMG. It is an easy way for detection of the muscle activity. It is defined as

  $$ARV = \frac{1}{N} \sum_{n=1}^{N} |x_n|$$

- Root mean square

  Root Mean Square (RMS) is related to the constant force and non-fatiguing contraction. It relates to standard deviation, which can be expressed as

  $$RMS = \sqrt{\frac{1}{N} \sum_{n=1}^{N} x_n^2}$$

- Variance

  Variance (VAR) uses the power of the sEMG signal as a feature. Generally, the variance is the mean value of the square of the deviation of that variable. However, mean of sEMG signal is close to zero. In consequence, variance of sEMG can be calculated by

  $$VAR = \frac{1}{N-1} \sum_{n=1}^{N} x_n^2$$

- Zero crossing

  Zero crossing (ZC) is the number of times that sEMG signal crosses the zero. In sEMG feature, the threshold condition is used to take into account the background noise. This feature provides an approximate estimation of frequency domain properties.

- Willison Amplitude

  Willison amplitude (WAMP) is the number of times that the difference between sEMG signal amplitude among two adjacent segments that exceeds a predefined threshold to reduce noise effects same as zero crossing. The definition is as

  $$WAMP = \sum_{n=1}^{N-1} f(|x_n - x_{n+1}|)$$

  $$f(x) = \begin{cases} 1 & if \quad x \geq threshold \\ 0 & otherwise \end{cases}$$

  WAMP is related to the firing of motor unit action potentials (MUAP) and the muscle contraction level.

- Waveform length

  Waveform length (WL) is the cumulative length of the waveform over the time segment.

- Autoregressive coefficients

  Autoregressive (AR) model describes each sample of sEMG signal as a linear combination of previous samples (plus a white noise error term). The model is basically described by the following form:

  $$x_n = -\sum_{i=1}^{p} \alpha_i \, x_{n-i} + e_n$$

  where x_n is a sample of the model signal, α_i is AR coefficients, e_n is white noise or error sequence, and p is the order of AR model.

- Cepstrum

  Cepstrum of a signal is defined as the inverse Fourier transform of the logarithm of the magnitude of the power spectrum of the signal data. Cepstral coefficients are given by

$$c_n = F^{-1} \log \left| X(f) \right|$$

for each channel after time window was applied to the signals. $X(f)$ represents the short-time frequency spectrum, while F^{-1} indicates the inverse Fourier transform. The lower Cepstral coefficients contain information about the spectral envelope.

Time-Frequency domain features

The purpose of feature extraction is to emphasize the important information in the measured signals while rejecting noise and irrelevant signal change. Time-frequency features allow accurate representation of the target physical phenomenon in a specific frequency range. However, time-frequency representation generally requires a transformation that lead to increase of computational cost.

- Short-time Fourier transform
  The short-time Fourier transform (STFT) consists of a series of DTFs. STFT at frequency m and time k can be expressed as

  $$STFT[k, m] = \sum_{i=1}^{L-1} x[i]\, g[i-k]\, e^{-j2\pi mi/L}$$

  where L is the length of the sequence, and $g[i]$ is the window function. The resolution in time and frequency is lower bounded by the time-bandwidth uncertainty principle or Heisenberg inequality.

- Mean frequency
  Mean frequency (MNF) is calculated based on power spectrum. It can be expressed as

  $$MNF = \sum_{i=1}^{M} f_i P_i \Bigg/ \sum_{i=1}^{M} P_i$$

  where f_j is the frequency of spectrum at frequency bin j, P_j is the sEMG power spectrum, and M is the number of frequency bins in the spectrum.

- Median frequency

  Median frequency (MDF) can be expressed as

  $$\sum_{i=1}^{MDF} P_i = \sum_{i=MDF}^{M} P_i = \frac{1}{2}\sum_{i=1}^{M} P_i$$

- Wavelet transform [42]

  The wavelet transform (WT) overcomes the main drawback of the STFT by varying the time-frequency aspect ratio and by producing a good frequency resolution in long time windows (low frequencies) and a good time localization at high frequencies. Wavelet transform method is divided into two types: continuous wavelet transform (CWT) and discrete wavelet transform (DWT). CWT is defined as

  $$CWT(\tau, a) = \frac{1}{\sqrt{a}} \int x(t)\, \Psi\left(\frac{t-\tau}{a}\right) dt$$

  where $\Psi(t)$ is the mother wavelet, a is the scale variable, and τ is the shift variable. DWT is a technique that iteratively transforms an interested signal into multi-resolution subsets of coefficients.

Chapter 3

Proposed sEMG Recording Method during Vowel Production

1. sEMG System Overview

1.1 Electrode Grid

In the experimental setup, we used an sEMG system developed by Hattori et al. with few modifications made on the electrode grid [20, 64]. The electrodes which consisted of silver bars in Hattori's study were substituted with spring connector pins (SK KOHKI Co.,Ltd., AX-12ENR-00), with each pin having a diameter of 0.8 mm, to absorb any dynamic displacement of the attached site (**Fig. 3.1**). The set of electrodes were arranged in an array of 8 rows by 8 columns, with the interelectrode distance (IED) set to 5.08 mm, from center to center, in both directions. To reduce skin impedance, a voltage follower circuit was built with each electrode.

1.2 Setup for sEMG measurement

The electric potential differences between each pair of electrodes neighboring in column direction were amplified up to 66 dB with band-pass filtering between 10 to 1500 Hz. Subsequently, the electric potential differences were digitized with a 16-bit analog-to-digital converter (National Instruments, NI USB-6255) and a laptop computer running MATLAB with its Data Acquisition Toolbox (MathWorks, 2010a). A microphone

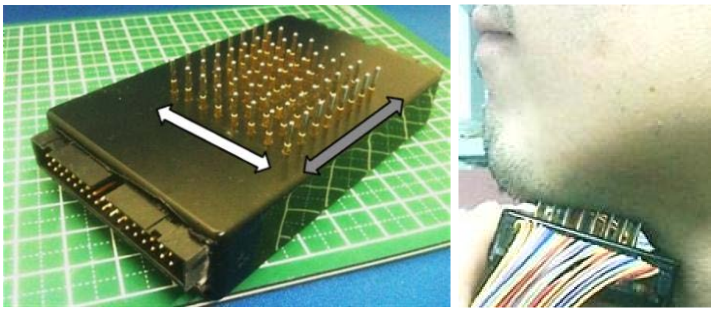

Figure 3.1. (*left*) The electrode grid. The white double-headed arrow indicates the row direction, and the gray double-headed arrow indicates the column direction.
(*right*) The location of the electrode grid on the submental region in lateral view.

(KNOWLES, SP0103NC3-3) was also attached in front of the electrode grid, so that acoustic signal could be simultaneously recorded along with the sEMG signals.

2. Recording Procedure

For this experiment, six adult Japanese native speakers were recruited as participants (two female and four male with mean age of 26.2 years. Refer to **Table 3.1** for more detail). All of the participants had no known speech impairment. In each trial, the subject was asked to produce each of the five Japanese vowels (/a/, /i/, /u/, /e/, and /o/) once in a random order. The task vowels were presented on a screen for 1 second with an interval of 2 seconds between each of them as shown in **Fig. 3.2**, and the subjects was instructed to start vowel production at the onset of a visual presentation and stop at the offset.

Table 3.1. Age and sex of participants

Subject ID	Age	Sex
1	33	Female
2	33	Male
3	24	Male
4	23	Male
5	22	Female
6	22	Male

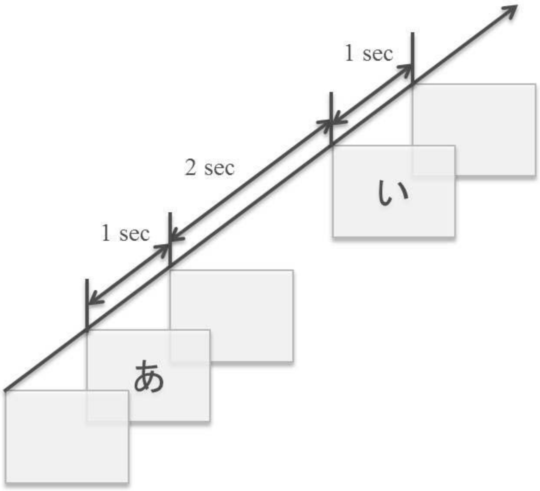

Figure 3.2. Experimental task presentation

Except for one subject (Subject 2), all other subjects (Subject 1, 3-6) conducted fifty trials in one day, while Subject 2 conducted fifty trials divided in half over two days. Every time subjects wanted to take a rest, enough time was given, while the electrode grid was removed. Though Subject 2 did not want to rest, other subjects took three to seven rest intervals throughout the experiment.

During vowel production, the sEMG signals were recorded with the electrode grid attached on the submental region as shown in **Fig. 3.1**. The grid's centerline in the column direction and the last row were aligned with the center of the mandible and the posterior edge of the submental triangle, respectively, by visual inspection. **Fig. 3.3**

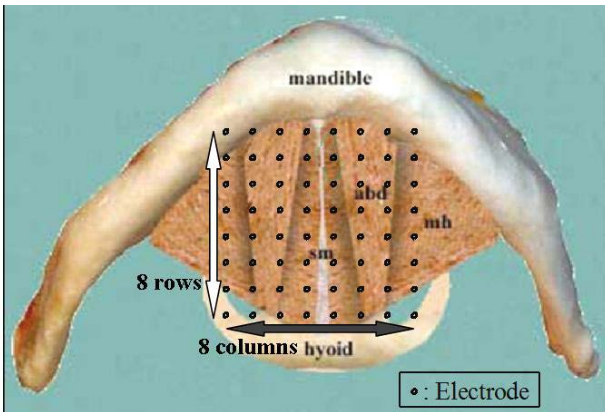

Figure 3.3. Anatomy of muscles in the submental region reprinted from [39] and corresponding positions of the electrodes. Black dots represent the positions of the electrodes. “abd”: the anterior belly of the digastric, “mh”: the mylohyoideus, “sm”: seam of the mylohyoideus (raphe of the mylohyoideus).

shows the muscles in the relatively superficial layer of the submental region and the corresponding positions of the electrodes. The anterior bellies of digastrics produce the sEMG signals whose amplitudes are relatively large. Therefore, the sEMG signals not only from the mylohyoideus but also from muscles in deeper layers, e.g. the geniohyoideus and the genioglossus, tend to be masked. In addition, there are innervation zones near the center of each muscle. Although the innervation zones and cross talks should be taken into account to avoid deterioration of the signal-to-noise ratio, it seems to be rather difficult to find appropriate locations using conventional bipolar electrodes whose diameters or lengths are approximately 1 cm.

As preparation, the skin on the submental region was cleaned with an alcohol swab prior to attaching the electrode grid. The electrode grid was fixed on a tripod, and the subject grasped the tripod’s legs, wrapped with stainless sheets, which served as the ground reference. Both the sEMG and acoustic signals were then captured and digitized at 16 kHz with an analog-to-digital converter.

Written informed consents were obtained from the subject prior to the experiment. This study was approved by the institutional ethics committee of Nara Institute of Science and Technology.

Examples of the signals coming from each vowel produced by Subject 1 are illustrated in **Fig. 3.4-3.8**. **Fig. 3.9** and **3.10** is horizontally magnified signals with the case of vowel /o/ for 200 msec and 50 msec, respectively. The signals coming from the anterior part seem to indicate similar patterns. Time delay caused by conduction can be regarded as short, given a common time frame length used for a conventional speech recognition.

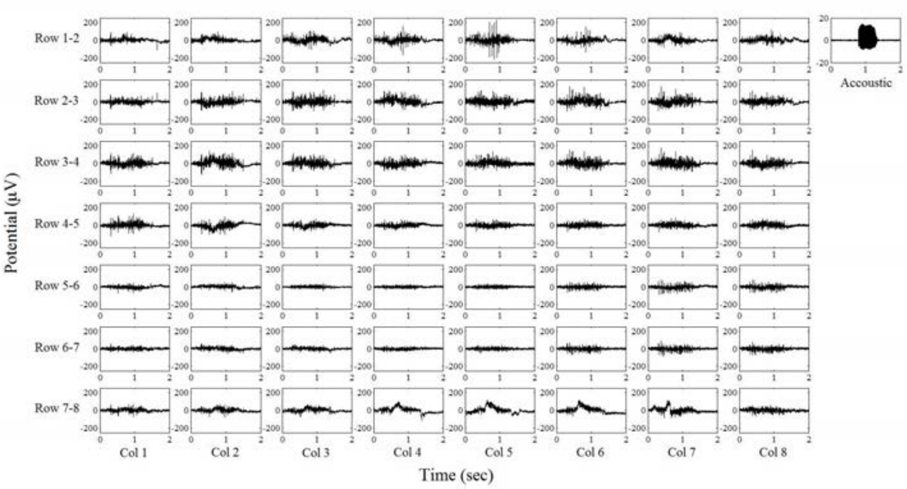

Figure 3.4. sEMG signals during producing vowel /a/

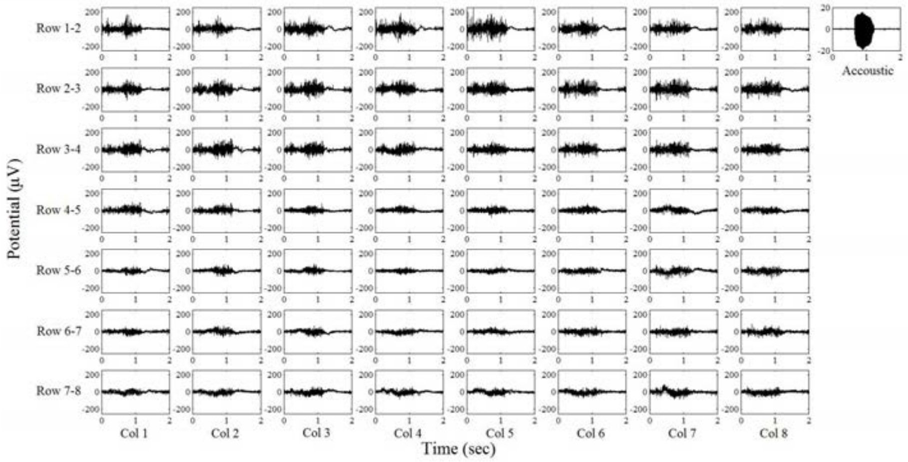

Figure 3.5. sEMG signals during producing vowel /i/

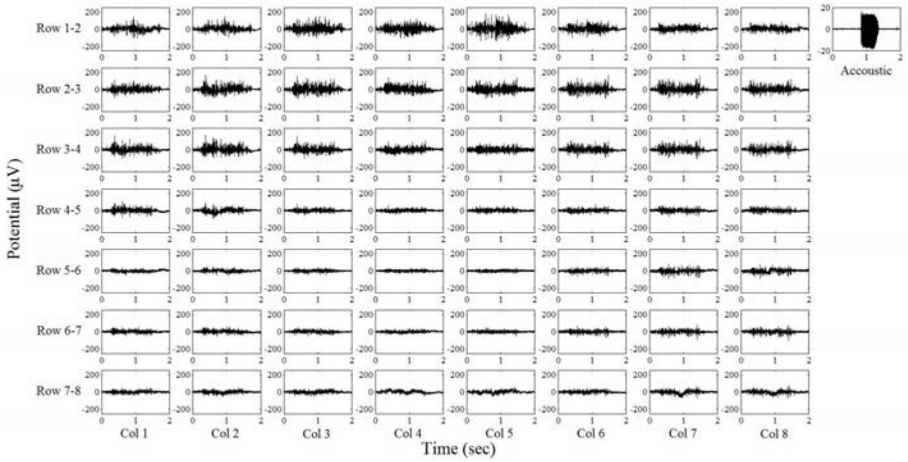

Figure 3.6. sEMG signals during producing vowel /u/

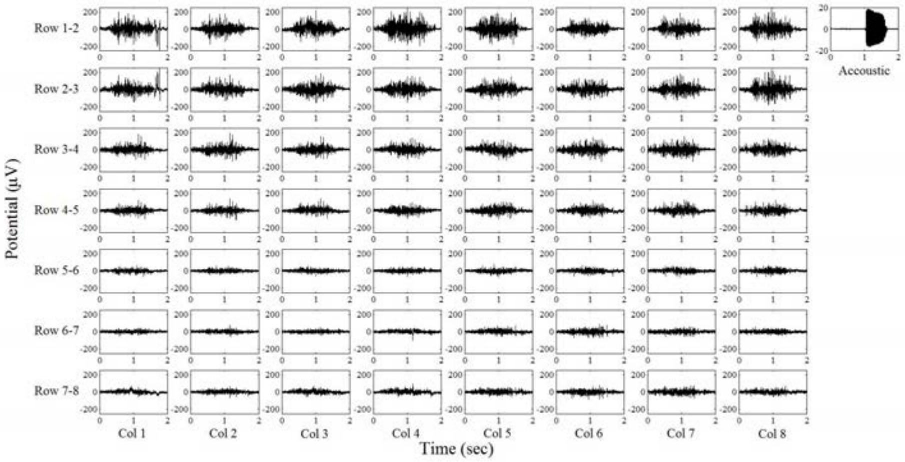

Figure 3.7. sEMG signals during producing vowel /e/

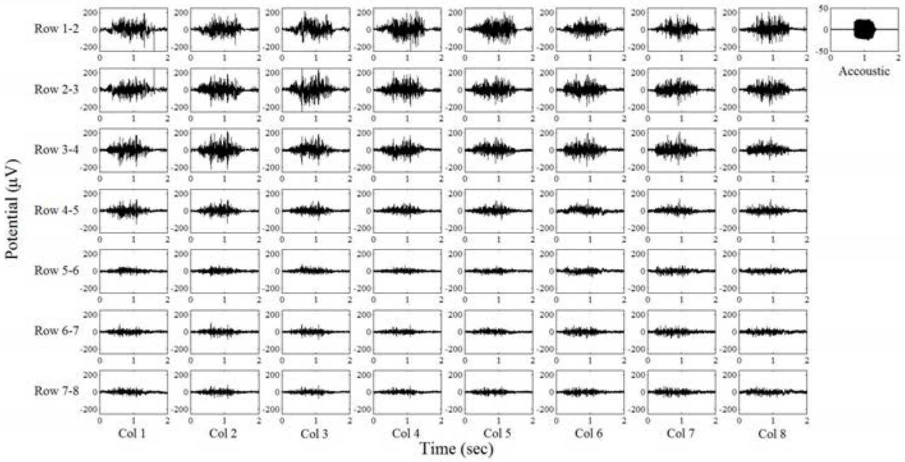

Figure 3.8. sEMG signals during producing vowel /o/

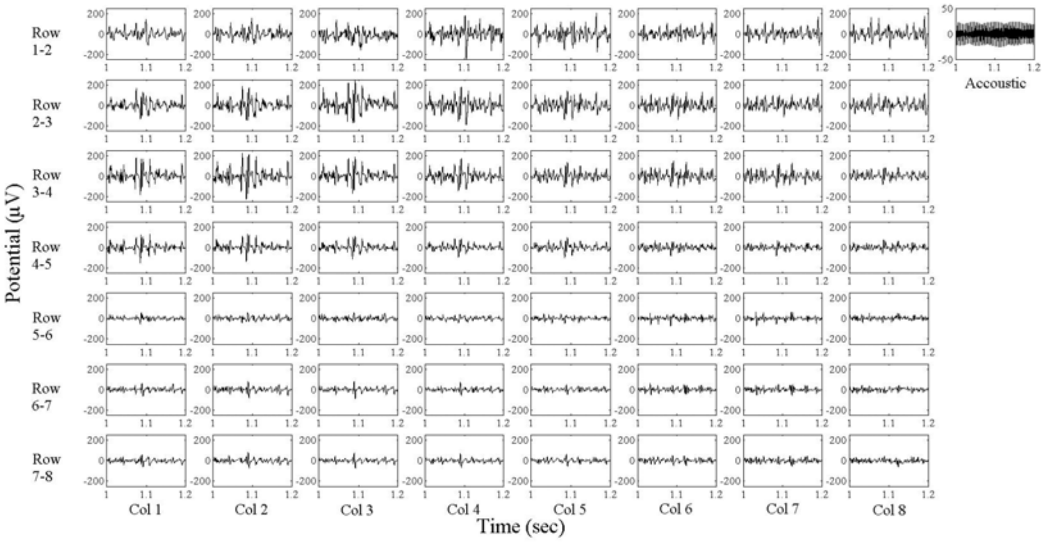

Figure 3.9. Magnified sEMG signals during producing vowel /o/ (200 msec)

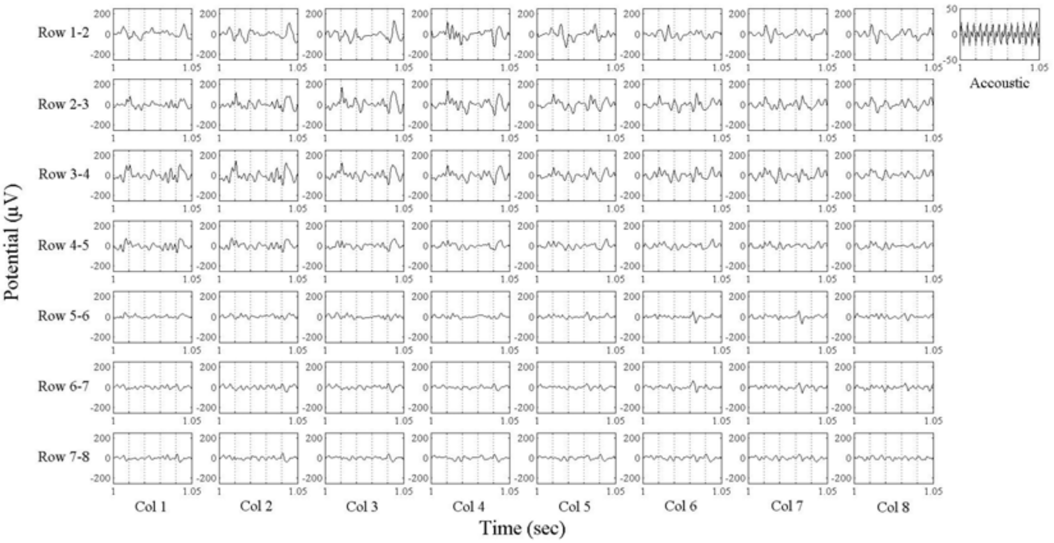

Figure 3.10. Magnified sEMG signals during producing vowel /o/ (50 msec)

Fig. 3.11 and **3.12** are spectrograms of the case same as **Fig. 3.9**. The latter is calculated from the channel “Row 3-4, Column 6”. A certain degree of stationarity is shown in these figures. The onset of the sEMG signals were precede that of the acoustic signal, and the offset of the sEMG signals follows that of the acoustic signal.

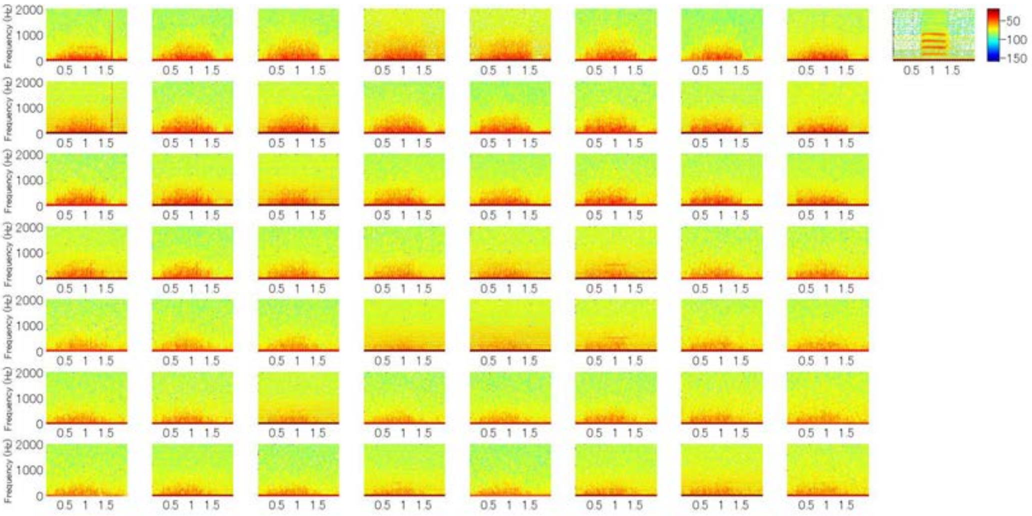

Figure 3.11. Spectrogram with the case producing vowel /o/

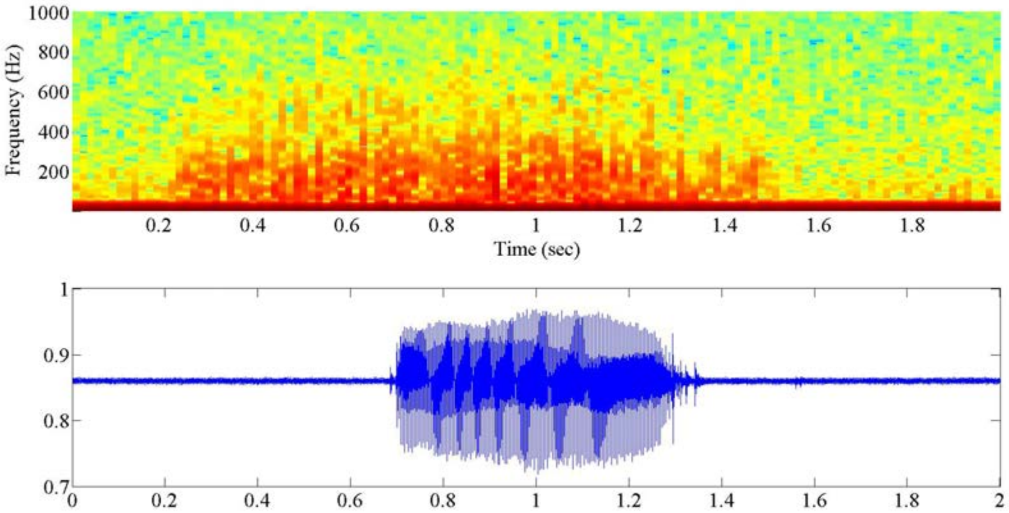

Figure 3.12. Magnified spectrogram with the case producing vowel /o/

3. Estimation of Innervation Zone

Hierarchical clustering, which was conducted according to Euclidean distances of the normalized signals, revealed that not only the anterior part but also the middle and posterior parts were clustered. The dendrogram of this hierarchical clustering is shown in **Fig. 3.13** and **3.14**. In addition, correlations between the representative channels and the whole channels during production of /o/ are shown in **Fig. 3.15** and **3.16**. The representative channels consist of the channels between the 3rd and 4th rows in the 6th column and between the 7th and 8th rows in the 7th column. Hereafter, these channels are denoted as “3-4, 6” and “7-8, 7”. While the channels “3-4, 6” and “7-8, 7” have positive correlations with the surrounding channels, these two channels have a significant negative correlation (-0.533, $p < 5.0 \times 10^{-291}$) with each other. **Fig. 3.17** represents the correlations between all possible combinations of channels.

These results might be relevant with the innervation zones of the anterior bellies of the digastrics from where the propagation of the motor unit action potentials starts (**Fig. 3.3** and **3.14**). Thus, by using electrode grid, the position of innervation zones can be estimated in the superficial muscles.

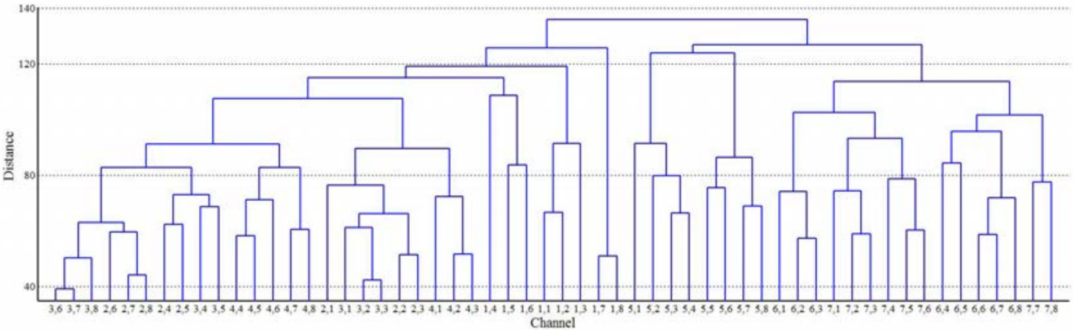

Figure 3.13. Dendrogram of hierarchical clustering on channels. Each label denotes the row and column of the channel, for example, “3, 6” denotes a channel between the 3rd and 4th electrodes in the row direction within the 6th column.

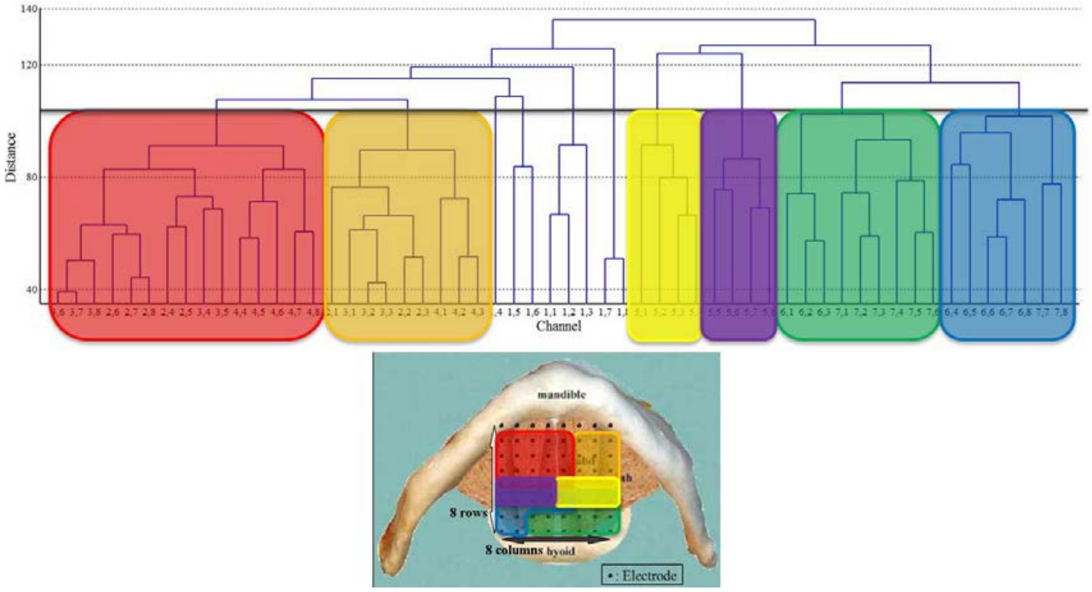

Figure 3.14. Cluster and corresponding anatomical locations. The correspondence between the cluster and the locations are represented by the colored shade. The anterior parts, the middle ones, and posterior ones are clustered. Each of them is subdivided into right and left.

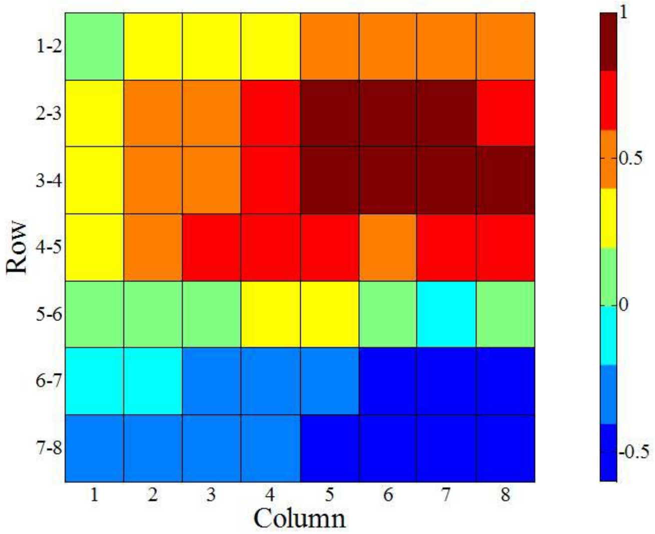

Figure 3.15. Correlations between the signal coming from channel “3-4, 6” and those coming from all channels of the electrode grid.

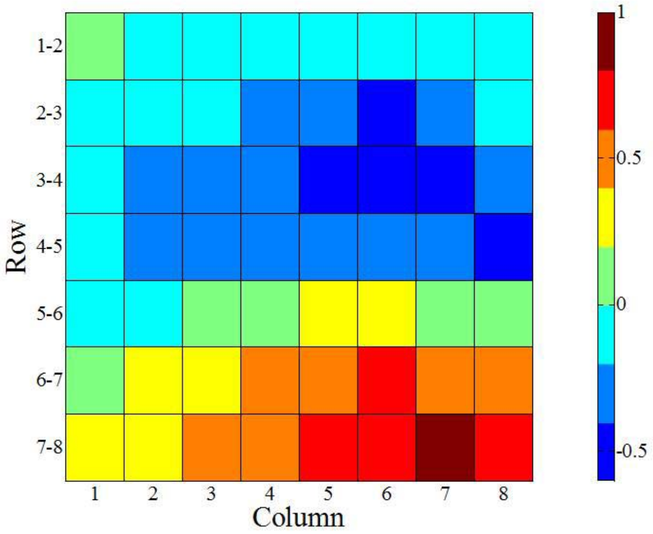

Figure 3.16. Correlations between the signal coming from channel “7-8, 7” and those coming from all channels of the electrode grid.

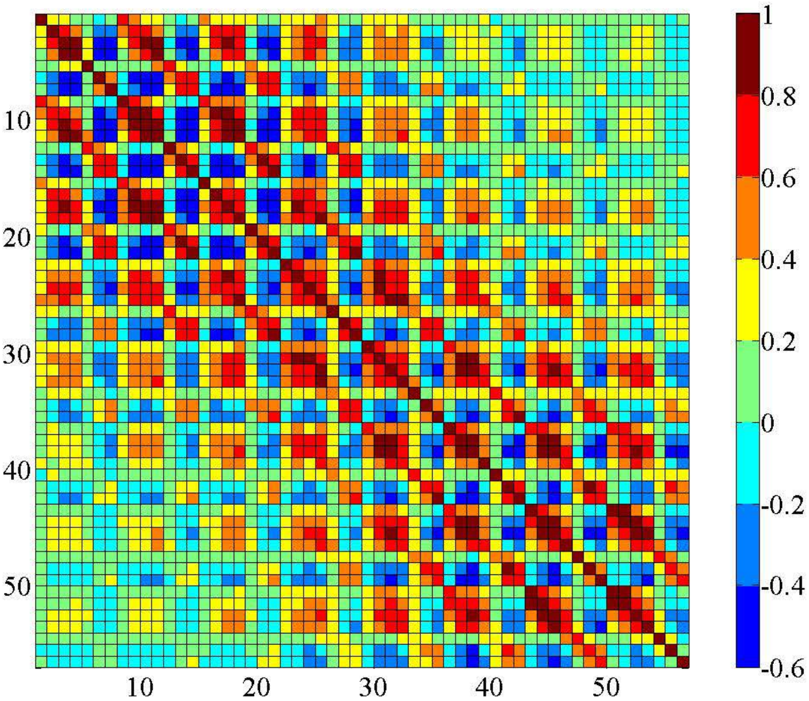

Figure 3.17. Correlations between all possible combinations of channels.

Chapter 4

Vowel Recognition Experiment

1. Introduction

To investigate whether electrode grid is more effective in extracting information for sEMG-based speech recognition than conventional electrodes, we compared the recognition accuracies between two methods: One was based on signals from all channels (hereafter,“all-channel method”) and the other was based on virtually reconstructed single bipolar signal (“single-channel method”).

2. Data Preprocessing

The sEMG signals were filtered with an 8th order low-pass Butterworth filter having a cut-off frequency of 500 Hz, and then downsampled to 2 kHz. The onsets and offsets of the acoustic signals were used as reference to determine those of the sEMG signals. The criteria applied in detecting the onsets and offsets of the acoustic signals were based on a set of amplitude thresholds, and these signals were then visually confirmed and corrected manually in only one onset. With the consideration of the delay between the sEMG signals and the acoustic signals [25], the onset of the sEMG signals were set to precede that of the acoustic signals by 150 msec, although the resting state could also be included. As for the offsets of sEMG signals, these were set to 150 msec after the offsets of the acoustic signals. These onsets and offsets of the sEMG signals were used to extract data for the following feature extraction process.

3. Feature Extraction

To be able to compare the single-channel method with the all-channel method, bipolar signals from all possible combinations of electrodes within the same column were virtually reconstructed from original signals, by adding signals from all channels between the two selected electrodes. Two types of feature sets were used in this study: (1) time domain features, and (2) cepstral coefficients. Features were extracted from the windowed signals of each channel. The window length was set to 25 msec with 50 samples, while the window period was set to 12.5 msec with 25 samples.

The time domain features consisted of the average rectified value (ARV), root mean square (RMS), zero-crossing rate of high-pass filtered signals, and the mean of the raw signals, along with the Δ and $\Delta\Delta$ features of these four features. To some extent, these features were similar to the features proposed by Jou et al. [26] except that less contextual information was used.

The real parts of the lower 15 cepstral coefficients, including the 0th coefficients, were used as features, along with the Δ and $\Delta\Delta$ features. Several researches have shown that Mel-frequency cepstral coefficients (MFCC), which are derived from cepstral coefficients by applying filter bank based on the Mel-scale, can also be used as features [15, 59]. However, there is no physiological plausibility to use MFCCs to parameterize sEMG signals, since the Mel-filter bank is designed to approximate human auditory perceptual response to acoustic signals. In addition, if the sampling rate, window length used in this study, and spectral features of the sEMG signals are taken into account, then the usefulness of Mel-filter bank will be reduced. Therefore, in this study, we employed cepstral coefficients instead of MFCC. For the all-channel method, features from all 56 channels were concatenated. This concatenation resulted in having more than several hundreds of dimensional features.

Dimensionality reduction was performed using linear discriminant analysis (LDA), which is commonly used to map the data onto a lower dimensional subspace keeping discriminative information as much as possible. The resulting final dimensions were reduced to four in both methods.

4. Vowel Recognition

Continuous HMM was adopted for vowel modeling, since it has been shown that the HMM is effective for sEMG-based speech recognition as well as for acoustic speech recognition. An HMM represents a stochastic process that takes sequential data as the inputs, and outputs the probabilities that the data are generated by the model. For each vowel, we used a nine state left-to-right HMM with three Gaussian mixture components, whose covariance matrices in each state are diagonal. Expectation maximization (EM) algorithm [13] was utilized in parameter estimation, and the vowel with the maximum likelihood was adopted as the recognition result. Hidden Markov Model Toolbox [38] was used to implement the HMMs in this experiment. 5-fold cross-validations were conducted to investigate the recognition accuracies.

5. Results

The comparison between the recognition accuracies of the different channels and feature conditions are shown in **Fig. 4.1**. For the single-channel method, the best recognition accuracies between all possible electrode combinations are indicated. The all-channel method outperformed the single-channel method. With respect to features, using cepstral coefficients indicated higher recognition accuracies than using time domain features. The all-channel method with cepstral coefficients achieved 85.6% recognition accuracy for Subject 1 and 79.6% recognition accuracy for Subject 2. **Table 4.1** shows recognition accuracy of all subjects under using cepstral coefficients from all channels. Although the recognition accuracies stay at 70% level with the subjects whose speech durations were short, those of the other subjects are almost over 80%. There is a positive correlation (0.827, $p < 0.05$ two sided t-test) between the mean speech durations and the recognition accuracies.

Table 4.1. Vowel recognition accuracy

Subject ID	Sex	Speech duration (ms)	Recognition accuracy (%)
1	F	880 ± 110	85.6
2	M	1150 ± 88	79.6
3	M	505 ± 24	71.2
4	M	537 ± 29	72.8
5	F	1260 ± 88	86.0
6	M	1042 ± 115	85.6

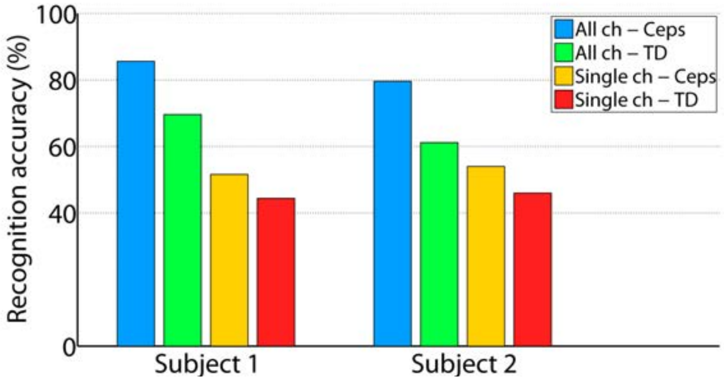

Figure 4.1. Comparison of the recognition accuracies for conditions with different features and channels used. All ch: all-channel method, Single ch: single-channel method, Ceps: cepstral coefficients, TD: time domain features.

Fig. 4.2 depicts the differences of recognition accuracies between the used electrode locations in the single-channel method, including the locations which were used in **Fig. 4.2**. The top of **Fig. 4.2** shows the result of the case when the participant was Subject 1 and the IED was 15.24 mm. The "Row" and "Column" labels denote the positions where the virtual bipolar electrodes in the grid were selected. In the following, location (i-j, k) denotes the bipolar signal between two (row, column) positions: (i, k) and (j, k). In the top of **Fig. 4.2**, the recognition accuracy reaches a maximum of 51.6%, at location (5-8, 2). However, it can be seen that (4-7, 1), (4-7, 2), and (4-7, 3), which are neighbors of the maximum point, indicate accuracies of 38.0%, 30.8%, and 38.4%, respectively. In some parts of the central locations in rows “2-5” or “3-6”, the accuracies are at 20 to 30%. In the center of **Fig. 4.2**, the location of the highest recognition accuracy is different from that of the top of **Fig. 4.2**. Yet some parts of the central locations in the row direction still indicated accuracies in the range of 20 to 30%. In the bottom of **Fig. 4.2**, tendencies of Subject 2 are shown. Here, the locations (3-5, 3) and (3-5, 6) reach the recognition accuracies of 54.0% and 53.6%, respectively. However, the location (3-5, 7) which is a neighbor of the location (3-5, 6) indicates an accuracy of 28.4%. In some posterior locations, the accuracies are at 20 to 30%.

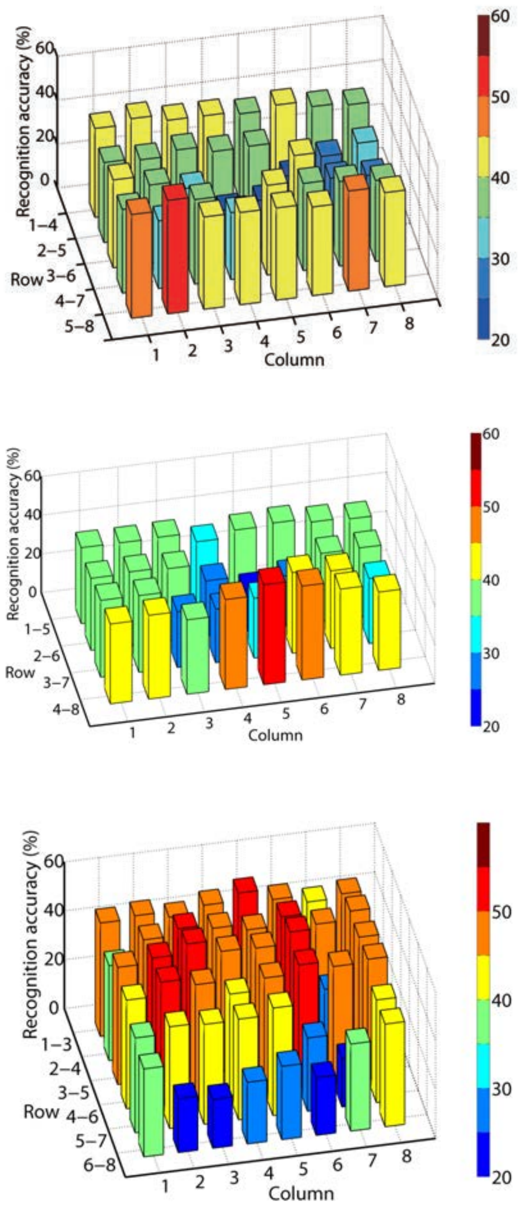

Figure 4.2. Changes of recognition accuracies with electrode locations under the single-channel method. (*top*) Subject 1 with IED = 15.24 mm, (*center*) Subject 1 with IED = 20.32 mm, (*bottom*) Subject 2 with IED = 10.16 mm. The “Row” and “Column” labels denote the rows and columns of the electrode grid from where virtual bipolar electrodes were selected.

Table 4.2 and **4.3** illustrate the confusion matrices, in the all-channel method and cepstral coefficients of Subject 1 and 2, respectively. The rows of the tables represent the actual spoken vowels, while the columns represent the vowels recognized by the HMMs. In both of Subject 1 and 2, there is a relatively high tendency that vowels /a/ and /e/ cannot be discriminated from each other. **Fig. 4.3-4.7** shows boxplots derived from each row, i.e. spoken vowel, of confusion matrices with respect to six subjects. From these boxplots, it is also shown that vowels /a/ and /e/ are hard to be discriminated from each other with high accuracy.

Table 4.2. Confusion matrix for the vowel recognition of Subject 1

		Recognized vowel					Accuracy
		/a/	/i/	/u/	/e/	/o/	(%)
Spoken vowel	/a/	40	0	1	6	3	80
	/i/	0	47	1	2	0	94
	/u/	0	1	48	0	1	96
	/e/	8	2	0	37	3	74
	/o/	3	0	1	4	42	84

Table 4.3. Confusion matrix for the vowel recognition of Subject 2

		Recognized vowel					Accuracy
		/a/	/i/	/u/	/e/	/o/	(%)
Spoken vowel	/a/	39	2	0	9	0	78
	/i/	1	38	3	8	0	76
	/u/	0	0	43	0	7	86
	/e/	6	3	0	38	3	76
	/o/	1	0	8	0	41	82

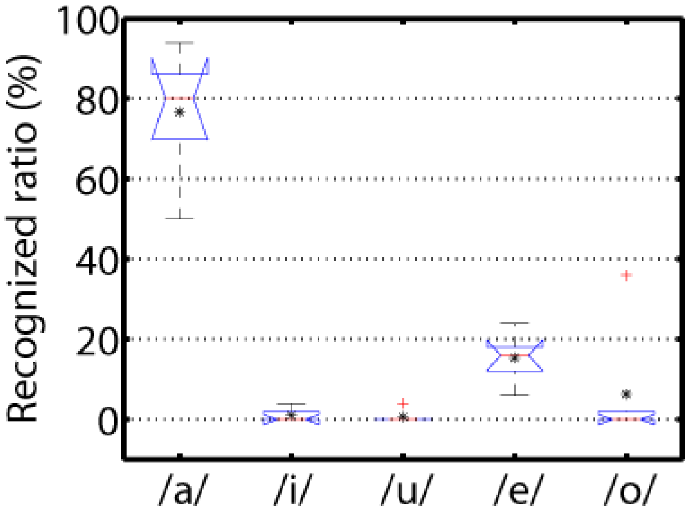

Figure 4.3. Recognized results of spoken vowel /a/. On each box, the central line is the median, the asterisk is the mean, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme data points not considered outliers, and outliers are plotted individually by red plus sign. Two medians are significantly different at the 5% significance level if their intervals which are represented by notches do not overlap.

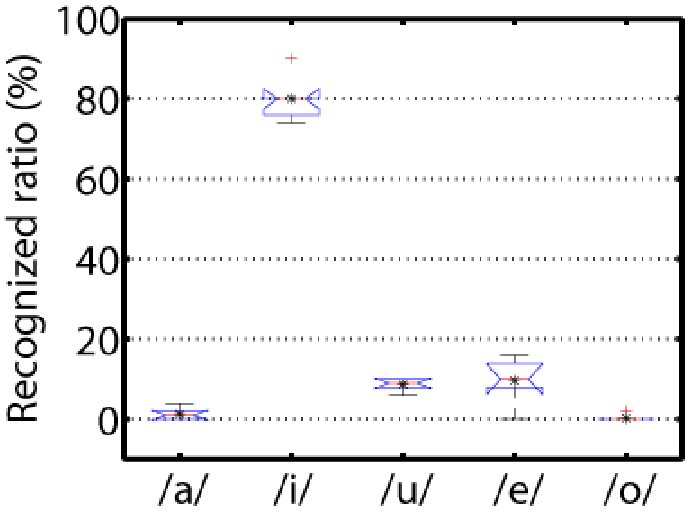

Figure 4.4. Recognized results of spoken vowel /i/.

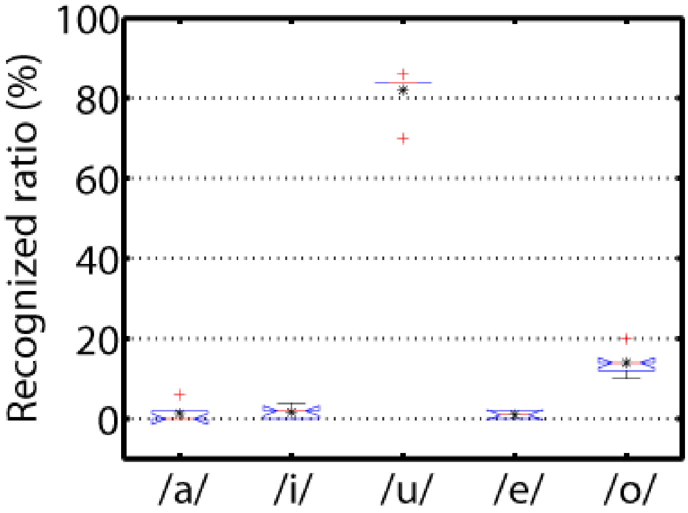

Figure 4.5. Recognized results of spoken vowel /u/.

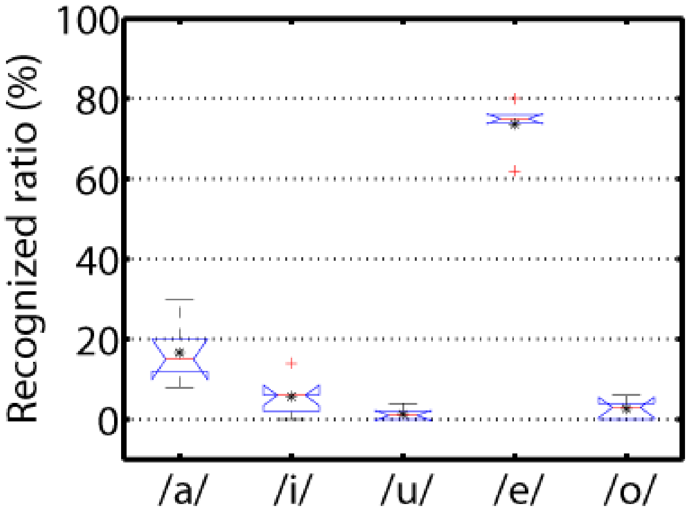

Figure 4.6. Recognized results of spoken vowel /e/.

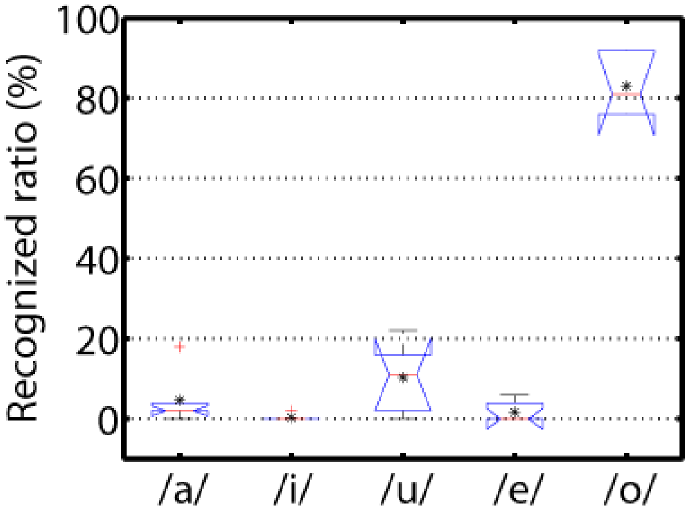

Figure 4.7. Recognized results of spoken vowel /o/.

6. Discussion

It has been confirmed in this experiment that the all-channel method has achieved considerably higher recognition accuracies for the five Japanese vowels than the single-channel method, although the oblique and lateral directions have not been investigated in this study. This result indicates that using electrode grid is more effective in extracting information for sEMG-based speech recognition than using conventional electrodes.

As shown in **Fig. 4.2**, the single-channel method is influenced by the locations of the selected electrodes. In addition, inter-individual variability is also shown in **Fig. 4.2**. Therefore, when conventional disc or parallel bar electrodes are used, it is highly important to carefully consider those locations with respect to each subject might be required in order to achieve higher recognition accuracies. But, doing such tests for each and every subject seems to be rather impractical. One of the reasons for this inter-individual variability is that there are differences in anatomical structures and muscular coordination patterns. To take into account the anatomical structure, magnetic resonance imaging (MRI) of the lower position of the face and neck should be useful [49].

From the confusion matrices shown by **Table 4.2** and **4.3** and the boxplots shown by **Fig. 4.3** and **4.6**, there is a possibility that vowels /a/ and /e/ cannot be discriminated with high accuracy from each other when only sEMG signals from the submental region are used. This finding is consistent with another previous study. By using three parallel bar electrodes, Manabe et al. [32] conducted an experiment of Japanese vowel recognition based on sEMG signals measured from the orbicularis oris, the zygomaticus major, and the anterior belly of the digastric during mouthed speech. However, only the RMS values of the signals were used as features. Although there seemed to be difficulty in vowel recognition using the RMS value from the anterior belly of the digastric, the RMS value from the orbicularis oris could contribute significantly to the discrimination of the vowels /a/ and /e/ (**Fig. 4.8**). Indeed, there is a difference in the condition of the usual voiced speech and voiceless mouthed speech between our experiment and that of Manabe et al. But their experiment implies that additional measurement from the orbicularis oris can improve the recognition accuracy of our proposed method in discriminating between vowels /a/ and /e/. Moreover, consonants should also be considered in future studies. In order to achieve it, sEMG signals from

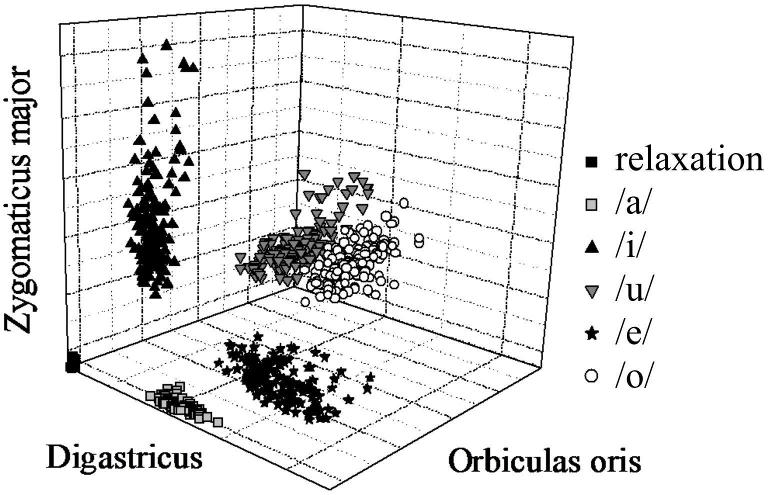

Figure 4.8. 3D distribution in feature space of a previous study. This image is reprinted from [32].

other perioral muscles must be considered as well.

On the other hand, the tradeoff for the dense measurement given by an electrode grid is that it may contain signal redundancy. It is therefore necessary to reduce this redundancy, considering the spatial inter-individual variability as well, especially when working with dysarthric patients. To this end, experiments must be conducted with more subjects.

7. Conclusion

This study proposed the use of an electrode grid for Japanese vowel recognition based on surface electromyography (sEMG). We compared the recognition accuracies of five Japanese vowels between two methods: the all-channel method which used an electrode grid, and the single-channel method which used a virtually reconstructed single bipolar signal. The former achieved recognition accuracies of approximately 80 to

85%, which was higher than that of the latter. This result indicates that using an electrode grid is more effective in extracting information for sEMG-based speech recognition than using a conventional disc or parallel bar electrode. Furthermore, future works on obtaining the findings for spatial inter-individual variability of sEMG signals and reducing the redundant electrodes are warranted.

Chapter 5

Feature Selection for Vowel Recognition

1. Introduction

In the vowel recognition experiments in chapter 4, using the electrode grid realizes denser measurements and brings more information about speech. However, in using the electrode grid, new problems such as unfavorable cost increase of both device and computation arise due to the redundancies of some signals which consequently lead to redundant features. To alleviate this problem, we introduce a feature selection method. We apply sparse discriminant analysis (SDA) [7–9] which was proposed by Clemmensen et al. as a solution, and investigate how this type of feature selection influences the accuracy of vowel recognition.

The cepstral coefficients are employed as features for this study, because the cepstral coefficients indicated higher recognition accuracies than the time domain features as shown in chapter 4. The cepstral coefficients were extracted from the windowed signals of each channel obtained from Subject 1. The real parts of the lower 15 cepstral coefficients (including the 0th coefficients), Δ features, and $\Delta\Delta$ features were used as features. The features from all 56 channels were concatenated. This concatenation resulted in having 2520 feature dimensions.

2. Sparse Discriminant Analysis

Although, in our preliminary study, dimension reduction was performed using linear discriminant analysis (LDA), in this study, we used sparse discriminant analysis (SDA) [7–9] proposed by Clemmensen et al. instead of LDA (SDA software in MATLAB is available from [7]). SDA can perform feature selection simultaneously with dimension reduction by imposing sparseness constraint.

Let $\mathbf{X}$ denote an $n \times p$ data matrix with observations down the rows and features in the columns, and let $\mathbf{Y}$ denote an $n \times K$ (classes) matrix of dummy variables which indicate belonging classes. Clemmensen et al. defined the sparse optimal scoring criterion as

$$\arg\min_{\boldsymbol{\theta},\boldsymbol{\beta}} n^{-1}\Big(\|\mathbf{Y}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\,\|\boldsymbol{\Omega}^{\frac{1}{2}}\boldsymbol{\beta}\|_2^2 + \gamma\,\|\boldsymbol{\beta}\|_1\Big)\ , \tag{5.1}$$

$$\text{subject to } n^{-1}\,\|\mathbf{Y}\boldsymbol{\theta}\|_2^2 = 1\ , \tag{5.2}$$

where $\boldsymbol{\beta}$ is a $p \times q$ matrix of parameters which leads to q components of directions, $\boldsymbol{\theta}$ is $K \times q$ matrix of scores, λ and γ are nonnegative tuning parameters, and $\boldsymbol{\Omega}$ is a symmetric positive definite matrix. This method involves recasting the classification problem as a regression problem by turning categorical variables into quantitative variables, via $\boldsymbol{\theta}$. Iterative algorithm is used for finding a local minimum of the criterion (5.1) with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. For fixed $\boldsymbol{\theta}$, $\boldsymbol{\beta}_j$, $j = 1, \ldots, q$, is obtained by solving the modified elastic net problem [60]:

$$\boldsymbol{\beta}_j = \arg\min_{\boldsymbol{\beta}_j} n^{-1}\left(\|\mathbf{Y}\boldsymbol{\theta}_j - \mathbf{X}\boldsymbol{\beta}_j\|_2^2 + \lambda\boldsymbol{\beta}_j^T\boldsymbol{\Omega}\boldsymbol{\beta}_j + \gamma\,\|\boldsymbol{\beta}_j\|_1\right)\ . \tag{5.3}$$

When γ is large, the L_1 penalty on $\boldsymbol{\beta}_j$ results in sparseness. For fixed $\boldsymbol{\beta}$, the criterion becomes

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} n^{-1}\,\|\mathbf{Y}\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta}\|_2^2\ , \tag{5.4}$$

$$\text{subject to } n^{-1}\,\|\mathbf{Y}\boldsymbol{\theta}\|_2^2 = 1\ . \tag{5.5}$$

Steps related to the equations (5.3) and (5.4) are iterated until convergence or until a maximum number of iterations is reached. In the SDA software, the desired number of features can be set instead of γ.

In this study, we investigated the relationships between the recognition accuracies and the numbers of selected features with λ set to 0, 0.01, 0.1, 1, 10, and 100. The number of components q was set to 4, which was equal to that of LDA in our preliminary study.

3. Results

Fig. 5.1 shows the changes in recognition accuracies with the numbers of selected features per component under varying value of λ. However, in this experiment, change in λ seems to have little influence on the recognition accuracies, especially in the range of higher accuracies. In the chapter 4, the recognition accuracy of 85.6% was achieved by applying LDA to all of the 2520 features, as shown in **Fig. 5.1** by the upper dashed line. Indeed, the LDA method outperformed the SDA method, but our main purpose is to investigate how feature selection influences the recognition accuracy. If there is redundancy in the obtained features, SDA based method can reduce the numbers of selected features without decline in the recognition accuracy. Actually, it can be seen in **Fig. 5.1** that even if the features are compressed to one fifth of the total features, recognition accuracies are still kept over 80%. However, as the number of selected features decrease from 100 to 20, the recognition accuracies decrease steeply. In the chapter 4, we also obtained the recognition accuracies by using LDA on virtually reconstructed bipolar single channel signals which were calculated from all possible combinations of electrodes within the same column of electrode grid. These virtually reconstructed signals could be equivalent to signals measured by one channel of conventional bipolar electrodes. The best recognition accuracy among all of the virtually reconstructed signals is shown to be 51.6% and is indicated by the lower dashed line in **Fig. 5.1**.

Additionally, we investigated the numbers of selected times of features with respect to the corresponding channels, orders of the cepstral coefficients, and the difference in cepstral coefficients, Δ features, and $\Delta\Delta$ features. The results of this investigation are shown in **Fig. 5.2** and **5.3**. These results are taken from the case when the number of selected features per component is 100 and λ is set to 0.01. In **Fig. 5.2**, there are 28 channels, equivalent to half of all the channels, whose numbers of selected times are 0 to 4. Most of the channels from column 4 to 8 are regarded as redundant. **Fig. 5.3** shows that the cepstral coefficients, Δ features, and $\Delta\Delta$ features whose orders are 6 or higher are selected 4 times or less. The difference in cepstral coefficients, Δ features, and $\Delta\Delta$ features has little or no influence on the numbers of selected times.

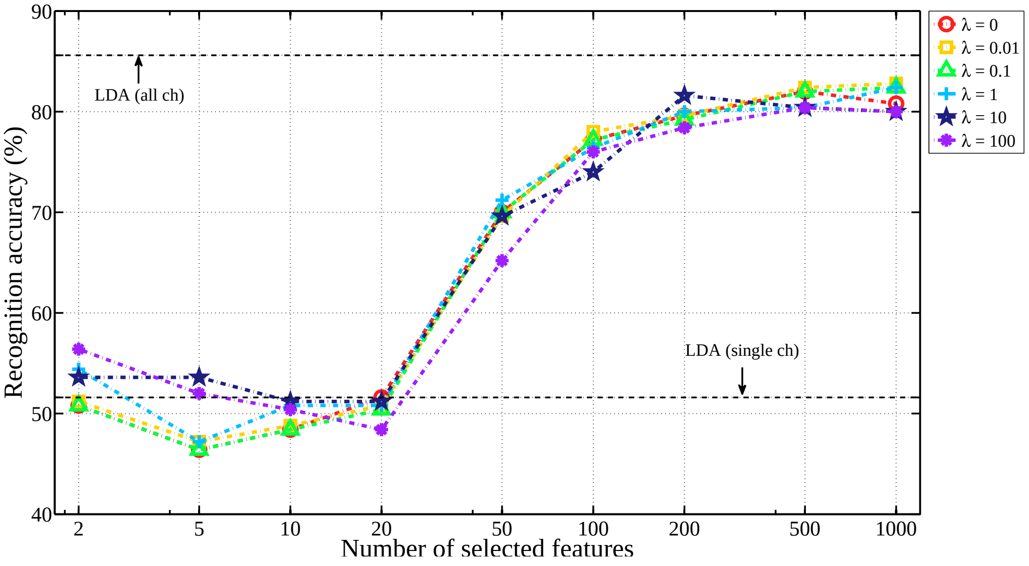

Figure 5.1. Changes of recognition accuracies with the numbers of selected features. "*LDA (all ch)*" denotes the recognition accuracy obtained by using LDA on the signals from all channels in the chapter 4. "*LDA (single ch)*" denotes the best recognition accuracy obtained by using LDA on virtually reconstructed bipolar single channel signal among all possible combinations of electrodes within the same column of electrode grid.

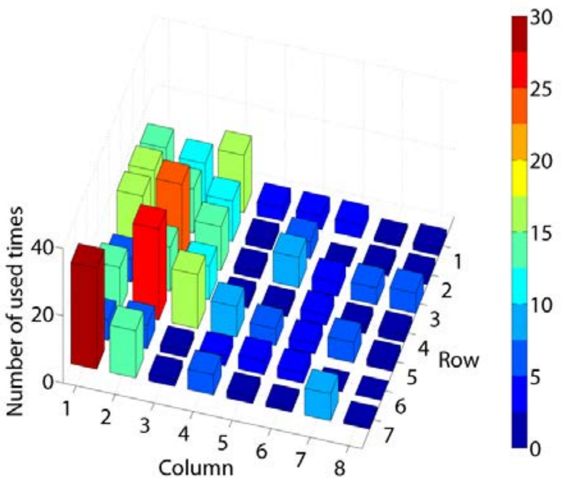

Figure 5.2. The numbers that electrode channels were selected. "*Row*" and "*Column*" denote the row and column of electrode grid, respectively.

4. Discussion

It was illustrated in **Fig. 5.1** that feature selection compressing to 100 or 200 feature dimensions can be realized while discriminative powers are kept relatively higher. From the point of view on the appropriate numbers of channels to extract information about speech, it was suggested from **Fig. 5.1** and **5.2** that using more than 28 channels were preferable and that 1 channel of conventional bipolar electrodes is insufficient to achieve acceptable recognition accuracy. As for features, regardless of difference in cepstral coefficients, Δ features, and ΔΔ features, higher order coefficients were regarded as redundant. Those values can have relatively higher correlation with each other due to the nature of the cepstral coefficients, therefore they may tend to be regarded as redundant. To deal with the tradeoff between cost for device and computation and recognition accuracy, combination of dense measurement based on the electrode grid and the feature selection method based on SDA is able to provide valuable information as shown in this study. Redundant channels and cepstral coefficients can be removed for the purpose to reduce device cost or computational cost.

However, the reason why the channels in column 4 to 8 in **Fig. 5.2** are regarded as redundant might be due to left-right symmetry of anatomical structure. **Fig. 5.4** is

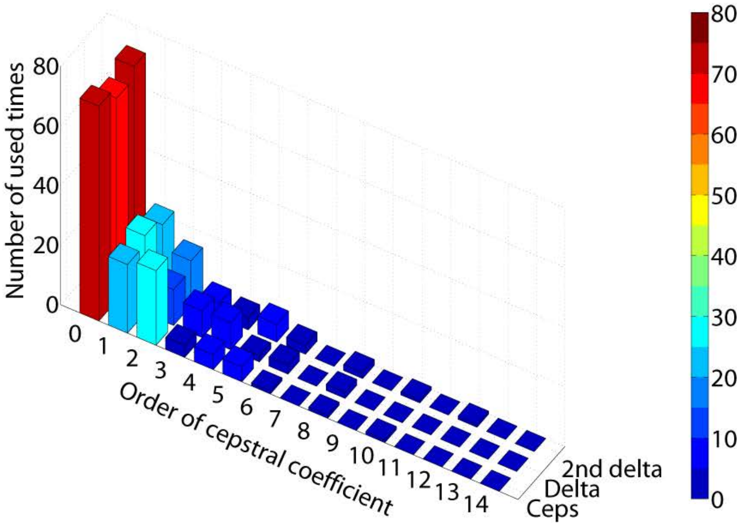

Figure 5.3. The numbers that cepstral coefficients, Δ features, and ΔΔ features were selected. “*Ceps*”: Cepstral coefficients, “*Delta*”: Δ features, “*2nd delta*”: ΔΔ features.

anatomical image of relatively superficial layer of submental region in the horizontal sectional view. Red shaded part covers column 1 to 3 and it is roughly in accordance with left one of the muscles called “anterior belly of the digastric”. On the other hand, it is possible that patients with dysarthria due to paralysis may have laterality in their sEMG signals. If that is the case, the channels regarded as redundant in this study should not be removed for dysarthric patients. Experiments with dysarthric patients are therefore essential in deciding appropriate electrode location for them.

5. Conclusion

This chapter investigated how feature selection influences the accuracy of vowel recognition based on sEMG derived with a multichannel electrode grid. We applied SDA for feature selection to cope with redundant signals. It was illustrated that feature selection compressing to one tenth or one twentieth of the total features could be achieved without steep decline in recognition accuracies. In addition, the redundant channels and features were specified by using SDA. Thus, combination of dense measurement based on the electrode grid and the feature selection based on SDA is an effective

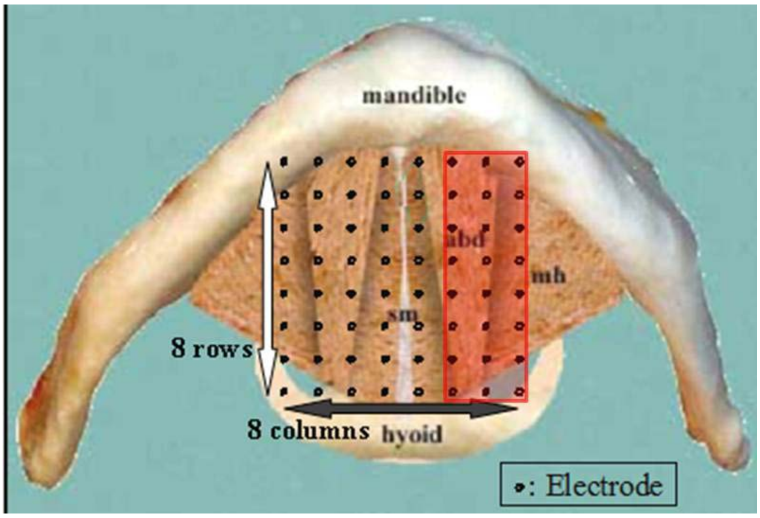

Figure 5.4. Relationship between selected channels and anatomical structure

approach for the researches on sEMG-based speech recognition which has not been established very well yet.

Chapter 6

Conclusions

1. Summary

This dissertation proposed the use of an electrode grid for Japanese vowel recognition based on surface electromyography (sEMG). First, we confirmed that innervation zone of anterior belly of digastic can be roughly estimated.

Next, we compared the recognition accuracies of five Japanese vowels between two methods: the all-channel method which used an electrode grid, and the single-channel method which used a virtually reconstructed single bipolar signal. The former achieved recognition accuracies of approximately 80 to 85%, which was higher than that of the latter. This result indicates that using an electrode grid is more effective in extracting information for sEMG-based speech recognition than using a conventional disc or parallel bar electrode.

Also, this dissertation investigated how feature selection influences the accuracy of vowel recognition based on sEMG derived with a multichannel electrode grid. We applied SDA for feature selection to cope with redundant signals. It was illustrated that feature selection compressing to some extent could be achieved without steep decline in recognition accuracies. In addition, the redundant channels and features were specified by using SDA. Hemi-lateral side of submental region was regarded as redundant in this dissertation.

2. Future works

For our future work, it is necessary to explore appropriate electrode locations not only on the submental region but also on the lower face and neck region, especially when considering recognition of consonants. Similarly, exploring with higher dimensional features which includes but is not limited to various time domain features, frequency domain features, and wavelet coefficients will be significant in extracting more information about speech. Because sEMG-based speech recognition has not been established very well yet, it seems unlikely that such explorations can be done without including redundant data. In addition, in this scenario, there will be high dimensional low sample size setting which will be problematic. Although LDA is more likely to cause overfitting in the high dimensional low sample size setting, SDA has the potential to be more effective because of its capability to reduce overfitting. Furthermore, SDA is straightforwardly extended to sparse mixture discriminant analysis (SMDA) [8, 9] which can deal with mixture of Gaussians. SMDA can be suitable for word recognition and continuous speech recognition. SMDA in MATLAB is also available from [7]. Thus, SDA and SMDA have great potentials to be effective tools for the researches on sEMG-based speech recognition.

Besides, obtaining the findings for spatial inter-individual variability of sEMG signals are warranted. This spatial inter-individual variability is largely due to difference in anatomical structure and coordination pattern of muscles. To take into account the anatomical structure, magnetic resonance imaging (MRI) of the lower position of the face and neck should be useful [49]. **Fig. 6.1** shows example of MRI image of them [1][1]. Source localization [52, 53] based on sEMG signals can be also useful for speech recognition, when the anatomical structure obtained by MRI is utilized as the constraint.

Our actual goal is clinical application of sEMG-based speech recognition to dysarthric patients. sEMG signals of patients with some kind of dysarthria present neurogenic change of EMG signals [61]. To achieve high recognition accuracy with dysarthric

[1] The MRI data used in this study is part of “ATR MRI data of Japanese vowel production” that were acquired at and released from ATR Human Information Science Laboratories under “Research of Human Communication” funded by the National Institute of Information and Communications Technology. The use of the database and release of the results are under the license agreement with ATR-Promotions Co. Ltd.

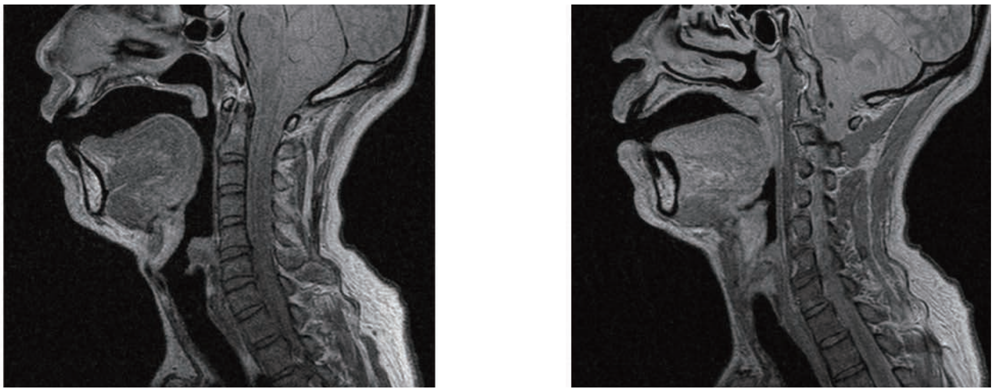

Figure 6.1. MRI image of vocal tract

patients, we have to consider change in firing rate and amplitude. Experiments with dysarthric patients and accumulation of their data are necessary to investigate feasibility of sEMG-based speech recognition with dysarthric patients. Our proposed method is also applicable as rehabilitation aid by providing feedback information derived from sEMG signals to the patients. Since tongue is less visible compared with upper limb or lower limb, feedback about tongue state will be more useful for rehabilitation.

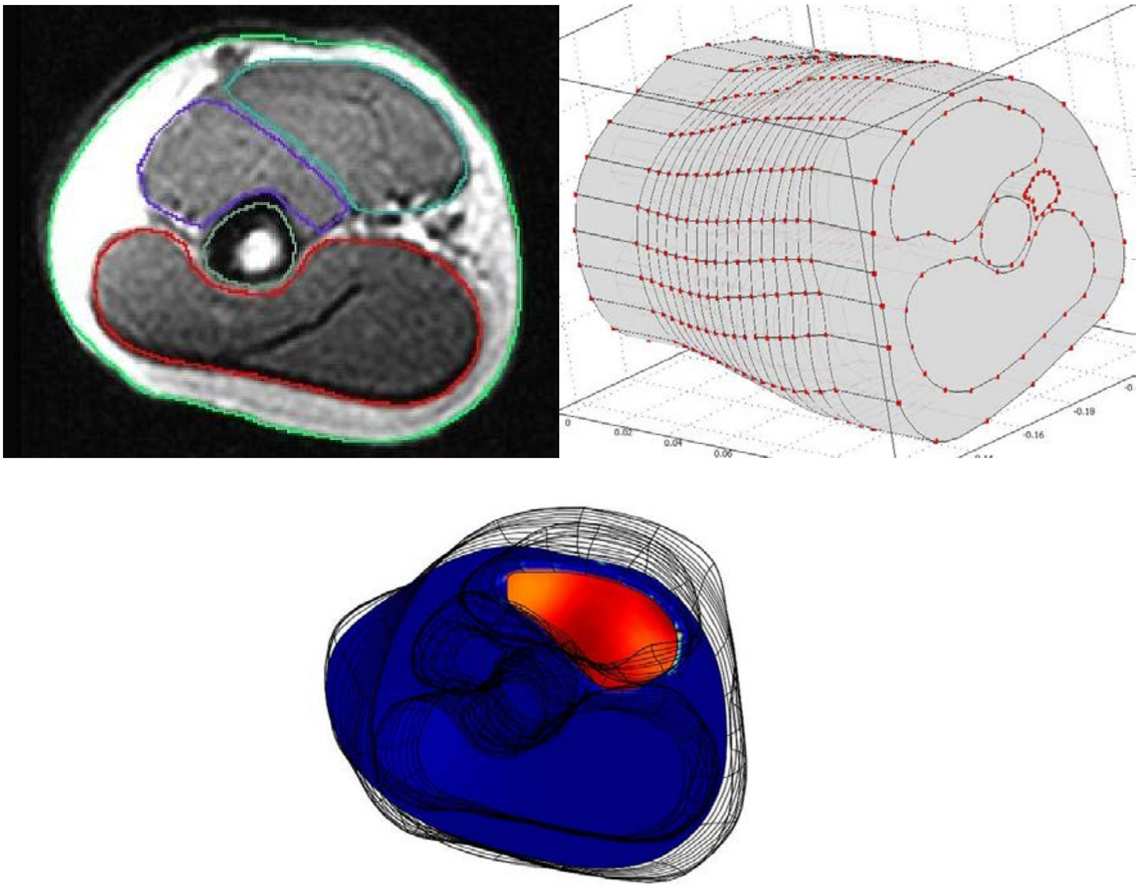

Figure 6.2. (*upper left*) MRI image of arm, (*upper right*) 3D geometry mode constructed from the MRI datal, (*bottom*) Reconstructed activities. These images are reprinted from [53]

.

Acknowledgements

I would like to thank Alphatec Corp., Mr. Yunbin Deng, Mr. Szu-Chen Jou, Mr. Sami Lemmetty, Mr. Wesley Norman, Mr. Hiroyuki Manabe, Mr. Kees van den Doel, ATR-Promotions Co. Ltd., Prof. Tadashi Masuda and Prof. Tohru Kiryu for giving me their kind permission to reprint the figures used in this dissertation.

謝辞

本研究の一部は、奈良先端科学技術大学院大学支援財団より2010年4月から2012年3月の間に助成を受け、実施しました。関係各位に心より深く感謝申し上げます。

主指導教官である池田和司教授には、正しく右も左も分からない状況で研究に身を投じた私に対し、灯台の如く導き続けて頂きました。今日に至ることができたことへの池田教授に対する感謝は筆舌に尽くし難いものです。誠にありがとうございました。この御恩を一生忘れることはありません。また、池田教授の助言や指摘の一言一言に蒙が啓れ、自分なりに吸収・成長できたと感じております。私の人生において非常に貴重で、有意義な経験でした。重ねて、厚くお礼申し上げます。

大阪電気通信大学　吉田正樹教授より生体計測、研究内容の論文化などの多岐に渡るご助言を幾度となく頂き、そのおかげを持ちまして、私の研究結果を曲がりなりにも世に出すことができました。心より感謝致しております。また、筋電図計測システム開発および発話時表面筋電図計測実験に多大なご協力を頂きました生命機能計測学講座　服部託夢特任助教（当時　福祉のまちづくり研究所）に厚くお礼申し上げます。

音声認識処理に関してご助言を頂きました、知能コミュニケーション研究室　戸田智基准教授、音情報処理学研究室　鹿野清宏教授に深く感謝致します。戸田准教授には、私のために多くの時間を割いて頂き、そしてその度に示唆に富んだ貴重なコメントを頂き、本当に感謝致しております。研究者としての在り方も、大変参考になりました。鹿野教授には、見通しが不明瞭な本研究始動段階において音声研究者の見地からのご意見・ご批判を頂き、研究内容のより深い思索を巡らせるための契機を与えて頂きました。心より感謝申し上げます。

本研究の遂行において有益なご助言を頂きました、柴田智広准教授、竹之内高志助教、渡辺一帆助教、林浩平さん、宮本敦志君を始めとする数理情報学研究室の諸氏、および計算神経科学講座所属の間島慶君、福嶋誠君に心よりお礼申し

上げます。また、Jimson Ngeo、Mauricio Burdelis には、論文の英語表現に関する指導をいつも快く引き受けて頂きました、ここに深謝致します。研究に必要な物品の購入等で、数理情報学研究室の谷本史さん、足立敏美さんに大変お世話になりました。本当にありがとうとざいました。

そして、最後にこの紙面を借りて、妻まどか、父母、義理の父母に感謝の気持ちを述べさせて頂きたいと思います。私の研究のために、多くの犠牲を厭わず、献身的に協力してくれた妻まどかの支えが無ければ、決して私の研究は現在の状況にまで至ることはできませんでした。本当に、感謝の気持ちで一杯です。ありがとう。この感謝の念をいつまでも胸に抱き、支えに報いて行きたいと思っています。父母、義理の父母には、いつも私達の生活を気遣ってもらい、様々な形で私の研究生活を支援して頂きました。過分な幸福に、申し訳なくすら思っております。本当にありがとうございました。父母には、この世に生を授けてもらい、育て上げて頂いたことを改めて心から感謝致します。ありがとうございました。

References

[1] ATR-Promotions. BAIC 発話 MRI データ.
Online: http://www.baic.jp/product/artic.html.

[2] T. Baer, J. Alfonso, and K. Honda. Electromyography of the tongue muscle during vowels in /pvp/ environment. *Ann Bull RILP*, Vol. 22, pp. 7–18, 1988.

[3] B. Blaney and J. Wilson. Acoustic variability in dysarthria and computer speech recognition. *Clinical Linguistics & Phonetics*, Vol. 14, No. 4, pp. 307–327, 2000.

[4] A. D. C. Chan, K. B. Englehart, B. Hudgins, and D.F. Lovely. Multiexpert automatic speech recognition using acoustic and myoelectric signals. *IEEE Trans Biomed Eng*, Vol. 53, No. 4, pp. 676–85, 2006.

[5] A.D.C. Chan, K. Englehart, B. Hudgins, and D.F. Lovely. Hidden markov model classification of myoelectric signals in speech. In *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, Vol. 2, pp. 1727–1730. IEEE, 2001.

[6] A.D.C. Chan, K. Englehart, B. Hudgins, and D.F. Lovely. Myo-electric signals to augment speech recognition. *Medical and Biological Engineering and Computing*, Vol. 39, No. 4, pp. 500–504, 2001.

[7] L. Clemmensen. Line Clemmensen.
Online: http://www2.imm.dtu.dk/ lhc/index.html.

[8] L. Clemmensen, T. Hastie, and K. Ersboell. Sparse discriminant analysis. Technical report, IMM, Technical University of Denmark, 2008.

[9] L. Clemmensen, T. Hastie, D. Witten, and B. Ersbll. Sparse discriminant analysis. *Technometrics*, Vol. 53, No. 4, pp. 406–413, 2011.

[10] ALPHAEC Corp. a-paso.com.
Online: http://www.a-paso.com.

[11] F.L. Darley, A.E. Aronson, and J.R. Brown. *Motor speech disorders*. Saunders, 1975.

[12] C. J. De Luca. Surface Electromyography: Detection and Recording. Online: http://www.delsys.com/Attachments_pdf/WP_SEMGintro.pdf.

[13] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B-Statistical Methodology*, Vol. 39, No. 1, pp. 1–38, 1977.

[14] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg. Silent speech interfaces. *Speech Communication*, Vol. 52, No. 4, pp. 270–287, 2010.

[15] Y. Deng, R. Patel, J. T. Heaton, G. Colby, L. D. Gilmore, J. Cabrera, S. H. Roy, C. J. De Luca, and G. S. Meltzner. Disordered speech recognition using acoustic and sEMG signals. In *INTERSPEECH*, pp. 644–647, 2009.

[16] G. Drost, D.F. Stegeman, B.G.M. van Engelen, and M.J. Zwarts. Clinical applications of high-density surface emg: A systematic review. *Journal of Electromyography and Kinesiology*, Vol. 16, No. 6, pp. 586 – 602, 2006.

[17] J.R. Duffy. *Motor Speech Disorders: Substrates, Differential Diagnosis, and Management*. Elsevier Mosby, St Louis, MO, second edition, 2005.

[18] S.K. Fager, D.R. Beukelman, T. Jakobs, and J.P. Hosom. Evaluation of a speech recognition prototype for speakers with moderate and severe dysarthria: A preliminary report. *Augmentative and Alternative Communication*, Vol. 26, No. 4, pp. 267–277, 2010.

[19] O. Fukuda, S. Fujita, and T. Tsuji. A substitute vocalization system based on emg signals. *IEICE Transactions on Information and Systems*, Vol. J88-D-2, No. 1, pp. 105–112, 2005.

[20] T. Hattori, T. Sato, K. Minato, H. Nakamura, and M. Yoshida. An identification method of motor units using a 3d template from grid surface electromyography. *Trans Jpn Soc Med Biol Eng*, Vol. 46, No. 2, pp. 268–274, 2008.

[21] M. Hawley, P. Enderby, P. Green, S. Brownsell, A. Hatzis, M. Parker, J. Carmichael, S. Cunningham, P. O'Neill, and R. Palmer. STARDUST; speech

training and recognition for dysarthric users of assistive technology. In *Proceedings of the 7th European Conference for the Advancement of Assistive Technology in Europe*, pp. 959–964, 2003.

[22] M.S. Hawley, P. Enderby, P. Green, S. Cunningham, S. Brownsell, J. Carmichael, M. Parker, A. Hatzis, P. O'Neill, and R. Palmer. A speech-controlled environmental control system for people with severe dysarthria. *Medical engineering & physics*, Vol. 29, No. 5, pp. 586–593, 2007.

[23] C. Jorgensen and K. Binsted. Web browser control using EMG based sub vocal speech recognition. In *System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on*, pp. 294c–294c, 2005.

[24] C. Jorgensen, D.D. Lee, and S. Agabont. Sub auditory speech recognition based on emg signals. In *Neural Networks, 2003. Proceedings of the International Joint Conference on*, Vol. 4, pp. 3128–3133. IEEE, 2003.

[25] S. C. Jou, L. Maier-Hein, T. Schultz, and A. Waibel. Articulatory feature classification using surface electromyography. In *ICASSP*, pp. 605–608, Toulouse, France, May 2006.

[26] S. C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel. Towards continuous speech recognition using surface electromyography. In *INTERSPEECH*, pp. 573–576, Pittsburgh, PA, Sep 2006.

[27] B. G. Lapatki, R. Oostenveld, J. P. Van Dijk, I. E. Jonas, M. J. Zwarts, and D. F. Stegeman. Topographical characteristics of motor units of the lower facial musculature revealed by means of high-density surface EMG. *Journal of Neurophysiology*, Vol. 95, No. 1, pp. 342–354, 2006.

[28] B. G. Lapatki, R. Oostenveld, J. P. Van Dijk, I. E. Jonas, M. J. Zwarts, and D. F. Stegeman. Optimal Placement of Bipolar Surface EMG Electrodes in the Face Based on Single Motor Unit Analysis. *Psychophysiology*, Vol. 47, No. 2, pp. 299–314, 2010.

[29] B. G. Lapatki, J. P. van Dijk, I. E. Jonas, M. J. Zwarts, and D. F. Stegeman. A thin, flexible multielectrode grid for high-density surface EMG. *Journal of Applied Physiology*, Vol. 96, No. 1, pp. 327–336, 2004.

[30] S. Lemmetty. Review of speech synthesis technology. Master's thesis, Helsinki University of Technology, 1999.

[31] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel. Session independent non-audible speech recognition using surface electromyography. In *2005 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 331–336. IEEE, 2005.

[32] H. Manabe, A. Hiraiwa, and T. Sugimura. Unvoiced speech recognition using emg - mime speech recognition. In *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, pp. 794–795, New York, NY, USA, 2003. ACM.

[33] T. Masuda and T. Sadoyama. Topographical map of innervation zones within single motor units measured with a grid surface electrode. *IEEE Transactions on Biomedical Engineering*, Vol. 35, No. 8, pp. 623–628, 1988.

[34] H. Matsumasa, T. Takiguchi, Y. Ariki, I. LI, T. Nakabayashi, et al. Integration of metamodel and acoustic model for dysarthric speech recognition. *Journal of Multimedia*, Vol. 4, No. 4, pp. 254–261, 2009.

[35] G. S. Meltzner, J. Sroka, J. T. Heaton, L. D. Gilmore, G. Colby, S. H. Roy, N. Chen, and C.J. De Luca. Speech recognition for vocalized and subvocal modes of production using surface EMG signals from the neck and face. In *Ninth Annual Conference of the International Speech Communication Association*, pp. 2667–2670, 2008.

[36] R. Merletti and P.A. Parker. *Electromyography: Physiology, Engineering, and Noninvasive Applications*. Wiley-IEEE Press, 2004.

[37] 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*, Vol. 16, No. 6, pp. 399–410, 1986.

[38] K. Murphy. Hidden Markov Model (HMM) Toolbox for Matlab.
Online: http://www.cs.ubc.ca/ murphyk/Software/HMM/hmm.html.

[39] W. Norman. The Anatomy Lesson. Online:http://www.wesnorman.com/lesson5.htm.

[40] S.H. Park and S.P. Lee. EMG Pattern Recognition Based on Artificial Intelligence Techniques. *Rehabilitation Engineering, IEEE Transactions on*, Vol. 6, No. 4, pp. 400–405, 1998.

[41] A. Phinyomark, C. Limsakul, and P. Phukpattaranont. A novel feature extraction for robustEMG pattern recognition. *Journal of Computing*, Vol. 1, No. 1, pp. 71–80, 2009.

[42] A. Phinyomark, C. Limsakul, and P. Phukpattaranont. Application of wavelet analysis in emg feature extraction for pattern classification. *Measurement Science Review*, Vol. 11, No. 2, pp. 45–52, 2011.

[43] P. Raghavendra, E Rosengren., and S. Hunnicutt. An investigation of different degrees of dysarthric speech as input to speaker-adaptive and speaker-dependent recognition systems. *Augmentative and Alternative Communication*, Vol. 17, No. 4, pp. 265–275, 2001.

[44] F. Rudzicz. Using articulatory likelihoods in the recognition of dysarthric speech. *Speech Communication*, Vol. 54, No. 3, pp. 430–444, 2012.

[45] E.J. Scheme, B. Hudgins, and P.A. Parker. Myoelectric signal classification for phoneme-based speech recognition. *Biomedical Engineering, IEEE Transactions on*, Vol. 54, No. 4, pp. 694–699, 2007.

[46] T. Schultz and M. Wand. Modeling coarticulation in EMG-based continuous speech recognition. *Speech Communication*, Vol. 52, No. 4, pp. 341–353, 2010.

[47] G.L. Soderberg, National Institute for Occupational Safety, and Health. *Selected Topics in Surface Electromyography for Use in the Occupational Setting: Expert Perspectives*. U.S. Dept. of Health and Human Services, Public Health Service, Centers for Disease Control, National Institute for Occupational Safety and Health, 1992.

[48] N. Sugie and K. Tsunoda. A speech prosthesis employing a speech synthesizer-vowel discrimination from perioral muscle activities and vowel production.

Biomedical Engineering, IEEE Transactions on, Vol. BME-32, No. 7, pp. 485–490, 1985.

[49] S. Takano and K. Honda. An mri analysis of the extrinsic tongue muscles during vowel production. *Speech communication*, Vol. 49, No. 1, pp. 49–58, 2007.

[50] F.J. Theis and G.A. García. On the use of sparse signal decomposition in the analysis of multi-channel surface electromyograms. *Signal processing*, Vol. 86, No. 3, pp. 603–623, 2006.

[51] N. Thomas-Stonell, A. L. Kotler, H. Leeper, and P. Doyle. Computerized speech recognition: Influence of intelligibility and perceptual consistency on recognition accuracy. *Augmentative and Alternative Communication*, Vol. 14, No. 1, pp. 51–56, 1998.

[52] K. van den Doel, U.M. Ascher, and D.K. Pai. Computed myography: Three-dimensional reconstruction of motor functions from surface emg data. *Inverse Problems*, Vol. 24, No. 6, p. 065010, 2008.

[53] K. van den Doel, U.M. Ascher, and D.K. Pai. Source localization in electromyography using the inverse potential problem. *Inverse Problems*, Vol. 27, No. 2, p. 025008, 2011.

[54] M. Walliczek, F. Kraft, S. C. Jou, T. Schultz, and A. Waibel. Sub-word unit based non-audible speech recognition using surface electromyography. In *INTERSPEECH*, pp. 1487–1490, 2006.

[55] M. Wand and T. Schultz. Towards speaker-adaptive speech recognition based on surface electromyography. In *BIOSIGNALS*, pp. 155–162, 2009.

[56] V. Young and A. Mihailidis. Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review. *Assistive technology the official journal of RESNA*, Vol. 22, No. 2, pp. 99–112; quiz 113–114, 2010.

[57] M. Zecca, S. Micera, MC Carrozza, P. Dario, et al. Control of multifunctional prosthetic hands by processing the electromyographic signal. *Critical Reviews in Biomedical Engineering*, Vol. 30, No. 4-6, p. 459, 2002.

[58] W. R. Zemlin. ゼムリン 言語聴覚学の解剖生理 原著第4版. 医歯薬出版, 東京都, 2006. 舘村 卓 監訳, 浮田 弘美, 山田 弘幸 訳.

[59] Q. Zhou, N. Jiang, K. Englehart, and B. Hudgins. Improved phoneme-based myoelectric speech recognition. *IEEE Trans Biomed Eng*, Vol. 56, No. 8, pp. 2016–23, 2009.

[60] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B-Statistical Methodology*, Vol. 67, pp. 301–320, 2005.

[61] 木村淳, 幸原伸夫. 神経伝導検査と筋電図を学ぶ人のために. 医学書院, 東京都, 2003.

[62] 西尾正輝. ディサースリアの基礎と臨床 第1巻 理論編. インテルナ出版, 東京都, 2006.

[63] 西尾正輝. ディサースリアの基礎と臨床 第3巻 臨床実用編. インテルナ出版, 東京都, 2006.

[64] 服部託夢. 格子状多点誘導表面筋電図の軸間情報を用いた運動単位同定手法に関する研究. 博士論文, 奈良先端科学技術大学院大学, 2009.

[65] 木塚朝博, 増田正, 木竜徹, 佐渡山亜兵. 表面筋電図. 東京電気大学出版局, 東京都, 2006.

[66] 木竜徹. EMG Website.
Online: http://earc.eng.niigata-u.ac.jp/.

List of Publications

Journal Paper

1. Takatomi Kubo, Tomoki Toda, Masaki Yoshida, Takumu Hattori, Kazushi Ikeda: Vowel Recognition Based on Surface Electromyography with Electrode Grid on Submental Region. Trans Jpn Soc Med Biol Eng, Vol. 50, No. 1, 2012. Accepted, to appear.

International Conference

1. Takatomi Kubo, Masaki Yoshida, Takumu Hattori, Kazushi Ikeda: Feature Selection for Vowel Recognition Based on Surface Electromyography Derived with Multichannel Electrode Grid. Workshop on Intelligence Science and Intelligent Data Engineering, Xi'an, China, October 2011. (To appear in "Lecture Notes in Computer Science".)

National Conference

1. 久保 孝富, 戸田 智基, 服部 託夢, 吉田 正樹, 池田 和司: 顔面・頚部から誘導した低周波電気信号の無発声音声認識における有用性. 第12回日本電気生理運動学会 大会抄録集, pp. 18-19, 横浜, 2011年3月.

2. 久保 孝富, 戸田 智基, 吉田 正樹, 服部 託夢, 池田 和司: Vowel recognition based on surface electromyography with electrode grid on submental region. 生体医工学シンポジウム2011 講演予稿集 CD-ROM, 2-1-7, 長野, 2011年10月.

Research Grant

1. 平成22年度 奈良先端科学技術大学院大学支援財団「教育研究活動支援」助成
   生体情報を利用した構音障害者への意思伝達支援デバイスの研究開発

2. 平成23年度 奈良先端科学技術大学院大学支援財団「教育研究活動支援」助成
   多チャンネル筋電位信号を利用した構音障害者への意思伝達支援デバイスの研究開発