2-2022

Speech Feature Analysis and Discrimination in Biological Information

Shogo Honda
sh4977@rit.edu

Speech Feature Analysis and Discrimination in Biological Information

Shogo Honda

Speech Feature Analysis and Discrimination in Biological Information

Shogo Honda

February 2022

A Thesis Submitted
in Partial Fulfillment
of the Requirements for the Degree of
Master of Science
in
Computer Engineering

RIT | Kate Gleason College of Engineering

Department of Computer Engineering

Speech Feature Analysis and Discrimination in Biological Information

Shogo Honda

Committee Approval:

Andres Kwasinski *Advisor* Date
Professor, Department of Computer Engineering

Cory Merkel *Co-Advisor* Date
Assistant Professor, Department of Computer Engineering

Minoru Nakazawa *Advisor* Date
Professor, Department of Computer Engineering, Kanazawa Institute of Technology

Acknowledgments

Dr. Andres Kwasinski

Dr. Cory E. Merkel

Dr. Minoru Nakazawa

Dr. Sayoko Takano

Dr. Masanori Higuchi

Mr. Celal Savur

Mr. Joseph A. Zonghi

Mr. Tomoki Tokida

RIT Computer Engineering Department

To all those who helped me along the way.

Abstract

A silent speech interface is a system that allows people doing speech communication without using their own speech sounds. Today, a variety of speech interfaces have been developed using biological signals such as the eye movement, and the articulatory. These interfaces are mainly for supporting people who have speech disorder to communicate with others, yet there are many speech disorder that have not been addressed by the current technologies. The possible cause of the issue is the limited numbers of the biological signals used for the speech interface. The uncovered issues with speech disorders can be addressed through identifying new biological signals for speech interface development. Therefore, we aim to find new biological signals that can be used for speech interface developments. The biological signals we focused on were the vibration of the vocal folds and brain waves. After measuring the data and extracting the features, we verified whether this data can be used to classify speech sounds through machine learning models: Support Vector Machine for the vocal folds vibration, and Echo State Network for the brain waves. As a result, using the vocal folds vibration signals, Japanese vowels could be classified with 71 % accuracy on average. Using the brain waves, five different consonants were classified with 28.3 % accuracy on average. These findings indicate the possibility that the vocal folds vibration signals and the brain waves can be used as new biological signals for speech interface developments. From this study, we were able to discover some needed improvements that should be considered in the future that may lead to further improvement in the classification accuracy.

Contents

Signature Sheet i

Acknowledgments ii

Dedication iii

Abstract iv

Table of Contents v

List of Figures vii

List of Tables ix

Acronyms x

1 Introduction 2

1.1 Motivation 2

1.2 Research Question 2

1.3 Contribution 3

1.4 Document Structure 4

2 Background 5

2.1 Vocal Folds Vibration 5

2.1.1 What Is Vocal Folds Vibration 5

2.1.2 Relations Between Vocal Folds Vibration and Speech 6

2.1.3 Speech Interface Study Using Vocal Folds Vibration 6

2.2 Electroencephalography 8

2.2.1 What Is Electroencephalography 8

2.2.2 Relations Between EEG and Speech 8

2.2.3 Study of Speech and EEG 9

2.3 Machine Learning Models 10

2.3.1 Support Vector Machine 11

2.3.2 Recurrent Neural Network 12

2.3.3 Echo State Network 14

3 Japanese Vowel Discrimination From Vocal Folds Vibration 17
3.1 Introduction . . . 17
3.2 Measurements . . . 18
3.2.1 Overview . . . 18
3.2.2 Devices for Experiment . . . 18
3.2.3 Data Collection Procedure . . . 20
3.2.4 Collected Data . . . 21
3.3 Feature Extraction . . . 24
3.3.1 Overview . . . 24
3.3.2 Signal Processing . . . 25
3.3.3 Extracted Data . . . 26
3.4 Classification Using Machine Learning . . . 27
3.4.1 Overview . . . 27
3.4.2 Methodology . . . 27
3.5 Result . . . 29
3.6 Conclusion . . . 31

4 Unvoiced Consonant Prediction from Pre-Speech EEG Data 33
4.1 Introduction . . . 33
4.2 Measurement . . . 34
4.2.1 Overview . . . 34
4.2.2 Devices and Software for Experiment . . . 34
4.2.3 Experiment . . . 37
4.3 Preprocessing . . . 39
4.4 Classification using machine learning . . . 40
4.4.1 Overview . . . 40
4.4.2 Data Structure . . . 41
4.4.3 Echo State Network Model . . . 41
4.4.4 Hyperparameter Adjustment . . . 42
4.5 Evaluation method . . . 49
4.6 Result . . . 50
4.7 Discussion . . . 52
4.8 Conclusion . . . 54

5 Conclusion and Future Work 56

List of Figures

2.1 Illustration of the larynx 5
2.2 Articulatory forms when human speaks /a/ and /i/ sounds 6
2.3 EGG device 7
2.4 The EGG wave of impedance 8
2.5 Measuring electrical activity of synapses in the cortex 9
2.6 Broca's area in the brain 10
2.7 SVM classification of two class dataset 11
2.8 The kernel method of SVM 12
2.9 General Recurrent Neural Network Model 13
2.10 The general Echo State Network model 15

3.1 The small wireless multi-functional sensor (TSND121) 19
3.2 Flow diagram of the data measurement 19
3.3 The software ALTIMA 19
3.4 The acceleration sensor attached to the throat 20
3.5 The procedure of the acceleration measurement 21
3.6 The plot of the acceleration data 22
3.7 Acceleration data of Japanese vowel /a/ 23
3.8 Acceleration data of Japanese vowel /i/ 23
3.9 Acceleration data of Japanese vowel /u/ 23
3.10 Acceleration data of Japanese vowel /e/ 23
3.11 Acceleration data of Japanese vowel /o/ 23
3.12 Flow for classification implementation 24
3.13 Power spectral density of vowel /a/ 26
3.14 Feature map with 2 class dataset: /a/ and /i/ 28
3.15 Feature map with 2 class dataset: /a/ and /u/ 28
3.16 Feature map with 2 class dataset: /a/ and /e/ 28
3.17 Feature map with 2 class dataset: /a/ and /o/ 28
3.18 Classification Learner app setting 29
3.19 Frequency distribution of vowels 31

4.1 EEG cap (EPOC X) 35
4.2 The locations of EMOTIV headset's electrodes 35
4.3 Data streaming on EMOTIV PRO 36

4.4 Setup of Lab Recorder 37
4.5 EEG experiment figure 38
4.6 EEG experimental procedure 39
4.7 The illustration of the data structure after preprocessing 41
4.8 Tanh graph 44
4.9 The nodes' output plot with the input weight's interval in the range of -0.01 to 0.01 45
4.10 The nodes' output plot with the input weight's interval in the range of -1 to 1 46
4.11 The nodes' output plot with the input weight's interval in the range of -100 to 100 47
4.12 Confusion matrix when $\alpha < 0.001$ 48
4.13 Confusion matrix when $\alpha > 0.1$ 49

List of Tables

3.1 Accuracy of Japanese vowel discrimination 30

4.1 Word prompts for the experiment 38
4.2 List of hyperparameters for ESN 43
4.3 Precision rate of each consonant classification 51

Acronyms

BPTT

Backpropagation Through Time

DFT

Discrete Fourier Transform

ECG

electrocardiogram

EEG

Electroencephalography

EGG

electroglottograph

EMG

Electromyography

EOG

Electrooculography

ESN

Echo State Network

GRU

Gated Recurrent Unit

LSL

Lab Streaming Layer

LSTM

Long Short Term Memory

MEG

Magnetoencephalography

MFCC

Mel Frequency Cepstrum Coefficients

PSD

Power Spectral Density

RNN

Recurrent Neural Network

SVM

Support Vector Machine

Chapter 1

Introduction

1.1 Motivation

A silent speech interface is a system that allows people doing speech communication without using their own speech sounds. Today, there has been a lot of research on new communication tools and systems (speech interface) [1] that use biological signals such as the eye movement [2], or the articulatory movement [3] to help people who have language disorders to communicate with others. However, the current speech interface technologies still have challenges, such as low recognition accuracy, instability, and difficulty in controlling, which leaves numerous cases where existing interfaces cannot help individuals with certain speech disorders. [4, 5].

Some current limitations with speech interfaces may be attributed to limitations in the technology but another contributing factor is the limited number of biological signals that have been explored so far. Therefore, in this study we aim to find new biological signals that could be used in silent speech interfaces.

1.2 Research Question

To find new biological signals useful in new speech interfaces, we analyzed biological signals that contain features associated with speech. Furthermore, we verified if these biological signals can be used to discriminate speech or not using machine learning

models. In this study, we focused on two biological signals: (1) the signal from the vocal folds vibration measured by an acceleration sensor, and (2) the brain waves measured by an EEG (electroencephalogram). We considered vocal folds vibration because there are people with speech impediments whose vocal cords work normally [6], so that the signal from the vocal folds vibration can be used as the biological signals of speech interfaces for many language disorders. Secondly, the vocal folds vibration is the first stage when producing speech. [7]. Hence, we expect the vocal folds vibration can be used as a new effective information source for speech interface developments for various language disorders. Regarding the brain waves, the brain controls many functions of the body including speech production [8], which makes conceivable that signals from the brain could be used to develop a speech interface. Therefore, we studied whether brain waves carry speech features or not, and whether it is possible to discriminate speech from the aforementioned biological signals using machine learning model. Specifically, for machine learning model we considered an Echo State Network (ESN).

As part of our research methodology, we measured, we measured biological signals while human subjects vocalized some specific words that contained certain speech sounds of interest.

1.3 Contribution

Our first study resulted 71% accuracy for Japanese vowels' classification by an accelerometer and a Support Vector Machine. This result showed a new finding that vocal fold vibration signals can be used for speech recognition in the field of speech interface research. Moreover, we found that an accelerometer can measure sufficient data from the vocal folds vibration.

In the field of speech and hearing science, the relationship between pre-speech brain activity and speech has been paid much attention. We implemented speech

sound classification using pre-speech Electroencephalography (EEG) data and found the possibility that EEG data contain features that discriminate speech sounds. In the field of computer engineering, optimal machine learning models for speech classification using EEG data have been researched. In this study, we built an Echo State Network model and found some tips for the implementation of speech sound classification using EEG data.

1.4 Document Structure

Chapter 2 provides a background of vocal folds vibration, brain waves, and machine learning models. Chapter 3 highlights and analyzes vocal folds vibrations for Japanese vowels classification using Support Vector Machine. Chapter 4 highlights and analyzes pre-speech EEG data for prediction of words using Echo State Network. Chapter 5 concludes by summarizing the key findings of this study and discuss future work.

Chapter 2

Background

2.1 Vocal Folds Vibration

2.1.1 What Is Vocal Folds Vibration

Vocal folds vibration is when the vocal folds produce a sound by vibrating during the exhalation of air from the lungs. Figure 2.1 shows an illustration of the larynx. The vocal folds are located within the larynx (voice box) at the top of the trachea. The vibration of vocal folds is measured as their frequency (Hz). For example, when vocal folds vibrate 100 cycles per second, the frequency is 100 Hz. In addition, the average fundamental frequency for a male voice is 125Hz, and for a female voice it is 200Hz.

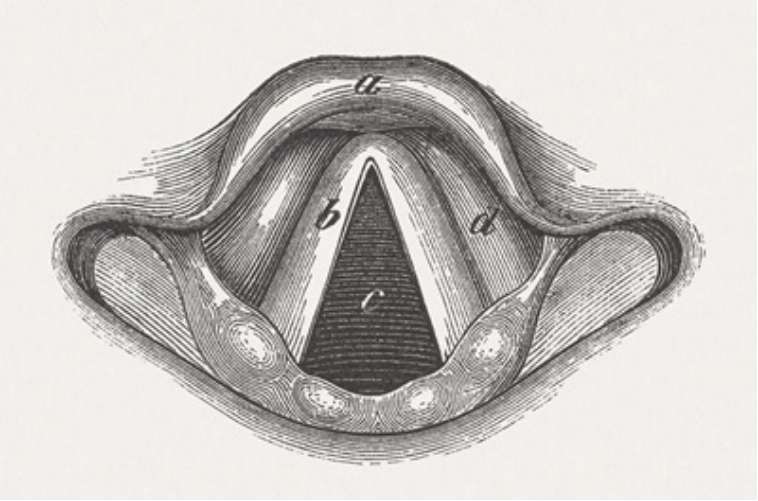

Figure 2.1: Illustration of the larynx: The part of b is vocal folds, a is Epiglottis, c is tracheae, d is ventricular fold [9].

2.1.2 Relations Between Vocal Folds Vibration and Speech

Sounds can be divided into two categories: Voiced and unvoiced. Voiced sounds are produced when the vocal folds vibrate during the vocalization of a phoneme. By contrast, unvoiced sounds do not require the use of the vocal folds. Voiced sounds are produced by vocal folds as described in the previous section. Secondly, the voiced sounds are amplified and altered by the vocal tract resonators (vocal tract, mouth cavity, and nasal passages) to produce a person's recognizable voice. Thirdly, the voiced sounds are modified by the vocal tract articulators (the tongue, soft palate, and lips) to produce recognizable words. For example, as shown in Figure 2.2, each form of resonators and articulators (vocal tract, mouth cavity, and nasal passages) are unique corresponding with the sounds of /a/ and /i/.

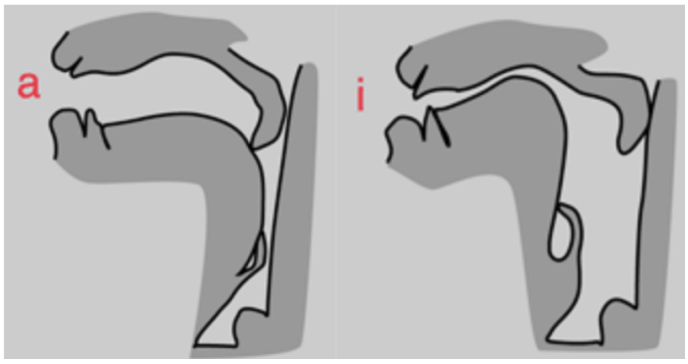

Figure 2.2: Articulatory forms when human speaks /a/ and /i/ sounds: The left is when speaking /a/, the right is when speaking /i/ [9].

2.1.3 Speech Interface Study Using Vocal Folds Vibration

As we highlighted above, speech and vocal folds vibration have a strong relationship and can be considered as biological information containing key features of speech.

There exists a number of techniques to capture the speech features embedded in vocal folds vibration. One of the methods is Electroglottograph (EGG), which measures the degree of contact between the vocal folds by measuring the electrical impedance from two electrodes plates placed on the neck at the level of the larynx

(Figure 2.3) [10] . As shown in Figure 2.4, when the electrical impedance is low, the degree of contact between the vocal folds is small (open) while when the impedance is high, the degree of contact is large (closed). In general, EGG is mainly used for diagnosis of larynx disorders.

Another method is electromyography (EMG). It measures the electrical potential with wire electrodes inserted in the different muscles. [11]. This technique is commonly used to evaluate the health condition of muscles and the nerve cells that control them.

At the same time, it is important to note that it is difficult for these methods to obtain the speech features that can be used for speech interface, which discriminate speech.

Figure 2.3: EGG device placed on throat

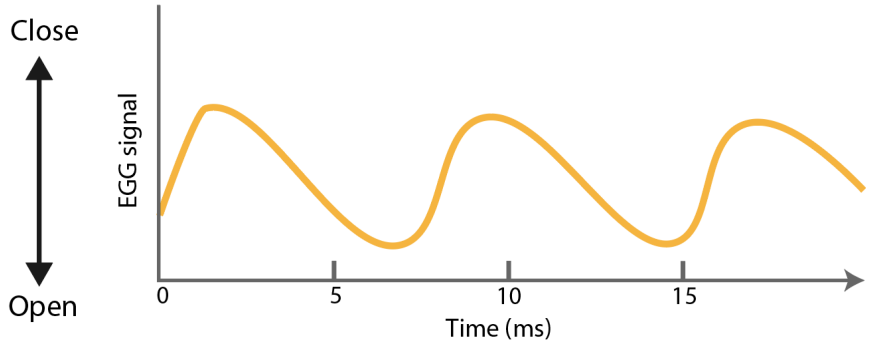

Figure 2.4: The EGG wave of impedance measured by EGG device. The x axis represents time, and y axis represents impedance)

2.2 Electroencephalography

2.2.1 What Is Electroencephalography

Electroencephalography (EEG) is a method to record the electrical activity of the brain using electrodes placed on the scalp. The electrical activity is produced by neurons communicating in the brain's cortex as shown in Figure 2.5. When neutrons communicate their information to each other, they generate an electrical potential (action potential and postsynaptic potential). The electrical potential fluctuations are measured by the EEG electrodes.

As described above, EEG signals represent patterns of brain activities. For example, EEG signals in the frequency of 8-14 Hz (called alpha wave) are detected when someone is relaxed; beta waves (14-30 Hz) appear when someone is actively thinking; gamma wave (30-50Hz) appears during meditation [13].

2.2.2 Relations Between EEG and Speech

The brain controls all the body's functions including speech [8]. In the brain, several areas are corresponding with each function. As shown in Figure 2.6, Broca's area, located in the left hemisphere, is associated with speech production and articulation.

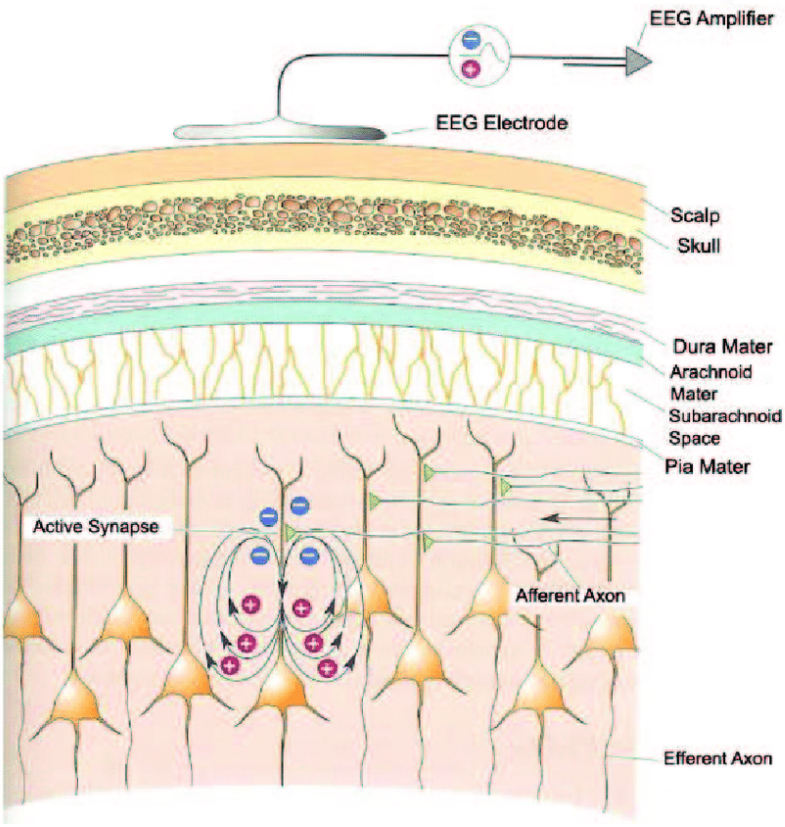

Figure 2.5: Measuring electrical activity of synapses in the cortex [12].

The speech information from the Broca's area is transmitted to the motor cortex located in the front lobe and the motor cortex tells the muscles of the mouth, tongue, lips, and throat how to move to form speech [14].

2.2.3 Study of Speech and EEG

Some studies verified that EEG signals have features associated to speech. For example, Ghane et al. [16] collected EEG signals while participants imagined English vowels without moving the oral cavity or jaws. With the EEG signals, they examined whether the vowels could be recognized or not, and the overall accuracy was 76.6 %. For another study, Moses et al. [17] recorded cortical activity while the participant attempted to say individual words from a vocabulary set. As a result, using the cortical activity data, they classified words with 47.1 % accuracy.

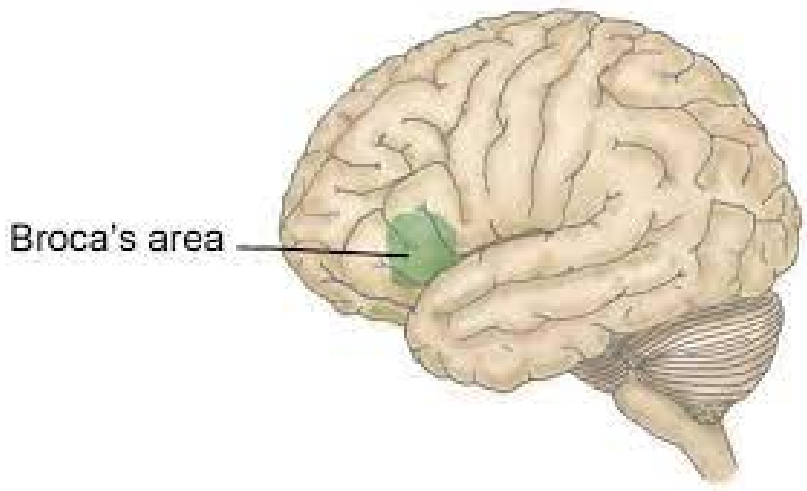

Figure 2.6: Broca's area in the brain [15].

From these studies, it was verified that EEG signals have features associated with speech, and could contribute to a new speech interface development by further improvements of the accuracy and the variety of words' recognition.

2.3 Machine Learning Models

Machine learning models have been used as tools for speech analysis and classification. In this study, we plan to utilize machine learning models to process vocal folds vibration and EGG signals [1]. We can consider two possible classifications for time series data. One is to utilize a machine learning model that uses features that ignore the temporal context of the data (e.g., the average Intensity of the data), and the other is to use machine learning that preserves the temporal context of the data. In this chapter, we highlight Support Vector Machines (SVMs) as the representation of the feature-based model, and describe Recurrent Neural Network, and Echo State Network as the time-maintained model.

2.3.1 Support Vector Machine

Support Vector Machines (SVMs) [18] is one of the algorithms often used for training a classifier which, given a training dataset with a set of features, can predict the class for previously unseen samples. Figure 2.7 highlights how SVM works for a binary classifier. When we have a two-class dataset, an SVM will find a hyperplane that will divide the space of features between the two classes. This hyperplane is chosen so that it maximizes the margin between itself and the closest data. Thus, all the points at one side of the hyperplane will belong to one class, and all the points at the other side will belong to the other class. This is how SVM classifies.

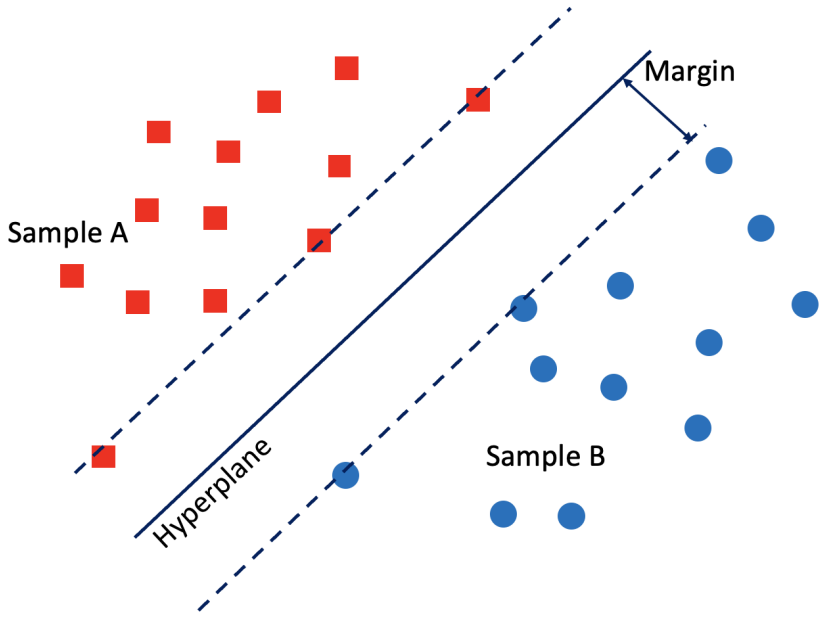

Figure 2.7: SVM classification of two class dataset.

Classification through a hyperplane is also possible for non-linearly separable data through a method called the kernel trick.. The kernel trick enables SVM to deal with non-linear separating hyperplanes since the input data are transposed to a higher-dimensional space (called the feature space) where a separating hyperplane is found. Figure 2.8 illustrates this principle.

Although SVM is an effective classifier in a wide array of classification tasks, it is

unable to retain the ordering in time of the sequence of inputs present at an input data. For this purpose, Recurrent Neural Networks (RNN) is commonly used for time series and sequential data classification.

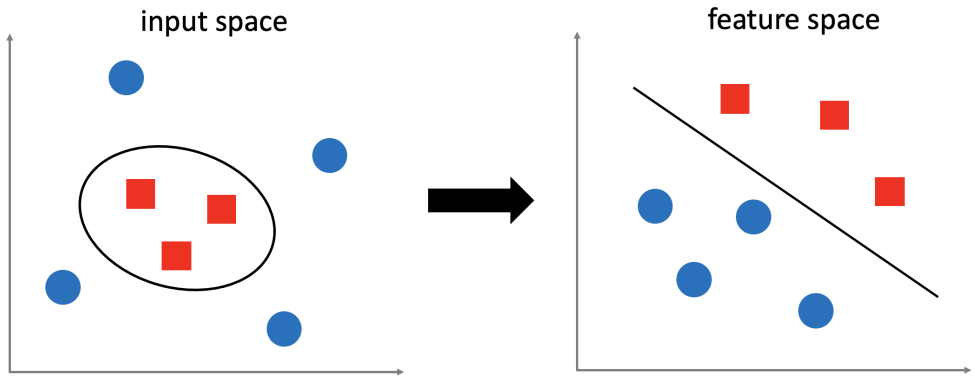

Figure 2.8: The kernel method finds a hyperplane in a higher dimensional space (feature space) of input data for classification.

2.3.2 Recurrent Neural Network

For time series pattern recognition, machine learning models, which can approximate a dynamic system, are necessary. One of the models is Recurrent Neural Network (RNN). In this section, we highlight the concept of RNN and its features.

Figure 2.9 illustrates a general recurrent neural network model consisting of the input layer, recurrent layer, and output layer. N_u, N_x, and N_y represent the number of nodes in the input layer, the recurrent layer, and the output layer, respectively. Due to handling time series data, each node state changes over a discrete time n ($n = 0, 1, 2, ...$). Input vector $u(n)$, the state vector of a node in the recurrent layer $x(n)$, and the output vector $y(n)$ are represented as:

$$u(n) = (u_1(n), ..., u_{N_u}(n))^T \in \mathbb{R}^{N_u} \tag{2.1}$$

$$x(n) = (x_1(n), ..., x_{N_x}(n))^T \in \mathbb{R}^{N_x} \tag{2.2}$$

$$y(n) = (y_1(n), ..., y_{N_y}(n))^T \in \mathbb{R}^{N_y} \tag{2.3}$$

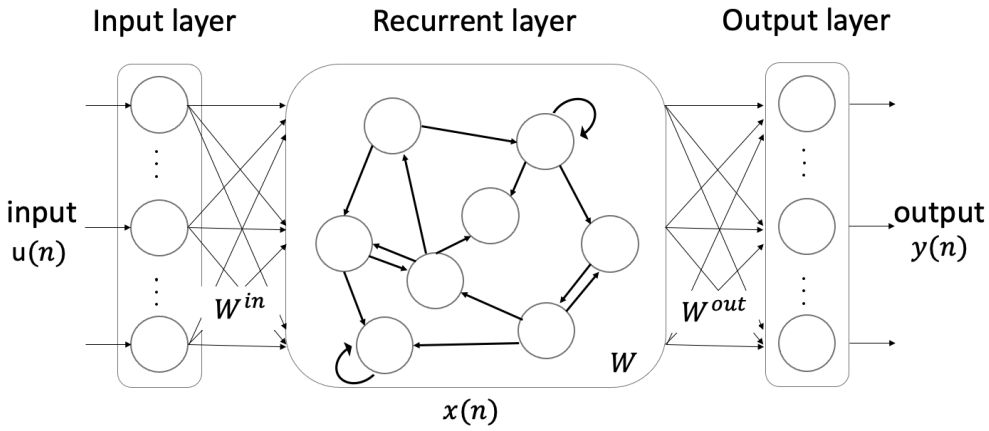

Figure 2.9: General Recurrent Neural Network Model

The weight connectivity between the input layer and recurrent layer is:

$$W^{in} = (w_{ij}^{in}) \in \mathbb{R}^{N_x \times N_u} \tag{2.4}$$

The weight connectivity of the inside recurrent layer is:

$$W = (w_{ij}) \in \mathbb{R}^{N_x \times N_x} \tag{2.5}$$

The weight connectivity between the recurrent layer and output layer is:

$$W^{out} = (w_{ij}^{out}) \in \mathbb{R}^{N_y \times N_x} \tag{2.6}$$

At time $n + 1$, the node state vector in the recurrent layer is:

$$x(n+1) = f(W^{in}u(n+1) + Wx(n)) \qquad (n = 0, 1, 2, ...) \tag{2.7}$$

Here, $f()$ represents the activation function for the artificial neuron nodes. Using the node state vector in the recurrent layer, the output vector is:

$$y(n+1) = f(W^{out}x(n+1)) \tag{2.8}$$

The standard algorithms used for training the RNN model is Backpropagation Through Time (BPTT) [19]. This algorithm updates the parameters of weight connectivity sequentially based on the gradient descent method. However, this algorithm may cause technical problems such as vanishing gradient or exploding gradient for long length input sequences. To solve this problem, RNN models have been designed such as LSTM [20], and GRU [21]. Since those methods are a combination of existing methods like BPTT, they still need to perform a very large number of calculations when applied on a model of large size. Therefore, there is a special model of RNN, called Echo State Network (ESN), which is an approach that drastically reduces the computational complexity for training with a different training algorithm. For our second study, this ESN was used for speech classification.

2.3.3 Echo State Network

Echo State Network (ESN) model belongs to the class of reservoir computing models discussed by [22]. In ESN model, an RNN model with fixed weights chosen randomly (reservoir) is used to generate a state in which the past information of the time series input remains echoed (echo state), and features of the input are read out (readout). To adjust the readout, a linear learner with low computational complexity is used as an output layer. The goal of the ESN model is to achieve both high computational performance and fast learning.

The basic model structure of ESN is shown in Figure 2.10. This is the same structure as that of a general recurrent neural network, except that the weight connectivity of the recurrent layer is fixed and the recurrent layer is used as a reservoir. For the learning algorithm, the output weight matrix W_{out} is only adjusted. The reservoir is a transformer of the input data, and the readout is a learner to properly read out the

input features from the state of the reservoir.

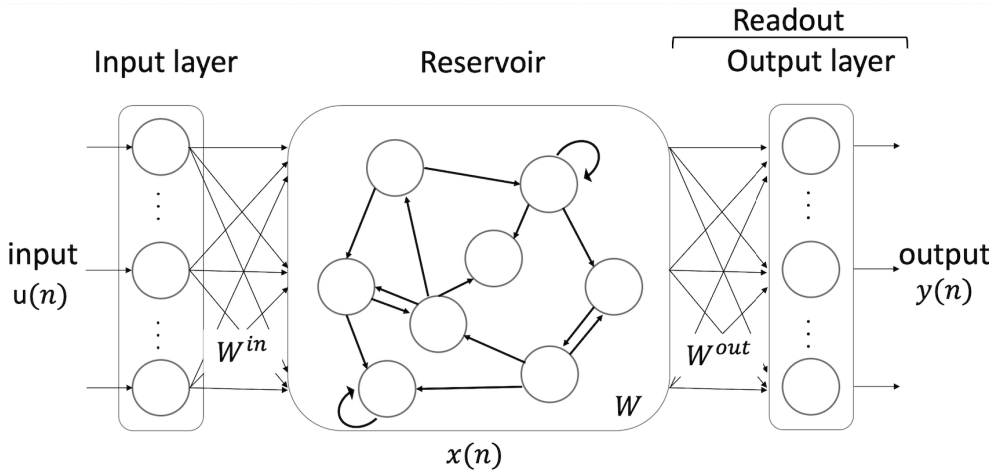

Figure 2.10: The general Echo State Network model: The same structure as that of a general recurrent neural network, except that the weight connectivity of the recurrent layer is fixed and the recurrent layer is used as a reservoir.

ESN performance depends on specific hyper-parameters. The sparsity of the reservoir weight matrix connectivity allows for signals to more easily propagate and terminate inside the reservoir. The sparsity is determined by setting the percentage of connectivity. The spectral radius of the reservoir is used to determine how much the previous time series information should be retained, and satisfy with the property of the reservoir, called Echo State Property (ESP) which is a condition of asymptotic state convergence of the reservoir network under the impact of controlling inputs [23]. The spectral radius $\rho(W)$ represents the maximum absolute value of the eigenvalues of W, as in the following Equation.

$$\rho(W) = \underset{i}{max}(|\lambda_i|) \tag{2.9}$$

The leaky rate α defines the speed of reservoir updating dynamics, described by Wu et al. [24]. With the leaky rate, the output of the reservoir state vector is written as:

$$y(n+1) = (1-\alpha)x(n) + \alpha f(W^{out}x(n+1)) \tag{2.10}$$

Where $\alpha \in (0, 1]$ is leaky rate.

Chapter 3

Japanese Vowel Discrimination From Vocal Folds Vibration

3.1 Introduction

Chapter 1 described the the need to find new biological information for speech interface developments, and Section 2.1 highlighted that the vocal folds vibration and speech have a relationship so that the vocal folds vibration can be a useful biological information because it includes speech features. However, we also noted that the existing methods for the measurement of the vocal folds vibration (EGG, MEG, and so on) cannot be used enough for the development of speech interfaces because they cannot measure features that can discriminate speech.

Therefore, in this part of study, we focused on using an acceleration sensor to measure the vibration of the vocal folds. By attaching the acceleration sensor to the throat, data can be measured without the need for invasive medical procedures. In addition, we verified whether the acceleration data can be used to discriminate specific speech sounds with the application of a Support Vector Machine model. This chapter presents this research in the following order: The measurements, the feature extraction, the classification method, and the result.

3.2 Measurements

3.2.1 Overview

To measure the vocal folds vibration data with an acceleration sensor, the sensor was attached to the throat at the level of Adam's apple. The acceleration data were recorded while the subject was vocalizing specific sounds (Japanese vowels in this study). The following sections explain the devices we used, data collection procedure, and the measured data.

3.2.2 Devices for Experiment

To measure the acceleration data, we used the TSND121 sensor (Figure 3.1), which is manufactured by the ATRAdvanced Telecommunications Research Institute International. This device consists of a small wireless multi-functional sensor, including an acceleration measurement. We can select the sampling frequency and the acceleration range. The acceleration sensor obtains the three dimensional acceleration data (x, y, z) but we only used the z-axis data. For sampling the vibration data clearly, the higher sampling rate should be set. Therefore, the sampling frequency was set 1000 to Hz (maximum). For the acceleration range, because the vibration of the vocal folds is quite small, the range was set $\pm$2G (minimum).

Figure 3.2 illustrates the flow diagram for the data measurement. For collecting the acceleration data from the sensor, we used the dedicated software, ALTIMA. As shown in Figure 3.3, the collected data sent to ALTIMA is displayed in the software and saved to an excel file.

Figure 3.1: The small wireless multi-functional sensor (TSND121), which can measure the acceleration data and transmit it through a Bluetooth wireless link.

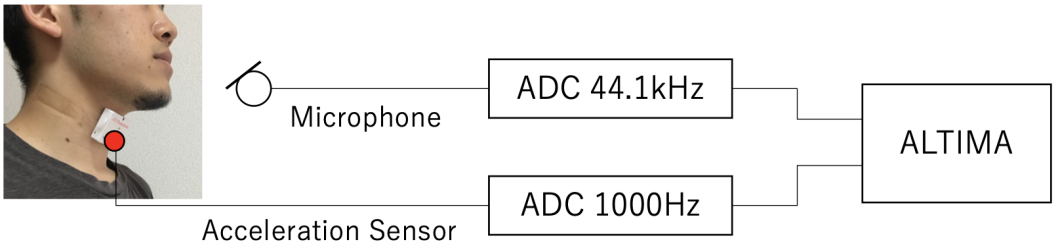

Figure 3.2: Flow diagram of the data measurement. The acceleration data, measured by the sensor, is converted into digital signal and is sent to ALTIMA via a Bluetooth wireless link. The voice data, measured by the microphone, is converted into digital signal and is sent to ALTIMA via a wired audio interface. The software, ALTIMA, saves both data to an MS Excel file.

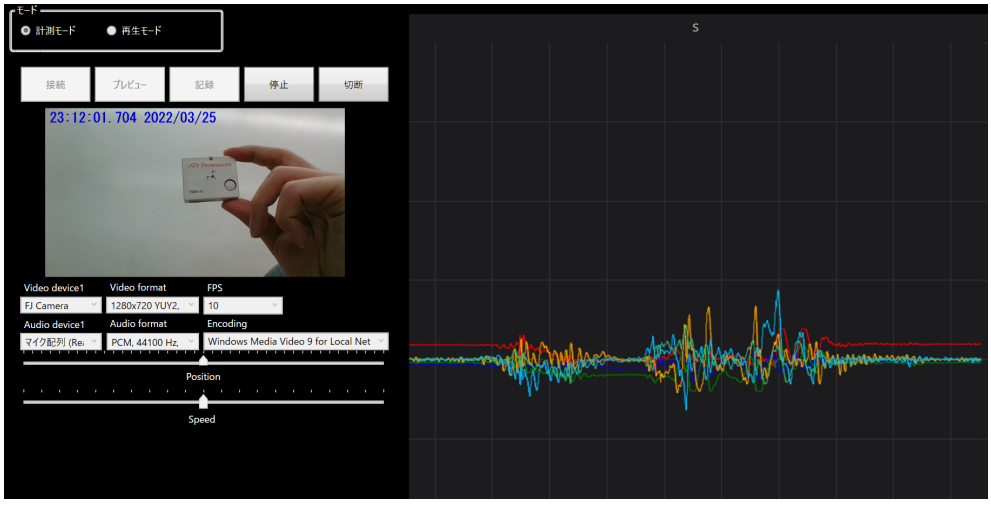

Figure 3.3: The software ALTIMA: The collected data are sent to this software and is displayed as shown on the right side. On the left side, the video is displayed although we did not use it in this study.

On account of the fact that we measure the acceleration data when the subject vocalizes, the voice data ought to be collected to detect the speech onset. For recording the voice data, we used SONY Dynamic Microphone F-720 at a sampling frequency of 44.1 kHz. This recorded voice data were converted into digital signal and was sent to ALTIMA.

3.2.3 Data Collection Procedure

To measure the acceleration data from the vocal folds vibration, the sensor was attached to the throat at the level of Adam's apple where the larynx is located as shown in Figure 3.4.

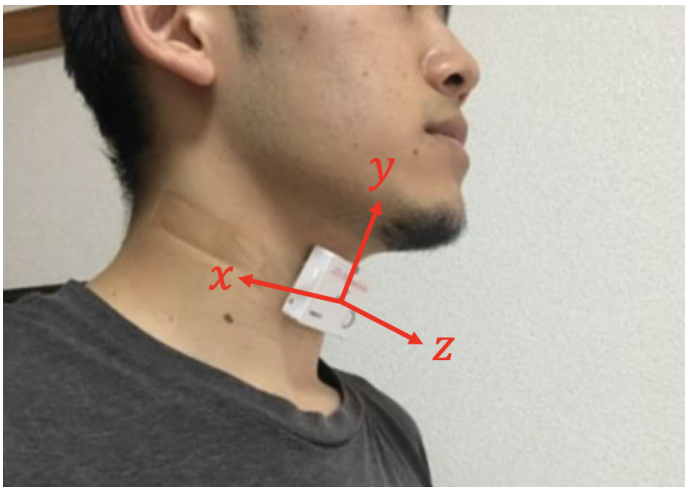

Figure 3.4: The acceleration sensor was attached to the throat at the level of Adam's apple where the larynx is located. We only used the z-axis acceleration data for measuring the vibration in this study.

The phonemes that the subject was asked to vocalized were Japanese vowels (/a/, /i/, /u/, /e/, and /o/). The reason why we used only (Japanese) vowels in this study is that there have not been any studies verifying the possibility of even vowel discrimination. In addition to this, since the vowels are the fundamental speech sounds, we decided to examine the vowel discrimination with the acceleration data of the vocal folds vibration.

The procedure of the measurement is shown in Figure 3.5. With the acceleration sensor attached to the throat, the subject was asked to vocalize each vowel in order. To ensure that the pitch of each vowel sound is the same, the subject listened to a tone (100 Hz tone) every time before the subject vocalizes. This process was repeated 10 times. For this experiment, a 22-year-old man participated in this study.

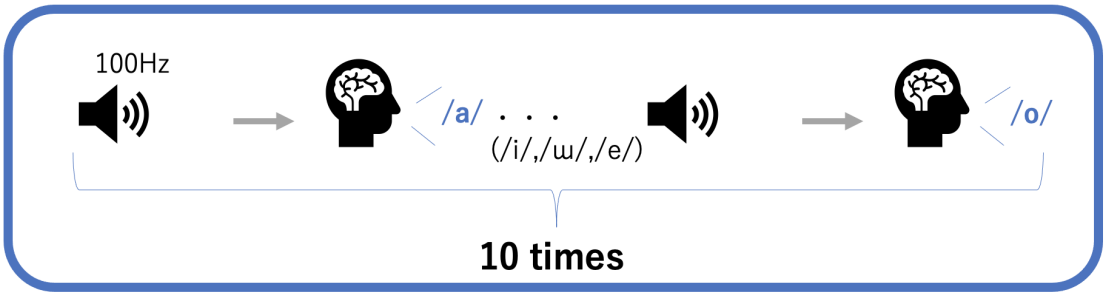

Figure 3.5: The procedure of the acceleration measurement. The participant was asked to vocalize each vowel sound in order. To ensure that the pitch of each vowel is the same, the subject listened to the 100 Hz tone every time before the subject vocalizes. This process was repeated 10 times.

3.2.4 Collected Data

Figure 3.6 shows the collected acceleration data when the participant vocalized each vowel once. The x-axis represents the time, and the y-axis represents the acceleration data. Because the five vowels were vocalized once in one block, we can see five large amplitude signals in the plot. From the left side, the phonemes follow /a/, /i/, /u/, /e/, and /o/.

To verify whether each signal has its unique wave form, we zoomed in on the part of the vocal folds vibration with the largest amplitude during each vowel vocalization to a range of 10 ms since one cycle is 10 ms. As seen in Figure 3.7 - 3.11, we could see the vibration form of the vocal folds in one cycle by taking a closer look at an interval of 10 ms though the sampling rate might not be sufficient rate to clearly observe useful details. It can be seen that the degree of kurtosis and concavity differs slightly for each vowel. However, we cannot still see significant differences among the

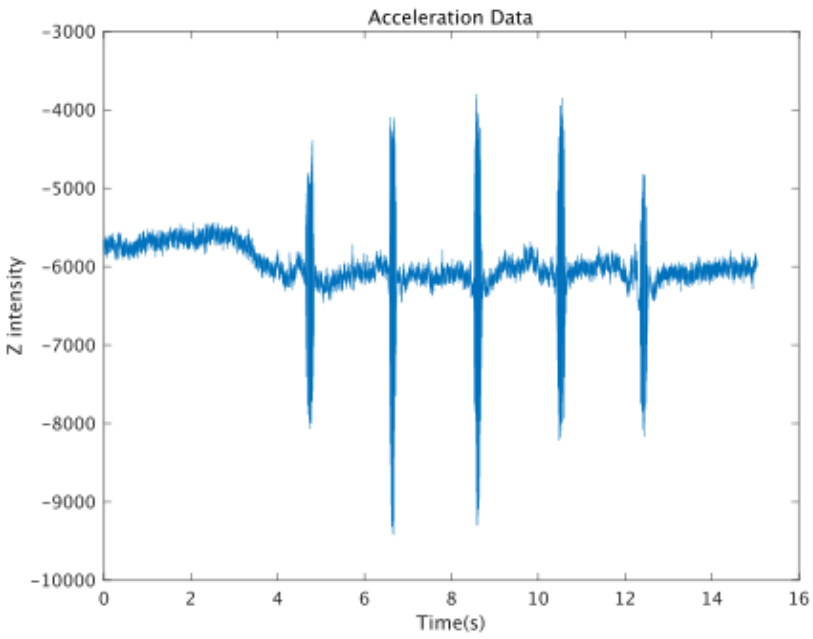

Figure 3.6: The plot of the acceleration data, collected when the participant vocalized each vowel once. The x-axis represents the time and the y-axis represents the acceleration data. From the left side, the order is /a/, /i/, /u/, /e/, and /o/.

signals, so that it may be challenging to discriminate each speech from the raw data. Therefore, we applied a feature extraction step before feeding the data to the machine learning model for speech classification.

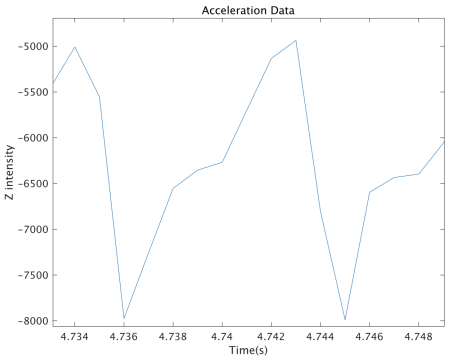

Figure 3.7: Acceleration data within 10 ms when the subject vocalized Japanese vowel /a/.

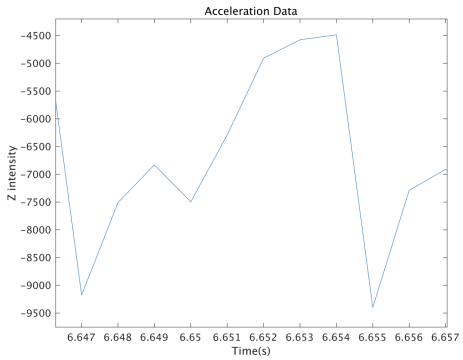

Figure 3.8: Acceleration data within 10 ms when the subject vocalized Japanese vowel /i/.

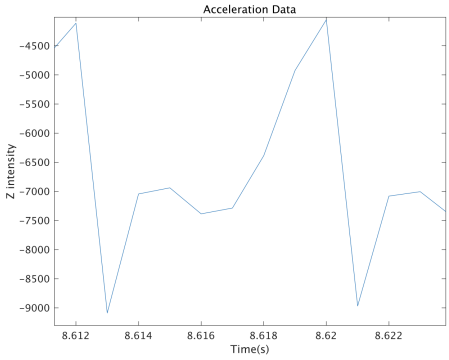

Figure 3.9: Acceleration data within 10 ms when the subject vocalized Japanese vowel /u/.

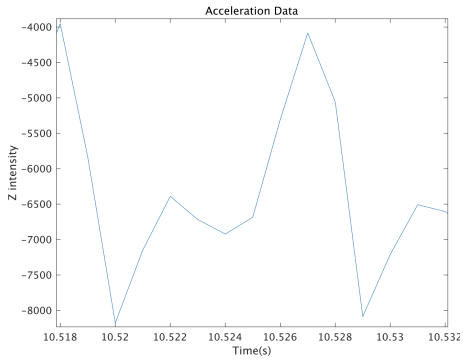

Figure 3.10: Acceleration data within 10 ms when the subject vocalized Japanese vowel /e/.

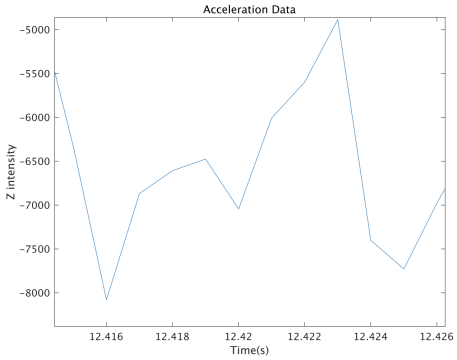

Figure 3.11: Acceleration data within 10 ms when the subject vocalized Japanese vowel /o/.

3.3 Feature Extraction

3.3.1 Overview

A common procedure for classification implementations follows *data capture*, *feature extraction*, and *classification* as shown in Figure 3.12. In the previous section, we described the data capture with the accelerometer. For the next, the feature extraction method we carried out is addressed in this section, and the classification method is explained afterwards.

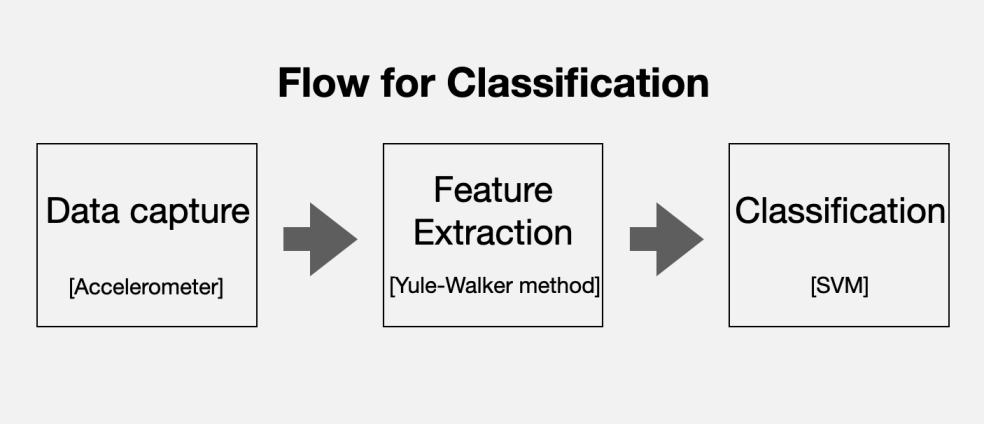

Figure 3.12: Flow for classification implementation

As described in Section 2.3.1, SVM needs a set of features to find an appropriate hyperplane that can discriminate classes. A common approach to extract features with time series data is to convert it to the frequency domain [1, 25, 26]. There exists lots of methods for the conversion [27, 28, 29, 30]. In this study, we used one of the methods called Yule-Walker method. This method generates the frequency features from time series signals and also was used for biological signal analysis in [31, 32]. Thus, we applied this technique to the acceleration data we collected.

In this section, we highlight the techniques we used for feature extraction, and how to apply it. At the end, we examine which features we decided to utilize for classification.

3.3.2 Signal Processing

There exists many techniques that estimate information in the frequency domain of a signal from the time series data, such as zero-crossing method mentioned in [27], auto-correlation method [28], and Cepstrum method [29, 30]. In this study, the Yule-Walker method was used to estimate the power spectral density (PSD) in view of the fact this method can deal with both stationary, and non-stationary [33], and is used for time series feature extraction. Kallas et al. [31] examined the prediction of signal changes in electrocardiograms (ECG) and used the Yule-Walker method to estimate the feature values. More details of the Yule-Walker method are discussed in [32].

The Yule-Walker method assumes an auto-regressive model. Therefore, using Yule-Walker method, we have to set the order of the AR (auto-regressive) model used to generate the PSD estimate as well as the input data and the number of samples to use in the Discrete Fourier Transform (DFT).

Before performing PSD estimation, we preprocessed the acquired data. First, 0.3-second (300 samples) intervals of the acceleration data during each vowel vocalization were cut from the measured data. The Hamming window was applied to each of the cut signals for the same number of samples. After the preprocessing, we performed the spectral analysis on the data to estimate its frequency components. To implement the Yule-Walker method, we used the function given in MATLAB:

```
[pxx,f] = pyulear(x,order,nfft, Fs)
```

This function requires the input signal, `x`, the arguments of the order of the AR (auto-regressive) model used to generate the PSD estimate, `order`, the number of samples to use in the Discrete Fourier Transform, `nfft`, and the sampling rate, `Fs`. We set the order as 10, and the number of samples as 2048. This procedure has been applied for all the acceleration data.

3.3.3 Extracted Data

As a result of the signal processing procedure just explained, we got the PSD output `pxx` with 1025 samples. Figure 3.13 shows the plot of the output (vowel /a/).

To choose from the frequency-domain information a useful feature for our intended classification purpose, we selected the frequency values of the first two peaks appearing on the plot (first and second harmonics with largest power content). These frequency values were extracted from all data.

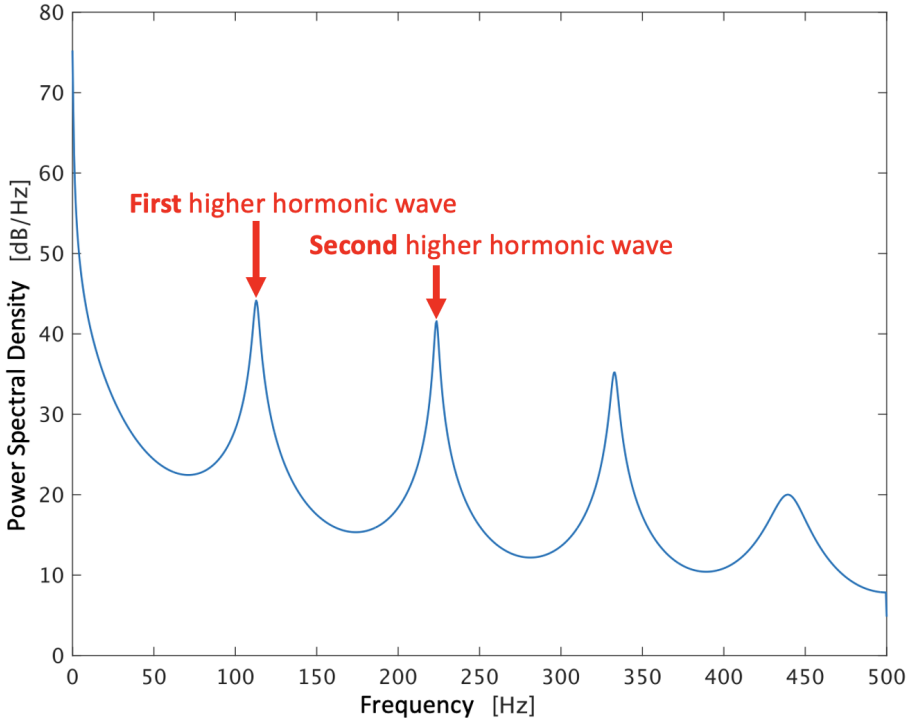

Figure 3.13: Power spectral density, obtained from the acceleration data of the vowel /a/. The x-axis represents the frequency and the y-axis represents the power spectral density. The first two peaks are chosen as the feature values (at about 115 Hz, and 220 Hz for this plot).

3.4 Classification Using Machine Learning

3.4.1 Overview

Thus far, the acceleration data were collected by the sensor placed on the throat and an spectral analysis was applied to extract the feature values. The feature values are the first and second harmonics with larger power content. In this section, using the feature values, we verified whether the feature values are useful in providing information to classify speech sounds using an SVM model.

3.4.2 Methodology

The purpose of this study was to verify if the acceleration data can discriminate each vowel, nevertheless there have not been any studies of vowel discrimination by vocal folds vibration. Therefore, we arranged the pattern of discrimination for the two-class classification between the vowel /a/ and the other vowels (i.e. /a/ vs. /i/, /a/ vs. /u/, /a/ vs. /e/, /a/ vs. /o/) to start with as simple classification as possible. The total number of samples we collected was 50 samples (10 samples per vowel).

Figures 3.14 – 3.17 show the plots of each feature values' pair to examine whether the feature values are appropriate for classifying classes visually. The x-axis (F1) and y-axis (F2) represent the first and second larger-power harmonic waves (feature values), respectively.

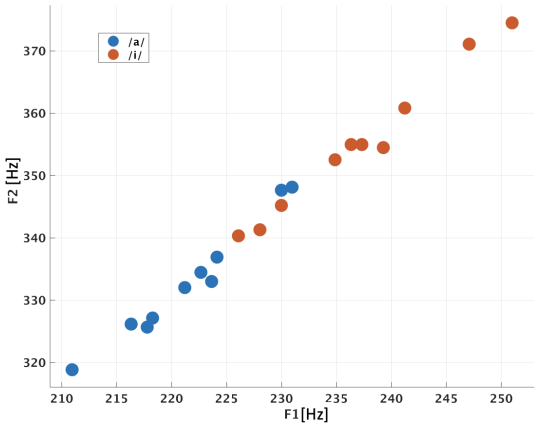

Figure 3.14: Feature map with 2 class dataset: /a/ and /i/. The blue dots represent the feature of /a/, and the orange dots represent the feature of /i/.

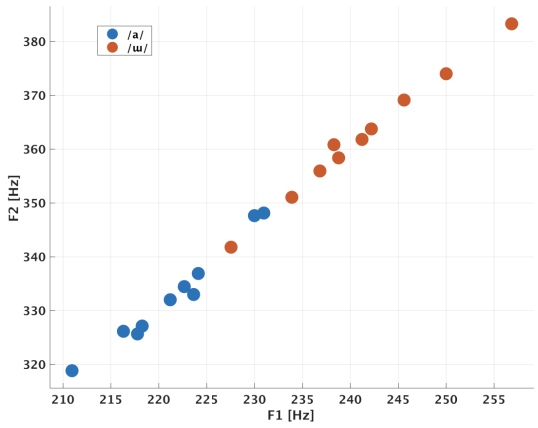

Figure 3.15: Feature map with 2 class dataset: /a/ and /u/. The blue dots represent the feature of /a/, and the orange dots represent the feature of /u/.

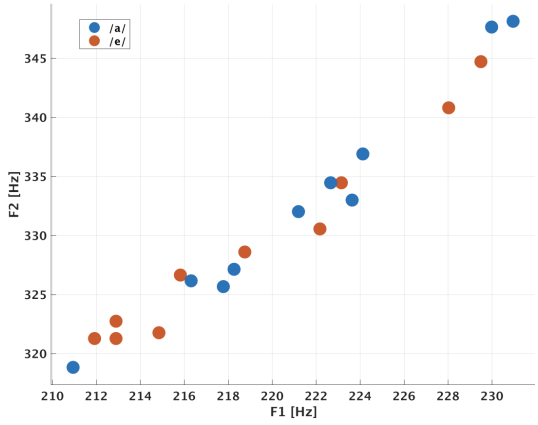

Figure 3.16: Feature map with 2 class dataset: /a/ and /e/. The blue dots represent the feature of /a/, and the orange dots represent the feature of /e/.

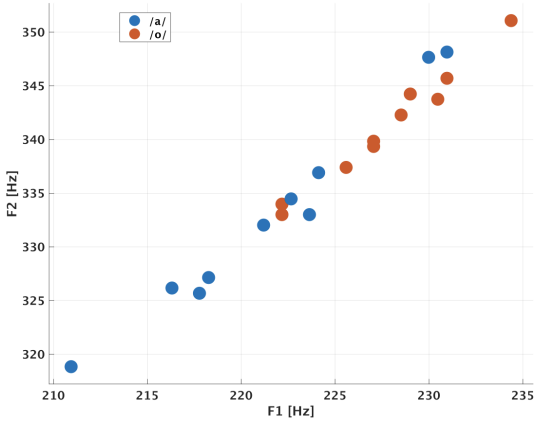

Figure 3.17: Feature map with 2 class dataset: /a/ and /o/. The blue dots represent the feature of /a/, and the orange dots represent the feature of /o/.

Using the classifier, SVM, the discrimination accuracy was tested for each two-class dataset. We implemented SVM by using the Classification Learner app in MATLAB. We trained the SVM model using cross-validation and labeled dataset. Labels were assigned in the dataset in the following way: `1:/a/, 2:/i/, 3:/u/, 4:/e/, 5:/o/`. Figure 3.18 shows the table that contains feature values and corresponding labels. For cross validation, because we did not have many sample data, we set it with 5 folds; The given dataset is split into five partitions of ten samples each, where each fold is used as a testing set at some point. Moreover, as shown in Figures 3.14 – 3.17, dots

of each class are mostly grouped together; Therefore, we used a linear SVM for the classification.

	F1	F2	Label
	Number	Number	Number
1	F1	F2	Label
2	230.9570	348.1445	1
3	250.9766	374.5117	2
4	250	374.0234	3
5	214.8438	321.7773	4
6	228.5156	342.2852	5
7	217.7734	325.6836	1
8	229.9805	345.2148	2
9	233.8867	351.0742	3
10	212.8906	321.2891	4
11	225.5859	337.4023	5
12	210.9375	318.8477	1
13	226.0742	340.3320	2
14	227.5391	341.7969	3
15	211.9141	321.2891	4
16	227.0508	339.3555	5
17	218.2617	327.1484	1

Figure 3.18: Classification Learner app setting: The first and second columns show the feature values, the first and second larger-power harmonic components, respectively. The third column represents the labels assigned to each sample in the dataset. Each class in the dataset, a vowel, was designated with a number 1 to 5 as follows: 1:/a/, 2:/i/, 3:/u/, 4:/e/, 5:/o/.

3.5 Result

Here, we implemented classifications between `/a/` and other vowels (`/i/, /u/, /e/, /o/`); We have not examined patterns such `/i/ vs. /o/ or /u/ vs. /e/`. For each two-class classification, the discrimination accuracy calculated by the classifier (SVM) is shown in Table 3.1. As a result, the average discrimination accuracy was recorded as 71 %.

The accuracy of `/a/ vs. /i/`, `/a/ vs. /u/`, and `/a/ vs. /o/` were fairly high, while that of `/a/ vs. /e/` was low. There are two possible reasons of the low classi-

Table 3.1: Accuracy of Japanese vowel discrimination

Pattern	Accuracy [%]
/a/ vs. /i/	75
/a/ vs. /u/	85
/a/ vs. /e/	55
/a/ vs. /o/	70

fication accuracy between `/a/` and. `/e/`.

First, the similarity of the frequencies associated with both input features might cause the low accuracy. Rue et al. [34] investigated the frequency differences of vowels and Figure 3.19 shows the frequency distribution. As we can see in the figure, the position of vowels /a/ and /e/ are close compared to others; their frequencies are similar. In addition to this, observing the feature map between /a/ and /e/ (Figure 3.16), we can tell most of feature dots of both classes are overlapped and not separated well as others were. Hence, although we used the frequency data as the feature values in this study, the frequency data may not be the best choice for the classification of this pair. To improving the accuracy, we need to find other feature values besides frequency that can can help to better discriminate /a/ and /e/. For example, another feature extraction method we can apply is Mel Frequency Cepstrum Coefficients (MFCC). MFCC is a method that was used to extract voice features values which have been widely used in the field of speech analysis, for speech recognition [35, 36, 37]. To apply the MFCC method, it requires Mel-cepstrum which is also one of the most preferred speech recognition features. The basic idea of the Mel-cepstrum reflects the human auditory perception mechanism for speech recognition.

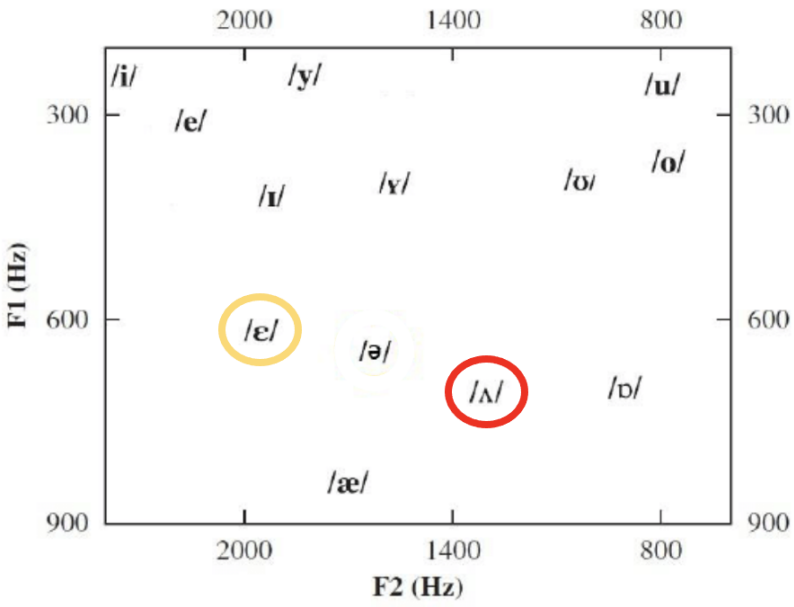

Figure 3.19: Frequency distribution of vowels. The y-axis (F1) represents the first formant frequency, and the x-axis (F2) represents the second formant frequency. The part circled in red is the Japanese vowel /a/, the part circled in yellow is the Japanese vowel /e/. The position of these two is close compared to others. [34]

Secondly, the number of samples we used for the classification might not be large enough. Since we only had 50 samples in total, we chose to use the machine learning model, SVM, that requires a few samples to train itself. But as shown in Figure 3.16 (the feature map between /a/ and /e/), the two classes are not sufficiently separated. Therefore, with more samples the SVM may be able to find an appropriate hyperplane that can separate the two classes better.

3.6 Conclusion

In this part of the thesis, the vocal folds vibration data of Japanese vowels were measured from the throat using an acceleration sensor, and the first and second harmonics with larger power content were extracted and used for the classification. As a result, we saw that following this procedure we were able to discriminate vowels using a biological signal (the vibration of the vocal folds).

In addition, as described in Section 2.1.3 , there is the method for the vocal folds vibration analysis called EGG to measure the degree of contact between the vocal folds by measuring the electrical impedance. In [38] and [39], it was described that it is possible to identify only whether the person speaks modal voice or falsetto voice by seeing the difference of the wave forms recorded by EGG. However, no further identification of speech has been done using EGG or any other method of measuring vocal fold vibration data. Through our study, the result shows the possibility that the vocal fold vibration data measured by the acceleration sensor provides feature data to discriminate speech. Though we only classified simple patterns such as /a/ and other vowels, we aim to address more variety of classification patterns as a future work.

The main purpose of this study is to find a new biological signal to develop new speech interfaces. As a result, we identified a new biological signals, the vibration data of the vocal folds measured by the acceleration sensor, that could classify Japanese vowels. Nevertheless, as most words are composed of a combination of vowels and consonants, it is necessary to discriminate consonants as well as vowels. However, it can be difficult to recognize consonants only by the vocal folds vibration data since the vocal folds vibration generates the fundamental frequency (vowels). Therefore, we focused on using Electroencephalography (EEG) because it is associated with the speech production process and has the potential to recognize consonants. This part of the study is further elaborated in the following chapter.

Chapter 4

Unvoiced Consonant Prediction from Pre-Speech EEG Data

4.1 Introduction

Our research goal was to find new biological signals that can be used for speech interface development. In the previous chapter, we focused on the vocal folds vibration, and verified the discrimination accuracy of Japanese vowels using the vibration signals from the vocal folds measured by the acceleration sensor. As a result, the accuracy of discriminating /a/ versus the rest of four vowels was 71 % on average. Thus, we found the possibility that the vocal folds vibration data can be used as a new biological signal for the development of speech interface. However, because most words are composed of vowels and consonants, it is necessary to discriminate consonants as well as vowels.

Therefore, for the next study we focused on the electroencephalogram (EEG) data as biological signals to recognize the consonants. This is because it is known that EEG signals, because of being originated in the brain during the process of speaking, contain the speech features as described in Section 2.2.2. Also, there have already been some studies that attempted to recognize what a human test subject wanted to say using EEG data [16, 17]. However, most studies have not presented high accuracy yet, and none of these studies carried out specifically the classification of consonants. For the study in this chapter, we measured EEG data while the

participants were vocalizing the words, and extracted the pre-speech EEG data (EEG data right before the vocalization). Using the pre-speech EEG data, which can be considered to have less noise, the classification of consonants has been conducted. To perform the classification of the EEG signals into certain target consonants, an Echo State Network (ESN) model, which is a subset of recurrent neural networks (RNN) and can deal with time series data well was utilized.

This chapter is organized in the following way: Firstly, the devices and software used for data measurement and the content of the experiment are introduced. Secondly, a preprocessing method to the measured data is explained, followed by the explanation for ESN model and the setting of the parameters. Thirdly, the result and discussion about the consonant classification are examined. Lastly, the conclusion is highlighted.

4.2 Measurement

4.2.1 Overview

In this study, the measured data was (pre-speech) EEG data. The participants were asked to wear a EEG head cap and vocalized words. While they were vocalizing, we measured their EEG data. In addition to EEG data, we recorded the voice data and trigger signals to calculate the speech onset (the timing when they have vocalized).

4.2.2 Devices and Software for Experiment

For EEG data measurement, an EPOC X EEG cap from Emotiv Inc. was used. The headset is shown in Figure 4.1. The sensor positions of EPOC X EEG cap are shown in the Figure 4.2. The brain activity at the Broca's area, which is known to process languages in the brain, can be measured through the combined 14 channels from the EPOC X EEG scalp. The EPOC X EEG cap measures measures with sampling

frequency of 256 Hz, a configurable bandwidth between 0.16 to 43 Hz, and digital notch filters at 50Hz and 60Hz.

Figure 4.1: EEG cap (EPOC X) with 14 channels. The sensor felts are hydrated with a saline solution, and then are inserted into each channel for improving the conductivity, reducing the skin-electrode impedance.

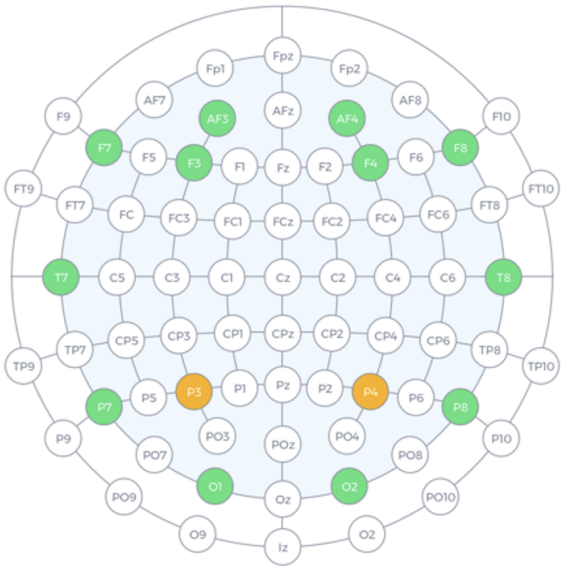

Figure 4.2: This image shows the sensor locations for EPOC X using the international 10-20 system. The green points are the electrode locations, and the orange points are the references.

The data collection process makes use of a dedicated software, EMOTIV PRO that receives the EEG data measured by the EPOC X EEG cap and displays the real

time EEG data stream on the screen as shown in Figure 4.3.

As part of the data collection process, we also recorded the voice data and trigger signals for detecting the speech onset. To record the voice data, an USB microphone was used. For the trigger signals, we built a program to send trigger signals when a word is displayed on a screen. The presentation of prompts for the words to be spoken on the screen was created by PsychoPy 3 [40].

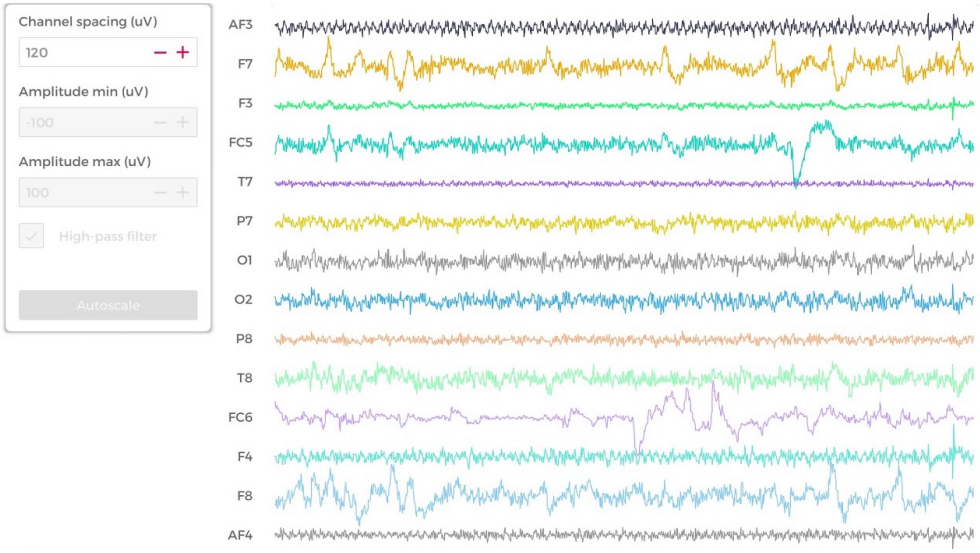

Figure 4.3: EEG data sent from the EEG cap is displayed on the screen. The window shows all 14 channels EEG activities as a time series.

To collect all data (EEG data, voice data, and trigger signals) together, we used a software called LabRecorder [41], that enables multiple signals to be recorded at the same time through the Lab Streaming Layer (LSL) protocol. For the setup, we set all data to be sent with LSL. For EEG data, we used the LSL Outlet mode in EMOTIV PRO to configure EMOTIV data streams that can communicate with the LabRecorder software. For the voice data, we used a software called AudioCapture [42] that enables the voice data to be sent through the LSL protocol. The sample rate was set to 44100 Hz. For the trigger signals, we used functions built in PsychoPy that we adapted to our own needs (PsychoPy program is in `https://github.com/sg1021/Thesis`). Figure 4.4 shows the LabRecorder's window where we can select

signals we want to record together.

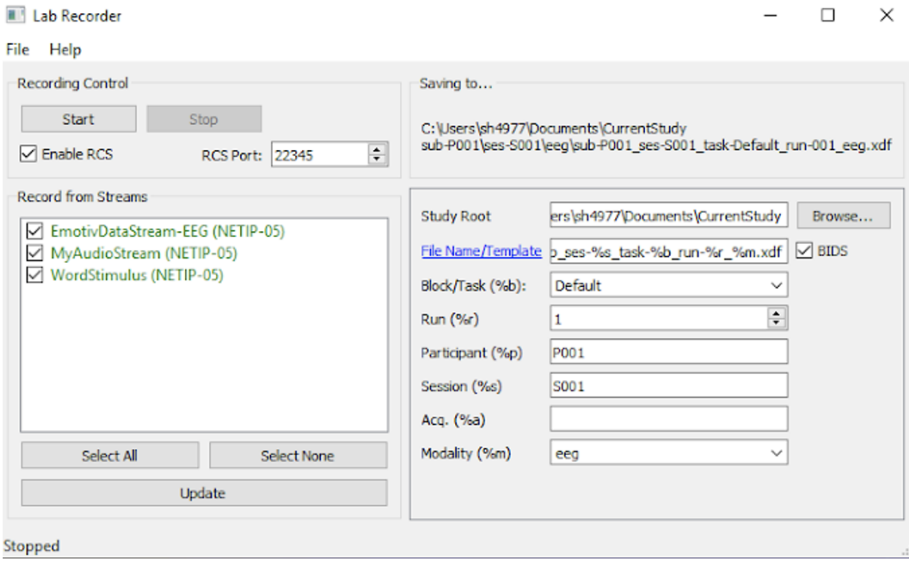

Figure 4.4: Setup window of Lab Recorder to select the data that we want to record together. Here we selected (checked) EEG data from EPOC X, audio data, and trigger data in order as we can see on the left side of the window.

4.2.3 Experiment

The experiment was carried out with 7 adult participants with no speech impairments, in the laboratory at Kanazawa Institute of Technology.

Wearing the EEG cap, each one of the participants were seated at approximately 40 cm from a 13-inch display monitor, and the microphone was located at 20 cm apart from the participant's mouth as shown in Figure 4.5.

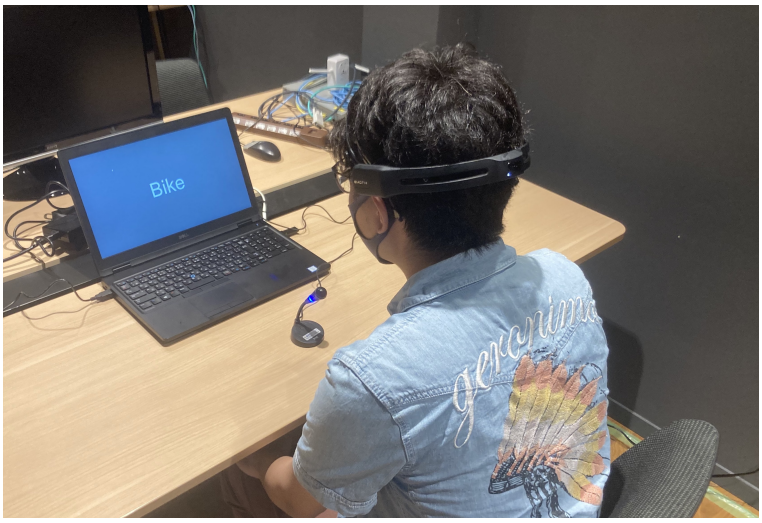

Figure 4.5: Seen in the figure how the participant wears the EEG cap. The microphone is set in front of the participant and a word to be spoken appears on the screen of the laptop.

Table 4.1 shows the word list that the participants were asked to vocalize during the EEG data measurement. This list contains 25 different words, and each word has only one syllable and are 3-5 letters long. For the purpose of this study, five different kinds of words, starting with a specific unvoiced consonant (F,B,P,M,S), were selected.

Table 4.1: Word prompts for the experiment

Phoneme Category	Word Prompt
F	Face, Fox, Fly, Faith, Free
B	Box, Bike, Body, Boom, Born
P	Pan, Pink, Push, Pool, Peace
M	Milk, Mix, Mind, Mood, Max
S	Sing, Soul, Sea, Six, Sweet

The participants wore the EEG cap and spoke words displayed on the screen following the instruction. As shown in Figure 4.6, each trial starts with a target fixation cross placed in the middle of the screen that is presented for 2 s. A word then appears for 2 s. A speech cue follows in the form of "(((> ", and "<)))" for 3 s. Participants are to speak the displayed word following the speech cue. The cue

is followed by the fixation cross of the next trial. This experiment ran in 250 trials per block, for two blocks. This is because the one block takes around 25 minutes and to allow the participants to have a rest. The trigger signals were sent every time the speech cue appeared on the screen.

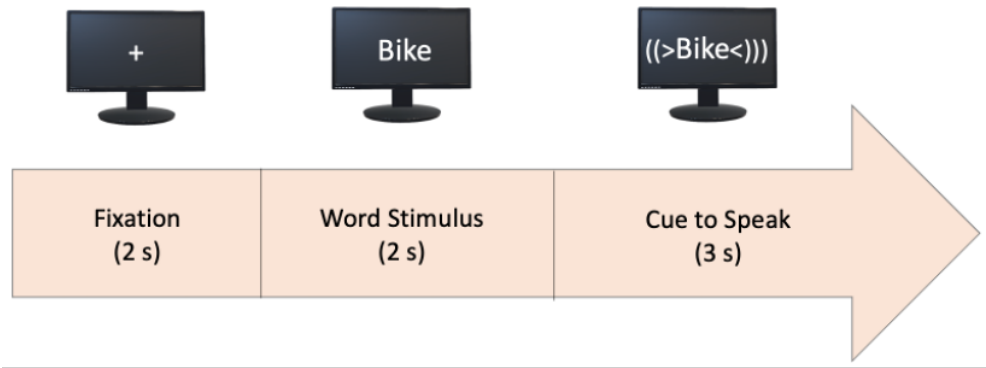

Figure 4.6: Experimental procedure: Each trial starts with a target fixation cross in the middle of the screen that is presented for 2 s. The word then appears for 2 s. A speech cue follows in the form of "(((> ", and "<)))" for 3 s. Participants are to speak the displayed word following the speech cue. The cue is followed by the fixation cross of the next trial. The experiment ran in 250 trials per block, for two blocks. The trigger signals corresponding with words were sent every time the speech cue appeared on the screen.

Before each recording session, the EEG measurement quality was calibrated by asking participants to look up, down, left and right to check whether each corresponding EEG signal was captured by the real time stream of EMOTIV PRO. Also, to record correctly, the participants practiced pronouncing each word with an English native speaker before the recording.

4.3 Preprocessing

The raw EEG data are preprocessed to enhance its quality so that features are more prominent. In this study, as we focused on the pre-speech EEG data, the EEG data were segmented into pre-defined time ranges and its amplitude was transformed into different amplitude ranges.

For the segmentation of EEG data, we used MATLAB and the EEGLAB toolbox

[43]. Before the segmentation, each speech onset of vocalization was calculated and found by using the voice data and trigger signals (the code for speech onset detection is in `https://github.com/sg1021/Thesis`). In addition, since LabRecorder causes a difference in the measurement starting time of each data, we implemented the segmentation by taking into account the time error (details are in `https://github.com/sg1021/Thesis`). Using the function built in EEGLAB, EEG data of each vocalization was segmented (epoched) into the range from -1000 ms to 0 ms with respect to the onset of speech. After the data were segmented, a technique called baseline correction was applied to remove the offset. The baseline interval was selected from -500ms to 0ms which is the first onset of speech. The features appearing in EEG data vary from person to person [44]. Therefore, the appropriate measurement location and analysis method should be chosen considering this fact. For this study, the amplitude of the EEG data was scaled into the range from -1 to 1 through the min-max scaling method. After this scaling process, we applied an IIR high pass filter with order 8 using the highpass MATLAB function, and cutoff frequency of 2 Hz to the EEG data. This is because high frequency EEG data have been observed for speech analysis [17, 45] rather than the low frequency EEG data; low frequency EEG data are ignored (removed) for EEG data analysis [16].

4.4 Classification using machine learning

4.4.1 Overview

In this part of study, we focused on verifying that the pre-speech EEG data can be used to recognize the consonants using a machine learning model called Echo State Network (ESN). We first highlight the data structure and how the hyper-parameters of ESN were adjusted based on the algorithms. Next, the evaluation method for the time series data classification is explained.

4.4.2 Data Structure

In total, we collected 3500 samples of pre-speech EEG data. Figure 4.7 shows how the data were arranged after preprocessing. The length of each EEG data is 1 second; Each EEG data contain 256 samples because of the sampling rate (256 Hz). The EPOC X EEG cap samples over 14 channels, so that the y-axis of each EEG data shows 14 different EEG signals. As our focus is to classify between the classes (F, B, S, P, M), there were five classes for classification, each class has 700 sets. For this classification, we randomly assigned 90 % of samples as the training sample, and 10 % as the test sample. With the data arrangement, we examined the consonant classification.

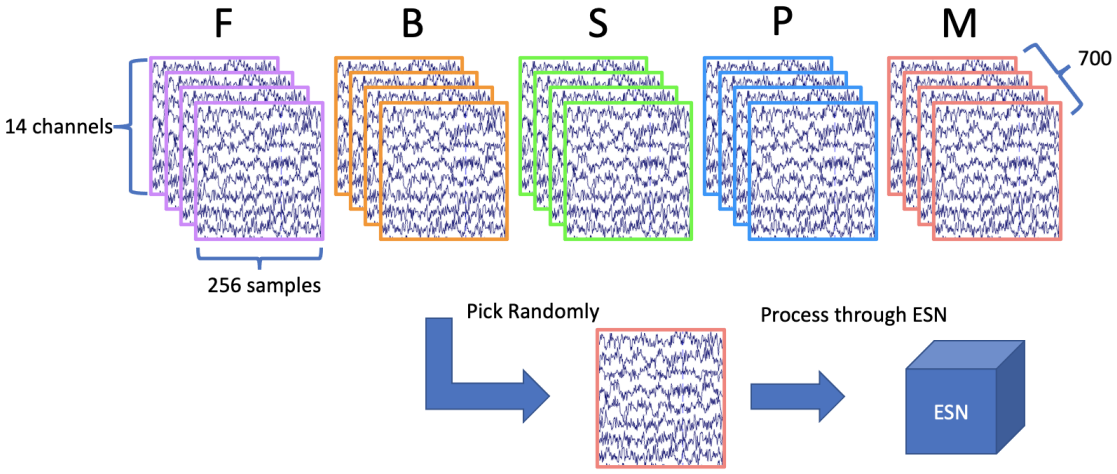

Figure 4.7: The illustration of the data structure after preprocessing: The data length of each EEG data is 256 samples (x-axis), and that of channels is 14 (y-axis). Each class of consonants has 700 samples. From the complete dataset, one EEG data were randomly picked and was processed through ESN model. This process was repeated to train the ESN model.

4.4.3 Echo State Network Model

To classify the five consonants F,B,P,M,S, we decided to use Echo State Network (ESN) model. As described in Section 2.3, the ESN model is a type of reservoir computing that uses RNN and is suitable for time series data analysis, such as EEG

data. Moreover, the ESN model dramatically reduces the computational complexity by training only the weights of the output layer. Therefore, we used a normal equations method to train the weights of the output layer. The normal equations method calculates the weights through the expression,

$$w = (M^T M)^{-1} M^T y \tag{4.1}$$

where M is the training data matrix, y is the target data vector, and w is the weight vector. However, in order to achieve high computational performance with such a method, the reservoir layer needs to be set appropriately. One of the most important settings is the hyperparameter setting.

4.4.4 Hyperparameter Adjustment

Many hyperparameter settings are required for ESN. Table 4.2 shows a list of hyperparameters we set in this part of study. We now describe the values we set for each hyperparameter, and the reason of the values. First, the number of nodes N_x, which represents the size of the reservoir, affects the number of training parameters because it is related to the size of the output weight matrix from the reservoir to the output layer, as in Equation 2.6. The larger N_x is, the more expressive the model becomes, but if N_x is too large, the computational complexity of training increases, which may lead to overfitting. Also, if N_x is too large compared to the length of the time series data T used for training, the calculations in the output layer may have negative affects [46]. Therefore, it is necessary to set N_x to as large a value as possible within the range where the computational cost of learning does not become a problem and where it is not too large compared to the length of the time series data T. Taking these factors into consideration, we set $N_x = 100$ in this study. In addition, as for the activation function, *tanh* is generally used to satisfy the ESN model requirements, so *tanh* was also set as the activation function in this study.

Secondly, the network structure can be set arbitrarily, but sparse and random networks are often used to eliminate arbitrariness [47]. It also prevents the uniformity of node states that tends to occur in dense networks, and allows each node to obtain a variety of responses to inputs data. However, if it is too sparse, the network may become unconnected. In this study, we set 90 %, which is relatively large value for the connectivity density. The reason is that the resulting network with more dense connections recorded better accuracy than the sparse network for the EEG data. Also, to determine the recurrent connectivity weight matrix in reservoir W, it is often generated randomly from some probability distribution to eliminate arbitrariness, and uniform, normal, and binary distributions are mainly used. In this study, the uniform distribution of [-1 1] was used.

Table 4.2: The list of hyperparameters we set for our study

Parameter	Meaning	Value
N_u	Number of input layer nodes	14
N_x	Number of reservoir layer nodes	100
N_y	Number of output layer nodes	5
W	Recurrent connectivity weight matrix in the reservoir	[-1 +1]
d	Density of connections in the reservoir	0.9
W^{in}	Input connectivity weight matrix	[-1 +1]
α	Leaky rate	0.009
ρ	Spectral radius of ρ	0.9

Next, we describe three hyperparameters that have a significant impact on ESN, and we have tuned these hyperparameters based on the ESN algorithm. To explain these hyperparameters, the node state vector in the recurrent layer described in Section 2.3 is noted again in Equation 4.2. We used a general model of ESN without the output feedback since the output feedback potentially cause a negative affect to the reservoir state vectors [47].

$$x(n+1) = f(W^{in}u(n+1) + Wx(n)) \qquad (n = 0, 1, 2, ...) \tag{4.2}$$

The first parameter we looked into was the input weight W^{in}. For the input weight, random numbers following an uniform distribution between -1 and +1 are used. W^{in} is one of the main parameters that affect the computational performance of ESN. The impact of W^{in} is determined by the combination of the activation function of each node in the reservoir. In this part of our study, the activation function was the tanh function. As shown in Figure 4.8, when the W^{in} interval is small (①), the transformation is likely to be performed in the almost linear part around the origin, while when it is large (②), the transformation is likely to be performed in the part with strong non-linearity. In general, the optimal range of W^{in} is determined empirically.

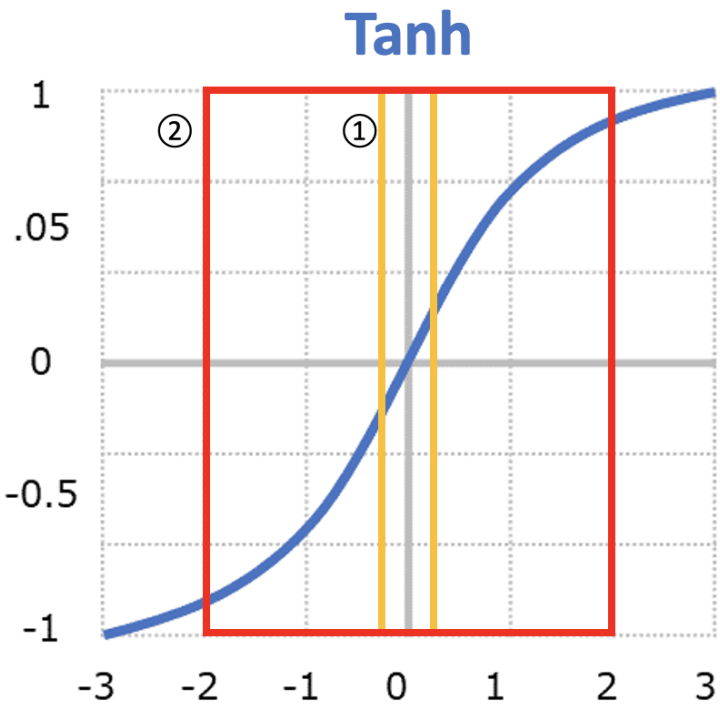

Figure 4.8: Tanh graph: The range of the tanh function is from -1 to 1. X axis represents the input of the activation function. ① When the W^{in} interval is small, the transformation is likely to be performed in the almost linear part around the origin. ② When it is large, the transformation is likely to be performed in the part with strong non-linearity.

After nodes of the reservoir computed all input data with the activation function, all these outputs were stored in array. By plotting a portion of these stored outputs, we could check whether each section of speech has a characteristic patterns. With this method, we searched for appropriate values of hyperparameters. Figure 4.9 shows the output plot when the input weight's interval was set in the range [-0.01 0.01]. Tanaka et al. [46] demonstrated the speech recognition using ESN model, and the recognition accuracy was fairly high with an interval of the input weight as the same range as that of the recurrent weight. Therefore, we also set the interval in the range [-1 1] in the same range as the recurrent weight. Figure 4.10 shows the output plot when the input weight's interval is the range between [-1 1].

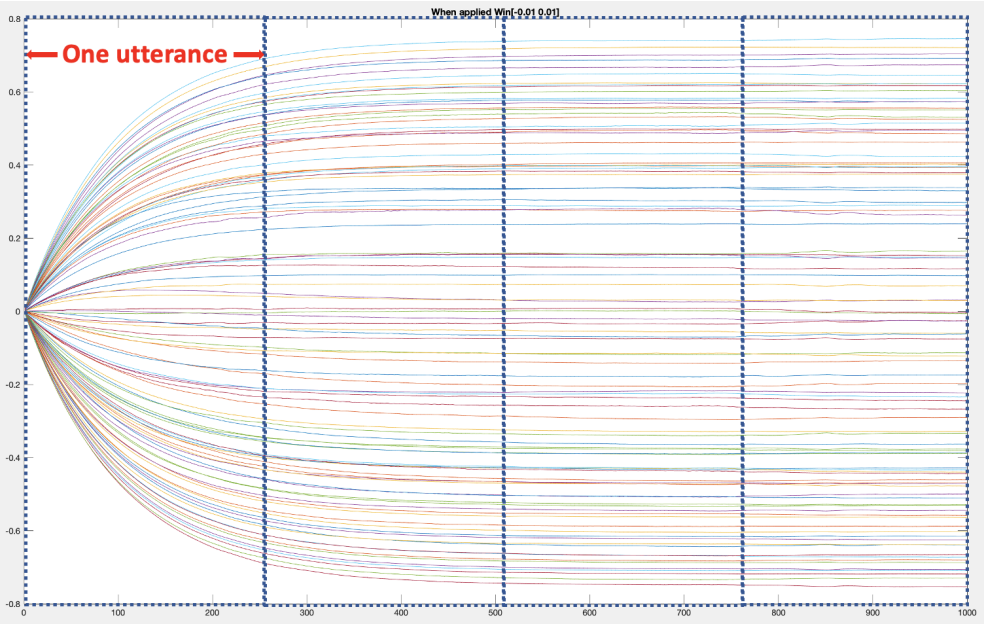

Figure 4.9: Part of the combined output of the EEG computed by the activation function of reservoir nodes with the input weight's interval was between [-0.01 0.01]. The x-axis shows the sample number, and the y-axis shows the output of the activation function. The vertical dashed lines indicates the duration when a single word was vocalized.

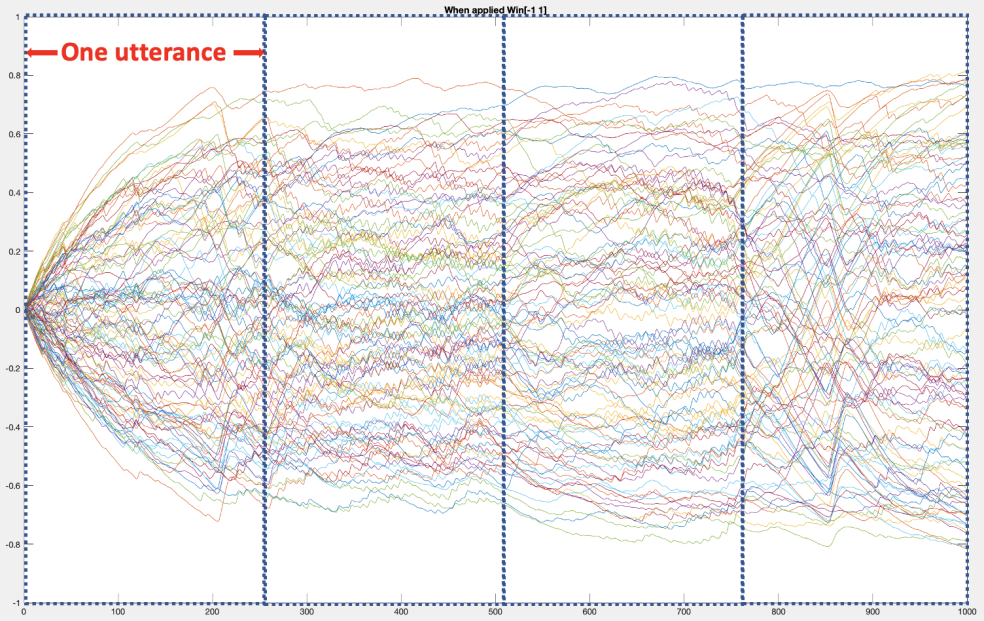

Figure 4.10: Part of the combined output of the EEG computed by the activation function of reservoir nodes with the input weight's interval was between [-1 1]. The x-axis shows the sample number, and the y-axis shows the output of the activation function. The vertical dashed lines indicates the duration when a single word was vocalized.

From the observation of these plots, we found that the output of Figure 4.9 expresses no difference between the sections. On the other hand, the output of Figure 4.10 shows the characteristics of the waveform in each section. Furthermore, we learned that when the interval range was set to more than [-1 1], the waveform becomes too expressive and no difference could be seen in each section as shown in Figure 4.11.

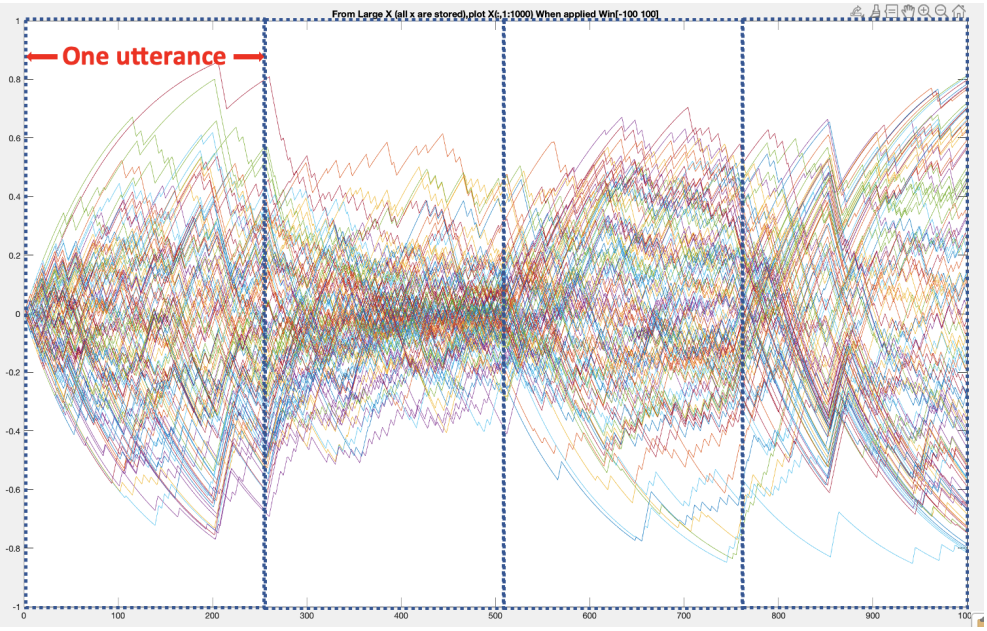

Figure 4.11: Part of the combined output of the EEG computed by the activation function of reservoir nodes with the input weight's interval was between [-100 100]. The x-axis shows the sample number, and the y-axis shows the output of the activation function. The vertical dashed lines indicates the duration when a single word was vocalized.

The second hyperparameter was the spectral radious of the recurrent weight matrix W. As described in Section 2.3.3, this parameter satisfies the Echo State Property (ESP). In addition, as we can see from Equation 4.2, the larger W gets, the more information each node retains from the previous state. Because we wanted nodes to keep the past state information of EEG data, the parameter was set to 0.9. However, though we tried changing W to several different values, we could not see significant changes from the outcome of classification and the output plot.

The third parameter was the leaky rate. As noted in Section 2.3.3, the leaky rate α defines the speed of reservoir updating. This is why the smaller α is, the slower the reservoir updating speed gets. By looking into the output plots, the parameter was set to 0.009, which consistently produced high discrimination accuracy. Regarding this hyperparameter adjustment, we found that within a small range of values, small changes in α have a large effect on the results. For example, if we set $\alpha < 0.001$, the predictions become scattered, as we can be seen in the confusion matrix of Figure

4.12. When we set $\alpha > 0.1$, there was a tendency for the predictions to be heavily concentrated on class 1 (F), as shown in Figure 4.13. This could be due to the fact that EEG data do not need to update the reservoirs as fast as it should, since the frequencies it handles are lower than those of voice data.

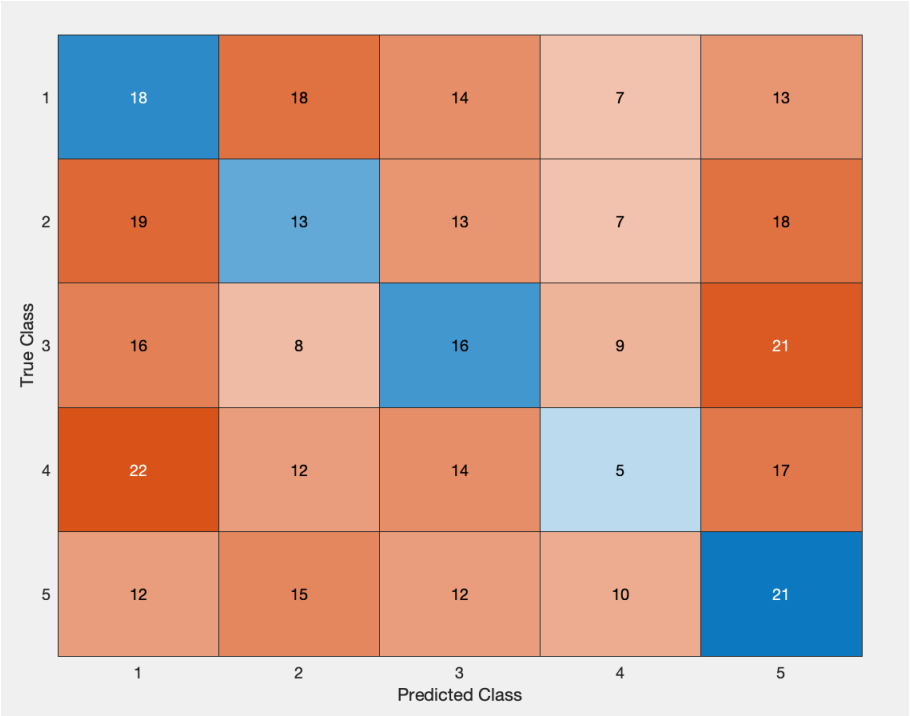

Figure 4.12: Confusion matrix when $\alpha < 0.001$: The x-axis is the prediction class, and the y-axis is the true classes. From 1 to 5, the numbers represent the following consonants: F, B, P, M, S.

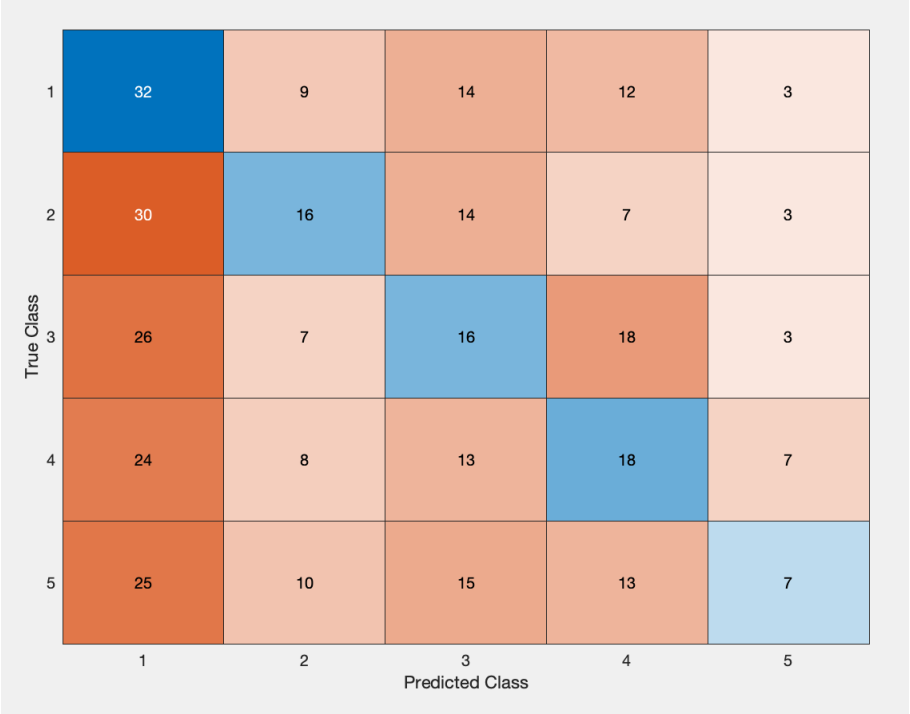

Figure 4.13: Confusion matrix when $\alpha > 0.1$: The x-axis is the prediction class, and the y-axis us the true classes. From 1 to 5, the numbers represent the following consonants: F, B, P, M, S. There was a tendency for the predictions to be heavily concentrated on class 1 (F).

4.5 Evaluation method

In a time series classification task, one class label is assigned to a part or the whole of the time series input data. In our case, one class label was given to each entire time series data. Therefore, in order to evaluate the classification performance, we need to convert the output of the trained model to a single class label.

For example, one class label $c \in 1, \ldots, N_c$ is given to the time series input data $u(n)$ $(n = 1, .., T)$ in a certain time range, where N_c represents the number of classes. To evaluate classification accuracy, the trained model's output $y(n)$ $(n = 1, .., T)$ should be converted to one class label. As the target output was applied with one-hot encoding, only the output of the c-th node $y_c(n)$ is expected to be close to 1 after training, while the other nodes are expected to be close to 0. Then, the class label

assigned to a sample $u(n)$ is the index to the output node with the largest output value.

$$\hat{k}(n) = \underset{k \in 1,...,N_y}{argmax} \quad \hat{y}_k(n) \tag{4.3}$$

Next, the most frequent value $\hat{k}(n)$ for $n = 1, .., T$ is:

$$\hat{k}(n) = MODE(\hat{k}(1), ..., \hat{k}(T)) \tag{4.4}$$

This is considered as the class label, classified by the trained model. If this element matches the correct class label $\hat{k} = c$, then the classification within the time range is successful. This procedure was performed to each time range in which the correct class label was assigned. After this process was applied, we evaluated the classification using the confusion matrix.

4.6 Result

To verify whether the pre-speech EEG data can discriminate the consonants, we measured the pre-speech EEG data and verified the classification using ESN model. We assessed classification performance by measuring the accuracy, which is the percentage of samples correctly classified by the model out of the total samples, as shown in Equation 4.5, where the correct label is i, the class label of the model is j, and such a sample is denoted as s_{ij}. The measured accuracy of consonant classification using pre-speech EEG data by ESN was 28.3% (compared to 20 % for random classification).

$$Accuracy = \frac{\sum_{i=1}^{N_c} s_{ii}}{\sum_{i=1}^{N_c} \sum_{j=1}^{N_c} s_{ij}} \tag{4.5}$$

Also, we measured the precision, calculated as shown in Equation 4.6, which

represents the percentage of samples that are correctly classified among the samples whose model classification label is j. Table 4.3 shows the measured precision rate for each of the classifications. The results show that B, P, and F were classified with relatively high precision among the classified consonants, while the classification precision for S was small.

$$PRE = \frac{s_{jj}}{\sum_{i=1}^{N_c} s_{ij}} \qquad (j = 0, 1, ..., N_c) \tag{4.6}$$

Table 4.3: Precision rate of each consonant classification

Consonant	Precision [%]
F	29.1
B	33.8
P	29.5
M	24.1
S	22.0

With regards to the above results, we note that one study found a similar, but not identical, tendency: Moses et al. [17] measured the brain activity of an anarthria participant while the participant was vocalizing to classify the specific words. Some of the words began with the five consonants that we used in our study. When we looked at their classification results for words starting with any of those five consonants, we found that, as in our study, the classification accuracy for words starting with F and B was relatively high, and the classification accuracy for word starting with S was the lowest. This could suggest that features of speech starting with B or F (consonants) are more likely to show up in the brain activity, while features of speech beginning with S (consonants) are less likely to be reflected in the brain activity.

Another possible explanation for our results of the consonant classification may be the differences in the movement of the articulators that form each speech sound.

There are two categories of consonants: Plosive and fricative. A plosive is a sound produced by completely blocking the path of the breath and then suddenly opening it as if to burst. In this part of study, P and B are examples of this. A fricative sound is made by narrowing the path of the breath and pushing the sound out of it. This is the case with F and S in this part of study. This means that when (right before) a plosive consonant is vocalized, the articulators become tense by blocking the path of the breath. Such movements of the articulators might be included in the EEG data as EMG activity, which appeared as a feature in the classification and affected the accuracy between B and P. Conversely, since the frictional sound F is vocalized with a breath path, there is no significant change in the movement of the articulatory organs, and as a result of it not being included in the EEG data as EMG signals, it has features of less significance than the plosive consonants, which may have reduced the classification accuracy. However, there are some contradictions in this consideration. This is because F, another friction sound, has a relatively high classification accuracy, and P, a plosive sound, recorded a relatively low classification accuracy in the study by Moses et al. Therefore, it may be difficult to conclude that in general the differences of the articulator movements affects the classification accuracy.

4.7 Discussion

Regarding the results of the consonant classification using ESN, there are several possible reasons and suggestions for improvement. The first possible reason is the training algorithm we used. This time we trained the model using linear regression, which is an off-line training algorithm (or called batch algorithm) to reduce the computational complexity. We speculate that the classification accuracy can be higher by a gradient-based learning algorithm with ESN model. For example, we can use another training algorithm called online training algorithm. In this algorithm, the

output weight W^{out} is updated adaptively over time. Wen et al. [48] stated that one of the online training algorithms called Least Mean Square (LSM) algorithm is suitable for ESN and easier to be achieved rather than other learning methods. They implemented a system with ESN using the LMS method. The LMS method calculates the error between the model output and the target output each time, and updates W^{out} iteratively by the gradient descent method to minimize the squared error. Therefore, applying this training algorithm to our ESN may lead to better classification accuracy.

Regarding EEG measurement, the location and number of the electrodes used in this study might not be the best for speech analysis. In [49], in order to capture the brain signals from the language area of the brain, 21 electrodes were located on the left hemisphere, over the area of Wernicke and Broca; these areas are related with prosecution of brain language. The EEG we used also has electrodes located close to those two areas, but not placed enough electrodes in those areas compared to their measurement. In addition, Montoya-Martínez et al. [50] determined the optimal number of electrodes, and the best position to place a limited number of electrodes on the scalp for language analysis. Regarding the number of electrodes, they have reported that 32 electrodes is the best, while the EEG we used has only 14 channels, which may be insufficient to fully measure EEG information for language analysis. Also, when we compared the best positions they reported with the positions of the electrodes used in our study, the EEG we used did not have the electrode at the top of the brain, and other positions were also slightly off.

Regarding preprocessing, in the fact that we initially wanted to see what the output would look like if we analyzed and classified the data as is, we did not conduct significant preprocessing. However, more preprocessings might be necessary to extract rich data from the EEG data. As artifacts from eye blinking or eye movements (known as EOG signals) are included in the EEG data measurement, these artifacts

are removed by preprocessing [51]. Moreover, as Jenson et al. [52] carried out for EEG data analysis, the multiple artifact rejection algorithms (MARA) could be applied to identify and remove components identified as an artifact.

4.8 Conclusion

To discriminate the consonants from pre-speech EEG data, we measured the EEG data and applied preprocessings, and then classified the consonants F, B, P, M, and S using an ESN model. As a result, the overall accuracy was 28.3 %. Through conducting this study, several challenges have been addressed. First of these challenges is the analysis of pre-speech EEG data. Studies of measuring and analyzing EEG data before and during speech are generally rare, but such implementation has been examined in studies of stuttering. For example, Toyomura et al. [53] recorded EEG data of participants receiving auditory feedback from their own voices during instruction in which they vocalized one vowel, and conducted an analysis of the EEG data in speech condition. Daliri et al. [54] examined whether different types of auditory stimuli have different effects on auditory processing of speech planning in an experiment in which participants were presented with sound stimuli prior to the onset of speech, and their EEG data were acquired during speech production. In a study using magnetoencephalography(MEG), another method used to analyze brain activity different from EEG, Mersov et al. [55] acquired and analyzed brain activity prior to speech to characterize neural oscillations in the speech motor network during preparation for and execution of overt speech production. However, most of these studies have focused on brain waves during and after speech, or even if they have measured brain waves before speech, they have not addressed the characteristics of speech and its classification. In this part of our study, we were able to examine speech classification by focusing on the pre-speech EEG data and the speech features, which has not been done before.

Another challenge challenge was how to apply an ESN model to classify speech from EEG data. There have not been many studies of speech identification from EEG data using an ESN model. The ESN has been used in classification research largely for speech data, 1-dimensional sensor data, and image data [56, 57, 58, 59]. In classification studies using EEG data, which is relatively complex in terms of the number of dimensions and the amount of noise, ESN has been often used to classify specific psychological states [60] or emotion recognition [61], but not speech classification. Through this part of our study, we were able to analyze the relationship between speech and the (pre-speech) EEG data through an ESN.

At the same time, in terms of speech classification accuracy, we need more improvements. Ghane et al. [16] examined vowel classification from EEG data, in which the EEG data of participants were measured while the participants were imagining vowels. As a result, the vowels were recognized with an accuracy of 76.6 %. In another study, though this focus was also not the pre-speech EEG data, Moses et al. [17] recorded cortical activity while the participant attempted to say individual words from a vocabulary set. The model classified words with 47.1 % accuracy. Though it is hard to compare our study with these studies because the data (pre-speech EEG) we used and the purpose (consonant classification) we focused on were different than their studies. Yet, the classification accuracy for consonant in this study cannot be considered as high. Therefore, we should address the improvements we described earlier as the future work for enhanced classification accuracy.

Chapter 5

Conclusion and Future Work

In this study, we have conducted two studies to find new biological signals that can be used for speech interface. First study was the Japanese vowel classification using the vocal folds vibration data. We measured the acceleration data of the vocal folds vibration from a sensor attached to the neck over the throat and applied the spectral analysis to the measured data to extract the feature values. We selected the first and second harmonics with larger power as the feature values to train the machine learning model. For the two-class classification, we used linear SVM for the classification of Japanese vowels. The classifier performance was of 71 % accuracy on average. All classification instances yielded fairly high accuracy, except that of between /a/ and /e/. This can be because both frequency values are similar, so that it was difficult to discriminate them in the frequency domain. Therefore, we observed that the combination with other features besides harmonics can lead to better accuracy.

The second study was the unvoiced consonant recognition using pre-speech EEG data. We measured the EEG data while the participants were vocalizing words. The words the participants vocalized were 25 different words, which start with a specific consonant (F, B, P, M, S). The EEG data were segmented and their amplitudes were transformed to a certain range. For the unvoiced consonant classification, we used an ESN model as the classifier, which processes fast and is able to handle time series data. As a result, the classification accuracy of 5 classes was 28.3 % (better than

the 20% accuracy corresponding to random classification). The possible reason for the accuracy is the training algorithm because the simple linear training method was used for reducing the computational complexity in this study. To obtain the better classification accuracy, another training algorithm such as a gradient-based learning algorithm can be addressed.

Through these two studies, we verified and found the possibility that two biological signals (the vocal folds vibration and the pre-speech EEG data) can be used to discriminate speech though there still is room for improvement. By enhancing the recognition accuracy more, these biological signals can be used for new developments of speech interfaces to help people who have speech disorders. In addition, this thesis describes a new method to capture the characteristics of vocal folds vibration from acceleration data, and to analyze the brain waves before speech to predict speech sound. We believe that these results together with further improvements would lead to advanced developments of speech interface.

In the future, we plan to combine the two biological signals that were studied in this thesis(the vocal folds vibration and the pre-speech EEG data) together to verify the accuracy of speech classification. We also plan to build a real-time speech classification system using these biological signals with the improvements we have learned.

Bibliography

[1] 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, April 2010.

[2] John J. Magee, Margrit Betke, James Gips, Matthew R. Scott, and Benjamin N. Waber, "A human-computer interface using symmetry between eyes to detect gaze direction," *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, vol. 38, no. 6, pp. 1248–1261, 2008.

[3] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, "Development of a (silent) speech recognition system for patients following laryngectomy," *Medical Engineering and Physics*, vol. 30, no. 4, pp. 419–425, May 2008.

[4] Jose A. Gonzalez-Lopez, Alejandro Gomez-Alanis, Juan M. Martín Doñas, José L. Pérez-Córdoba, and Angel M. Gomez, "Silent speech interfaces for speech restoration: A review," *IEEE Access*, vol. 8, pp. 177995–178021, 2020.

[5] Qinwan Rabbani, Griffin Milsap, and Nathan E. Crone, "The Potential for a Speech Brain–Computer Interface Using Chronic Electrocorticography," *Neurotherapeutics 2019 16:1*, vol. 16, no. 1, pp. 144–165, January 2019.

[6] "Types of Speech Disorders and Therapy Options — Insight Medical Campus," (2021). [Online]. Available: `https://www.iinn.com/types-of-speech-disorders-therapy/`.

[7] Brad H. Story, "An overview of the physiology, physics and modeling of the sound source for vowels," *Acoustical Science and Technology*, vol. 23, no. 4, pp. 195–206, July 2002.

[8] Medical Research Council, ""THE BRAIN"-MRC research for lifelong health," 2009, [Online]. Available: `http://www.mrc.ac.uk/`. Accessed: January 29, 2022.

[9] "Vowel Sounds," 2017. [Online]. Available: `http://hyperphysics.phy-astr.gsu.edu/hbase/Music/vowel.html`.

[10] Ronald J. Baken, "Electroglottography," *Journal of Voice*, vol. 6, no. 2, pp. 98–110, January 1992.

[11] Robert Thayer Sataloff, Steven Mandel, Yolanda D. Heman-Ackah, and Mona Abaza, "Laryngeal electromyography," [Online]. Available: `https://books.google.com/books/about/Laryngeal_Electromyography_Third_Edition.html?hl=ja&id=j-x6DwAAQBAJ`.

[12] Barry W. Connors Bear, Mark F. and Michael A. Paradiso, "Neuroscience: exploring the brain," 2001, Baltimore, Md: Lippincott Williams Wilkins.

[13] Hanshu Cai, Jiashuo Han, Yunfei Chen, Xiaocong Sha, Ziyang Wang, Bin Hu, Jing Yang, Lei Feng, Zhijie Ding, Yiqiang Chen, and Jürg Gutknecht, "A Pervasive Approach to EEG-Based Depression Detection," *Complexity*, vol. 2018, 2018.

[14] Adeen Flinker, Anna Korzeniewska, Avgusta Y. Shestyuk, Piotr J. Franaszczuk, Nina F. Dronkers, Robert T. Knight, and Nathan E. Crone, "Redefining the role of Broca's area in speech," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 9, pp. 2871–2875, March 2015.

[15] Robert Teasell, Norhayati Hussein, Mbbs Mrehabmed, Ricardo Viana, Sarah Donaldson Bhsc, and Mona Madady Msc, "Stroke Rehabilitation Clinician Handbook," (2020). [Online]. Available: `http://www.ebrsr.com/sites/default/files/Chapter%201_Clinical%20Consequences_0.pdf`.

[16] Parisa Ghane and Gahangir Hossain, "Learning Patterns in Imaginary Vowels for an Intelligent Brain Computer Interface (BCI) Design," 2020, [Online]. Available: `http://arxiv.org/abs/2010.12066`.

[17] David A. Moses, Sean L. Metzger, Jessie R. Liu, Gopala K. Anumanchipalli, Joseph G. Makin, Pengfei F. Sun, Josh Chartier, Maximilian E. Dougherty, Patricia M. Liu, Gary M. Abrams, Adelyn Tu-Chan, Karunesh Ganguly, and Edward F. Chang, "Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria," *New England Journal of Medicine*, vol. 385, no. 3, pp. 217–227, July 2021.

[18] Nello Cristianini and John Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*, Cambridge University Press, 2000.

[19] Paul J. Werbos, "Backpropagation Through Time: What It Does and How to Do It," *Proceedings of the IEEE*, vol. 78, no. 10, pp. 1550–1560, 1990.

[20] Sepp Hochreiter and Jürgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, November 1997.

[21] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014.

[22] Gouhei Tanaka, Toshiyuki Yamane, Jean Benoit Héroux, Ryosho Nakane, Naoki Kanazawa, Seiji Takeda, Hidetoshi Numata, Daiju Nakano, and Akira Hirose, "Recent advances in physical reservoir computing: A review," *Neural Networks*, vol. 115, pp. 100–123, July 2019.

[23] Nils Bertschinger and Thomas Natschläger, "Real-time computation at the edge of chaos in recurrent neural networks," *Neural Computation*, vol. 16, no. 7, pp. 1413–1436, July 2004.

[24] Qiuyi Wu, Ernest Fokoue, and Dhireesha Kudithipudi, "On the statistical challenges of echo state networks and some potential remedies," 2018, [Online]. Available: `https://arxiv.org/abs/1802.07369`.

[25] Gyungmin Toh and Junhong Park, "Review of vibration-based structural health monitoring using deep learning," *Applied Sciences*, vol. 10, pp. 1680, March 2020.

[26] Jose A. Gonzalez, Lam A. Cheah, Angel M. Gomez, Phil D. Green, James M. Gilbert, Stephen R. Ell, Roger K. Moore, and Ed Holdsworth, "Direct Speech Reconstruction from Articulatory Sensor Data by Machine Learning," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 25, no. 12, pp. 2362–2374, December 2017.

[27] D. C. Toledo-Perez, Juvenal Rodriguez-Resendiz, and Roberto A. Gomez-Loenzo, "A study of computing zero crossing methods and an improved proposal for EMG signals," *IEEE Access*, vol. 8, pp. 8783–8790, 2020.

[28] Lawrence R. Rabiner, "On the Use of Autocorrelation Analysis for Pitch Detection," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 25, no. 1, pp. 24–33, 1977.

[29] A. Michael Noll, "Cepstrum Pitch Determination," *The Journal of the Acoustical Society of America*, vol. 41, no. 2, pp. 293, July 2005.

[30] A. Michael Noll, "Short-Time Spectrum and "Cepstrum" Techniques for Vocal-Pitch Detection," *The Journal of the Acoustical Society of America*, vol. 36, no. 2, pp. 296, July 2005.

[31] Maya Kallas, Paul Honeine, Cédric Richard, Clovis Francis, and Hassan Amoud, "Prediction of time series using Yule-Walker equations with kernels," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 2185–2188, 2012.

[32] Maya Kallas, Clovis Francis, Paul Honeine, Hassan Amoud, and Cédric Richard, "Modeling electrocar-diogram using Yule-Walker equations and kernel machines," pp. 1–5, 2012.

[33] M. C. Chevalier and Y. Grenier, "Autoregressive models with time-dependent log area ratios," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pp. 1049–1052, 1985.

[34] Nadine de Rue, Alejandrina Cristia, Paula Fikkert, and Sho Tsuji, *Directional asymmetries in vowel perception*, Ph.D. thesis, January 2014.

[35] Sandeep Kumar and Jainath Yadav, "Implementation of audio recognition using mel frequency cepstrum coefficient and dynamic time warping in wirama praharsini," [Online]. Available: `https://ui.adsabs.harvard.edu/abs/2021JPhCS1722a2014W/abstract`.

[36] Bharti Gawali, Santosh Gaikwad, Pravin L. Yannawar, Suresh C. Mehrotra, and Babasaheb Ambedkar, "Marathi isolated word recognition system using mfcc and dtw features," 2011, [Online]. Available: `https://www.bibsonomy.org/bibtex/114a59b543f73e5b42e54bd55d05b77ac/ideseditor`.

[37] Jorge Martinez, Hector Perez, Enrique Escamilla, and Masahisa Mabo Suzuki, "Speaker recognition using Mel Frequency Cepstral Coefficients (MFCC) and Vector quantization (VQ) techniques," *CONIELECOMP 2012 - 22nd International Conference on Electronics Communications and Computing*, pp. 248–251, 2012.

[38] Alexander Mayr, "Parameters of Flow Glottogram and EGG for Vocal Registers-Modal, Falsetto and voce faringea," 2014. [Online]. Available: `https://phaidra.kug.ac.at/view/o:10962`.

[39] Uezu Yasufumi, "A study on the effect of source–filter interaction on the production of speech," [Online]. Available: `https://doi.org/10.15017/1807042`.

[40] Jonathan Peirce, Jeremy R. Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv, "PsychoPy2: Experiments in behavior made easy," *Behavior Research Methods*, vol. 51, no. 1, pp. 195–203, February 2019.

[41] "labstreaminglayer/App-LabRecorder: An application for streaming one or more LSL streams to disk in XDF file format.," 2021. [Online]. Available: `https://github.com/labstreaminglayer/App-LabRecorder`.

[42] "labstreaminglayer/App-AudioCapture: Capture audio and stream it over LabStreamingLayer," 2021. [Online]. Available: `https://github.com/labstreaminglayer/App-AudioCapture`.

[43] Arnaud Delorme and Scott Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, pp. 9–21, 2004.

[44] Shin-ichi Ito, Yasue Mitsukura, Minoru Fukumi, and Norio Akamtsu, "Proposal of the EEG Analysis Method Using the Individual Characteristic of the EEG," IEEJ Transactions on Electronics, Information and Systems, 2004, Volume 124, Issue 6, Pages 1259-1266.

[45] Anna Maria Mersov, Cecilia Jobst, Douglas O. Cheyne, and Luc De Nil, "Sensorimotor Oscillations Prior to Speech Onset Reflect Altered Motor Networks in

Adults Who Stutter," *Frontiers in Human Neuroscience*, vol. 10, no. SEP2016, September 2016.

[46] Gōhei Tanaka, Ryōshō Nakane, and Akira Hirose, "Reservoir Computing: Theory and Hardware of Fast Machine Learning for Time Series Pattern Recognition," 2021. [Online]. Available: `https://www.jstage.jst.go.jp/article/jsoft/33/2/33_65/_article/-char/ja`.

[47] Mantas Lukoševičius, "A Practical Guide to Applying Echo State Networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 7700 LECTU, pp. 659–686, 2012.

[48] Shiping Wen, Rui Hu, Yin Yang, Tingwen Huang, Zhigang Zeng, and Yong Duan Song, "Memristor-Based Echo State Network with Online Least Mean Square," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 9, pp. 1787–1796, September 2019.

[49] L. C. Sarmiento, P. Lorenzana, C. J. Cortés, W. J. Arcos, J. A. Bacca, and A. Tovar, "Brain computer interface (BCI) with EEG signals for automatic vowel recognition based on articulation mode," *ISSNIP Biosignals and Biorobotics Conference, BRC*, 2014.

[50] Jair Montoya-Martínez, Jonas Vanthornhout, Alexander Bertrand, and Tom Francart, "Effect of number and placement of EEG electrodes on measurement of neural tracking of speech," *PLoS ONE*, vol. 16, no. 2, February 2021.

[51] Ayoub Daliri and Ludo Max, "Electrophysiological evidence for a general auditory prediction deficit in adults who stutter," *Brain and Language*, vol. 150, pp. 37–44, 2015.

[52] David Jenson, Andrew L. Bowers, Daniel Hudock, and Tim Saltuklaroglu, "The Application of EEG Mu Rhythm Measures to Neurophysiological Research in Stuttering," *Frontiers in Human Neuroscience*, vol. 13, no. January, pp. 1–22, 2020.

[53] Akira Toyomura, Daiki Miyashiro, Shinya Kuriki, and Paul F. Sowman, "Speech-Induced Suppression for Delayed Auditory Feedback in Adults Who Do and Do Not Stutter," *Frontiers in Human Neuroscience*, vol. 14, pp. 150, April 2020.

[54] Ayoub Daliri and Ludo Max, "Modulation of auditory responses to speech vs. Nonspeech stimuli during speech movement planning," *Frontiers in Human Neuroscience*, vol. 10, May 2016.

[55] Ayoub Daliri, Ludo Max, Gang Li, Bao Jian Li, Xu Guang Yu, Chun Tian Cheng, Ayoub Daliri, Ludo Max, Jose A. Gonzalez-Lopez, Alejandro Gomez-Alanis, Juan M. Martín Doñas, José L. Pérez-Córdoba, Angel M. Gomez, Joseph Kalinowski, Tim Saltuklaroglu, Parisa Ghane, Gahangir Hossain, Akira Toyomura, Daiki Miyashiro, Shinya Kuriki, Paul F. Sowman, Anna Maria Mersov, Cecilia Jobst, Douglas O. Cheyne, Luc De Nil, B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, J. S. Brumberg, L. C. Sarmiento, P. Lorenzana, C. J. Cortés, W. J. Arcos, J. A. Bacca, and A. Tovar, "Sensorimotor oscillations prior to speech onset reflect altered motor networks in adults who stutter," *Frontiers in Human Neuroscience*, vol. 8, no. 4, pp. 1–16, April 2020.

[56] Stuart E. Lacy, Stephen L. Smith, and Michael A. Lones, "Using echo state networks for classification: A case study in Parkinson's disease diagnosis," *Artificial Intelligence in Medicine*, vol. 86, pp. 53–59, March 2018.

[57] Harsh Shrivastava, Ankush Garg, Yuan Cao, Yu Zhang, and Tara Sainath, "Echo state speech recognition," 2021, [Online]. Available: `https://arxiv.org/abs/`

2102.09114.

[58] Nils Schaetti, Michel Salomon, and Raphaël Couturier, "Echo State Networks-Based Reservoir Computing for MNIST Handwritten Digits Recognition," *2016 19th IEEE Intl Conference on Computational Science and Engineering (CSE), IEEE 14th Intl Conference on Embedded and Ubiquitous Computing (EUC), and 15th Intl Symposium on Distributed Computing and Applications for Business Engineering (DCABES)*, pp. 484–491, August 2016.

[59] Simone Scardapane and Aurelio Uncini, "Semi-supervised Echo State Networks for Audio Classification," *Cognitive Computation*, vol. 9, no. 1, pp. 125–135, February 2017.

[60] Dong Hwa Jeong and Jaeseung Jeong, "In-Ear EEG Based Attention State Classification Using Echo State Network," *Brain Sciences*, vol. 10, no. 6, June 2020.

[61] Rahma Fourati, Boudour Ammar, Javier Sanchez-Medina, and Adel M. Alimi, "Unsupervised learning in reservoir computing for eeg-based emotion recognition," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.