テーマ番号	1D01			
論文題目	和文	生体情報に含まれるスピーチの特徴分析と識別	指導教員	中沢　実 教授
	英文	Speech Feature Analysis and Discrimination in Biological Information		
氏名		2D1-10　本田　彰吾 (Shogo Honda)		

Abstract Today, a variety of speech interfaces have been developed using biological signals such as eye movement, heart rate, and so on. However, there are many language disorder problems that have not been solved yet by the current technologies. The possible reason we focused on was the small number of the biological signals used for the speech interface. Finding new biological signals that have speech features, more various speech interfaces can be invented. Therefore, we aim to find new biological signals that can be used for speech interface developments. What we focused on in this study were the vocal folds vibration and brain waves. After measuring the data and extracting the features, we verified whether these data can classify speech by machine learning models. As a result, using the vocal folds vibration data, the Japanese vowels could be classified with 71 % accuracy on average and using the brain waves, five classes of consonants were classified with 30 % accuracy on average.

Keywords Machine Learning, Speech interface, Signal Processing, Biological Signal, Classification

1. Introduction

Today, there has been a lot of research on new communication tools and systems (speech interface) that use biological signals such as eye movement [1], or the articulatory movement [2] to help people who have language disorders communicate with others. However, these technologies still have limitations that cannot apply to all speech disorders; There are many speech disorders not covered yet.

The cause of the problem can be considered to be the technology, but one of the possible reasons would be the small number of biological signals that can be used for speech interface development. Finding new biological signals corresponding to speech, we could develop speech interfaces that solve more language disorders. Therefore, in this study, we aim to find new biological signals that contribute to the development of communication interfaces.

For finding new biological signals used for speech interface, we analyzed some biological signals whether they include features associated with speech. In addition, we verified if these biological signals can be used to discriminate speech or not. In this study, two biological signals we focused on were (1) the vibration data of the vocal folds acquired by an acceleration sensor, and (2) the brain waves measured as EEG (electroencephalogram). The reasons focusing on the vocal folds vibration are the vocal folds cannot be damaged by language disorders in most cases, so that the data of the vocal folds vibration can be used as a resource of speech interface for most language disorders. Secondly, the vocal folds vibration is the origin of producing the voice. Hence, we expect the vocal folds vibration can be used as the new effective information for speech interface developments for various language disorders. Regarding the brain waves, the brain controls many functions of the body including speech production, so that it is persuasive to use the brain waves for speech interface. Therefore, we verified whether brain waves have speech features or not and whether the biological signals can discriminate speech.

As the research approach, we measured biological signals while participants vocalized some specific words categorized in classes. After measuring the biological signals, the data were applied signal processing to have more efficient features corresponding to speech. Using these features, we trained machine learning models and verified if the model can discriminate speech or not.

This study outcomes contribute to various aspects. First, this study would lead to new developments of speech interfaces to help people who have language disorders that have not been covered yet by the current interfaces. Moreover, more studies that find new biological signals could be carried out for speech interface developments.

The remainder of this document is organized as follows: Section 2 provides a background of vocal folds vibration, brain waves, and machine learning models. section 3 highlights and analyzes vocal folds vibrations for Japanese vowels classification using Support Vector Machine. section 4 highlights and analyzes pre-speech EEG data for the prediction of words using Echo State Network. section 5 concludes the key findings of this study and illustrates the future work.

2. Background

In this section, we highlight the necessary background to understand the study we have conducted. First, the concept of the vocal folds vibration and EEG is described, and the relationship with speech is illustrated. For the last, the machine learning model, Echo State Network, we used to be introduced.

2.1 Vocal Folds Vibration

Vocal folds vibration is a process that the vocal folds produce sound when they periodically vibrate as air passes through them during the exhalation of air from the lungs. The vocal folds are located within the larynx (voice box) at the top of the trachea. The vibration of vocal folds is represented by frequency rate (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; for a female voice, it is 200Hz.

2.1.1 Relation with Speech

First, the voiced sound is produced by vocal folds as described above. Secondly, the voiced sound is 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 sound is modified by the vocal tract articulators (the tongue, soft palate, and lips) to produce recognizable words. For example, each form of resonators and articulators (vocal tract, mouth cavity, and nasal passages) are unique when we speak the sounds of /a/ and /i/.

2.1.2 Speech Interface Study of Vocal Folds Vibration

As we have seen that vocal folds vibration is the resource of speech above section, Speech and vocal folds vibration have a strong relationship and can consider that vocal folds vibration contains the features of speech. For capturing the feature of vocal folds vibration, there are some existing methods.

One of the methods is Electroglottograph (EGG), which measure the degree of contact between the vocal folds by

measuring the electrical impedance from the two electrodes plates placed on the neck at the level of the larynx [3]. 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). EGG is mainly used for diagnosis of the larynx disorders. Another method is electromyography (EMG). With wire electrodes inserted in the different mussels, EMG measures the electrical potential created by muscle cells when these cells are active and are at rest [4]. This technique is commonly used to evaluate the health condition of muscles and the nerve cells that control them.

By contrast, it is difficult for these methods to capture the speech features that can be used for speech interface, which discriminate speech.

2.2 Electroencephalography

Electroencephalography (EEG) is a method to record the electrical activity of the brain from the scalp. The electrical activity is produced by neurons communication in the cortex. When neutrons communicate their information to each other, generate the electrical potential (Action potential and Postsynaptic potential). The electrical potential fluctuations are measured by 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 [5].

2.2.1 Relation with Speech

The brain controls all the body's functions including speech. In the brain, several areas are corresponding with each function. Broca's area, located in the left hemisphere, is associated with speech production and articulation. Information of Broca's area is taken 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 [6].

2.2.2 Study of Speech and EEG

Some studies verified that EEG signals have features associated to speech. For example, Ghane et al. [7] used EEG signals measured 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 %. Another example: Moses et al. [8] recorded cortical activity while the participant attempted to say individual words from a vocabulary set. As a result, they classified words with 47.1 % accuracy.

From these studies, it is 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 for recognition.

2.3 Echo State Network

Echo State Network is one of the representative models of reservoir computing. About ESN model, a recurrent neural network with fixed weights connectivity (Reservoir) is used to generate a state in which the past information of the time series input remains echoed (Echo State), from which the features of the input are read out (Readout). To adjust the readout, a linear learner with low computational complexity is used. The goal of this system is to achieve both high computational performance and fast learning.

The basic model structure of ESN is shown in Figure 1. This is the same structure as that of a general recurrent neural network, except that the weights connectivity of the recurrent layer is fixed and the recurrent layer is used as a reservoir. The only thing that is adjusted by the learning algorithm is the output weight matrix Wout. 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.

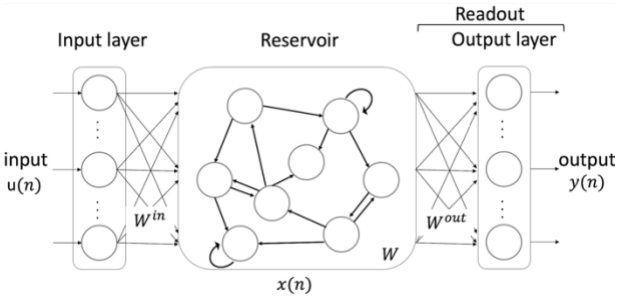

Fig.1 General ESN model structure

In Figure 1, due to handling time series data, each node state changes according to the 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:

$$u(n) = \left(u_1(n), \dots, u_{N_u}(n)\right)^T \in R^{N_u}$$
$$x(n) = \left(x_1(n), \dots, x_{N_x}(n)\right)^T \in R^{N_x}$$
$$y(n) = \left(y_1(n), \dots, y_{N_y}(n)\right)^T \in R^{N_y}$$

Where Nu, Nx, and Ny represent the number of nodes in the input layer, the recurrent layer, and the output layer, respectively. The weight connectivity between the input layer and recurrent layer is:

$$W^{in} = \left(w_{ij}^{in}\right) \in R^{N_x \times N_u}$$

The weight connectivity of the inside recurrent layer is:

$$W = \left(w_{ij}\right) \in R^{N_x \times N_x}$$

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

$$W^{out} = \left(w_{ij}^{out}\right) \in R^{N_y \times N_x}$$

At the moment, the node state vector in the recurrent layer is:

$$x(n+1) = f\left(W^{in}u(n+1) + Wx(n)\right) \qquad (n = 0,1,2, \dots)$$

Here, f represents an activation function. Using the node state vector in the recurrent layer, the output vector is:

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

3. STUDY 1: Japanese Vowel Discrimination by Throat Vibration

Section 1 described the necessity of finding 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 fold vibration can be efficient biological information that includes speech features. However, we also learned that the existing methods for the measurement of the vocal folds vibration (EGG and EMG) cannot be used enough for the development of speech interfaces because the measured data does not have enough features that can discriminate speech.

Therefore, in this study, we focused on using an acceleration sensor to measure the vibration data of the vocal folds. By attaching the acceleration sensor to the throat, data can be measured without any surgeries. In addition, we verified whether the acceleration data can be used to discriminate specific speech with the Support Vector Machine model. This section illustrates the following order: The measurement, the feature extraction, the classification method, and the result.

3.1 Measurement

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 was recorded while the subject was vocalizing specific sounds (Japanese vowels).

3.1.1 Devices for Measurement

To measure the acceleration data, we used TSND121, which invented by [9]. This device is equipped in a small wireless multi-functional sensor, including an acceleration measurement. The sampling frequency is 1000 Hz and the acceleration range is ±2G. For collecting the acceleration data from the sensor, we used the dedicated software, ALTIMA [9]. The collected data sent to ALTIMA is displayed in the software and saved to an excel file.

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 timing where the participant vocalized (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 is also sent to ALTIMA.

3.1.2 Experiment Procedure

For measuring 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.

The phonemes the subject was asked to vocalized are Japanese vowels (/a/, /i/, /u/, /e/, and /o/). The reason why we use only (Japanese) vowels in this study is that there have not been any studies verifying the possibility of even vowel discrimination by the vibration of the vocal folds. In addition to it, 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 2. 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.

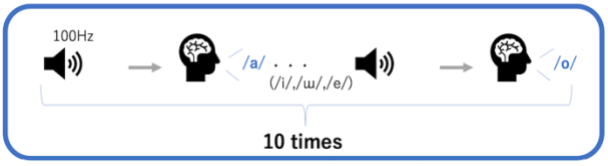

Fig.2 The procedure of the acceleration measurement. The participant was asked to vocalize each vowel sound in order. To ensure that the pitch of each vowel is the same, the subject listened to the 100 Hz tone every time before the subject vocalizes. This process was repeated 10 times.

3.1.3 Measured Data

Figure 3 shows the collected acceleration data at one block. The x-axis represents the time, and the y-axis represents the acceleration data. Because the five vowels are 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 the unique wave form, we observed each vowel's signal close by zooming it within 0.05 seconds. By zooming it within 0.05 seconds, we could see the vibration form of the vocal folds in one cycle. It can be seen that the degree of kurtosis and concavity differs slightly for each vowel. However, we cannot still see the significant differences among the signals, so that it may be hard to discriminate each speech with the raw data. Therefore, we applied the feature extraction before feeding the data to the machine learning model for speech classification.

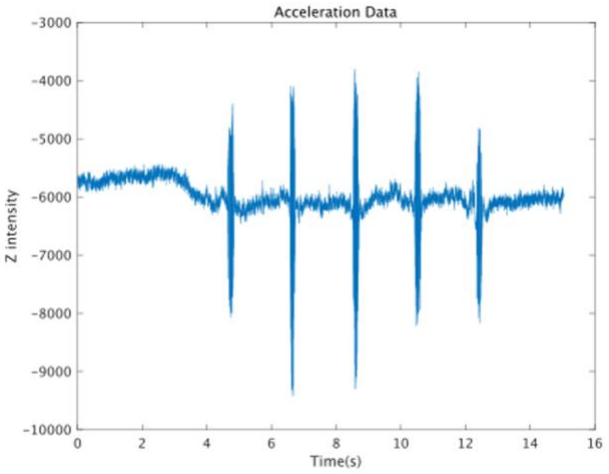

Fig.3 The plot of the acceleration data, collected when the participant vocalized each vowel once. The x-axis represents the time and the y-axis represents the acceleration data. From the left side, the order is /a/, /i/, /u/, /e/, and /o/.

3.2 Feature Extraction

Support Vector Machine needs a set of features to find an appropriate hyperplane that can discriminate classes. For the feature extraction from time-series data, it is a common method to transform time series data to frequency data [10,11,12]. Thus, we applied one of the techniques, Yule-Walker method, to the acceleration data we collected for the data transformation.

3.2.1 Signal Processing Technique

Many techniques that estimate the frequency value from the time-series data exist, such as Zero crossing method mentioned in [13], auto-correlation method [14], and Cepstrum method [15]. 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 [16], and is used for time-series feature extraction. Kallas et al. [17] 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 noted in [17].

When the Yule-Walker method is used, we have to set the input data, the order of the AR (auto-regressive) model used to generate the PSD estimate, and the number of samples to use in the Discrete Fourier Transform (DFT).

3.2.2 Procedure of Feature Extraction

Before performing PSD estimation, we preprocessed the acquired data. First, 0.3-second (about 300 samples) intervals of the acceleration data during each vowel utterance 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 the frequency values. For the use of the Yule-Walker method, we used the function "pyulear" given in MATLAB. This function requires the input signal, the arguments of the order of the AR (auto-regressive) model used to generate the PSD estimate, the number of samples to use in the Discrete Fourier Transform, and the sampling rate. We set the order as 10, and the number of samples as 2048. This procedure has been done for all the acceleration data.

3.2.3 Extracted Data

From the experiment procedure, we got the PSD output with 1025 samples. Figure 4 shows the plot of the output (vowel /a/) multiplied by 10log10 in the frequency domain.

To choose the efficient feature values from the data in the

frequency domain, we selected the frequency values of the first two peaks appearing on the plot (first and second higher harmonic waves). These frequency values were extracted from all data.

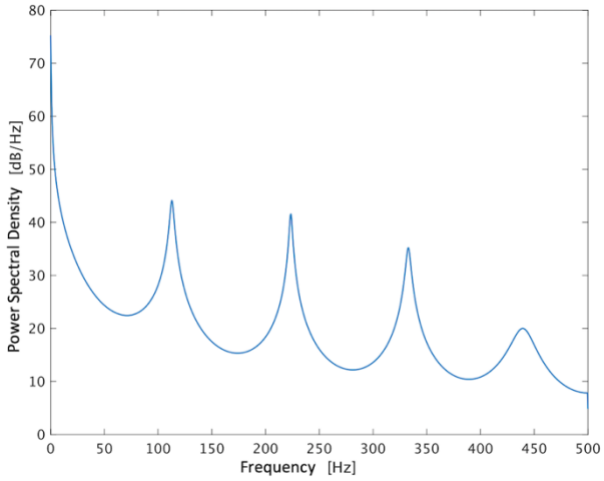

Fig.4 Power spectral density, which transformed from the acceleration data of the vowel /a/, in the frequency domain. The first two peaks are chosen as the feature values (at about 115 Hz, and 220 Hz for this plot).

3.3 Classification

Thus far, the acceleration data was collected by the sensor on the throat and was applied the spectral analysis method to extract the feature values. The feature values are the first and second higher harmonic waves. In this section, using the feature values, we verify whether the feature values can be biological information that discrimination each speech or not with SVM model.

3.3.1 Methodology

The total number of samples we collected was 50 samples (10 samples per vowel). Because the purpose of this study is to verify if the acceleration data can discriminate each vowel, we arranged the pattern of discrimination to the two-class classification between the vowel /a/ and the other vowel (/a/ vs. /i/, /a/ vs. /u/, /a/ vs. /e/, /a/ vs. /o/). Therefore, there are 4 patterns for classification in total.

With these pairs, the feature maps, which plot each vowel's feature values, are plotted in Figure 5. The x-axis (F1) and y-axis (F2) represent the first and second higher harmonic waves (feature values), respectively.

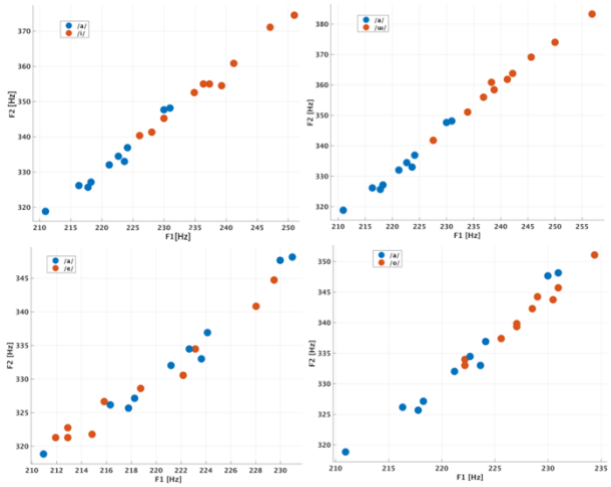

Fig. 5 Feature map with 2 class data set: /a/ vs. /i/ (upper left), /a/ vs. /u/ (upper right), /a/ vs. /e/ (bottom left), and /a/ vs. /o/ (bottom right). The blue plots represent the feature of /a/, and the orange plots represent the feature of the other.

Using the classifier, SVM, the discrimination accuracy was tested for each 2-class data set. We constructed SVM by using the Classification Learner app in MATLAB. For building the classifier, it requires data (feature values) set with the labels and the number of folds for the cross-validation. For the assignment of the labels, the data set was assigned with the numbers as follows: 1:/a/, 2:/i/, 3:/u/, 4:/e/, 5:/o/. In addition, the cross-validation was set with 5 folds.

3.4 Result

For each two-class classification, the discrimination accuracy calculated by the classifier (SVM) is shown in table 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 classification accuracy between /a/ and /e/.

First, the similarity of both frequencies might cause low accuracy. Rue et al. [18] investigated the frequency differences of vowels and Figure 6 shows the frequency distribution. As we can see in the figure, the position of vowels /a/ and /e/ is close compared to others; Their frequencies are similar. Hence, although we used the frequency data as the feature values in this study, the frequency data may not be the best choice for this pair classification. For improving the accuracy, we need to search for other feature values besides frequency that can discriminate /a/ and /e/ well.

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, which requires a few samples to train itself. Yet, as shown in Figure 5 (the feature map between /a/ and /e/), there are few outliers widely separated from the main cluster of each class. Therefore, with more samples, the SVM may be able to find an appropriate hyperplane that can separate the two classes better.

Table.1 Accuracy of Japanese vowel discrimination

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

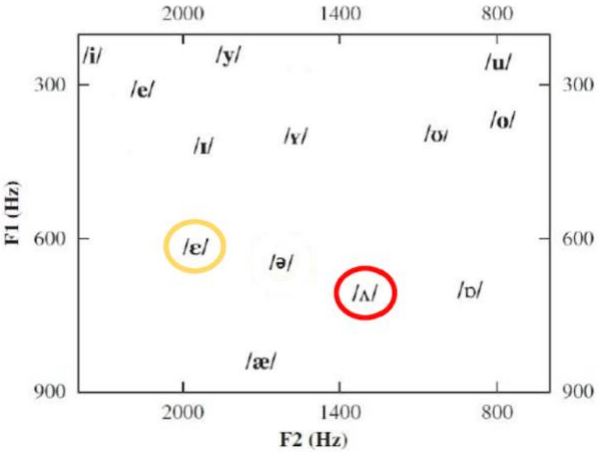

Fig.6 Frequency distribution of vowels [18]. The y-axis represents the first formant frequency, and the x-axis represents the second formant frequency. The part circled in red is the Japanese vowel /a/, the part circled in yellow is the Japanese

vowel /e/. The position of these two is close compared to others.

In this study, the vocal folds vibration data of Japanese vowels were measured from the throat using an acceleration sensor, and the first and second higher harmonics were extracted and used for the classification. Consequently, we could find the speech features in the data and the possibility of being able to discriminate vowels using the biological signal (the vibration of the vocal folds).

The main purpose of this study was to find new biological signal to develop new speech interfaces. As a result, we found a new biological signal, the vibration data of the vocal folds measured by the acceleration sensor, that could classify Japanese vowels. However, 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) data that associates with the speech process and has the potential to recognize consonants. This study is introduced in the following section.

4. Unvoiced Consonant Prediction from Pre-Speech EEG Data

Our research question was to find new biological signals that can be used for speech interface development. In the previous study, we focused on the vocal folds vibration, and verified the discrimination accuracy of Japanese vowels using the vibration data of the vocal folds measured by the acceleration sensor. As a result, the accuracy was 71 % on average. Thus, we found the possibility that the vocal folds vibration data can be used as new biological signals for the development of speech interfaces. 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. There have been already some studies on trying to recognize what the participant wants to say using EEG data [7,8]. However, most studies have not presented high accuracy yet, and none of those studies carried out the classification of consonants in particular. For this study, we measured EEG data while the participants vocalized specific 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 (artifact) comparatively, the classification of consonants has been conducted. For testing the classification accuracy, the Echo State Network (ESN) model, which is a subset of recurrent neural networks (RNN) and can deal with time-series data well was used.

4.1 Measurement

In this study, the data we want to measure was (pre-speech) EEG data. The participants were asked to wear an EEG head cap and vocalize some words. While they are vocalizing, we measured their EEG data. In addition to EEG data, we recorded the voice data and the trigger signals to calculate the speech onset.

4.1.1 Devices and Software for Experiment

For EEG measurement, EPOC X (Emotiv Inc., San Francisco, U.S.A.) was used. EPOC X has 14 channels from the 128 standard sites on the scalp, so that the brain activity at the Broca's area, which is known to process languages in the brain, can be measured. In addition, EPOC X measures EEG data with sampling frequency at 256 Hz, and its bandwidth is 0.16 – 43Hz and digital notch filters at 50Hz and 60Hz. We also recorded the voice data and trigger signals for detecting the speech onset. For the voice data, USB Microphone (Sanwa Supply Co.) was used. For the trigger signals, we built the program for the trigger signal to be sent every time the word prompt appears on the screen. The presentation of word prompts on the screen is created by PsychoPy 3 [19].

For collecting all data together, we used the software called LabRecorder [20], which enables multiple signals to be recorded at the same time with Lab Streaming Layer (LSL) protocol. For doing this, we set all data to be sent with LSL. For EEG data, we used the EMOTIV PRO function that allows the data to be sent with LSL protocol. For the voice data, we used a software called AudioCapture [21] that enables the voice data to be sent with LSL protocol. For the trigger signals, we used functions built-in PsychoPy and modified them for the purpose.

4.2 Experiment

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

Wearing the EEG cap, each one of the participants was seated approximately 40 cm from a 13-inch display monitor, and the microphone was located 20 cm apart from the participant's mouth as shown in Figure 7.

Fig.7 Experiment Figure: The participant wears the EEG cap. The microphone was set in front of the participant and the word appeared on the screen of the laptop.

Table 2 shows the word list that the participants vocalized during the EEG data measurement. This word list contains 25 different words, which have only one syllable and are 3-5 letters long. For the purpose of this study, five different kinds of words, starting with a specific letter (F,B,P,M,S), were selected.

Table.2 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 participant wears the EEG cap and speaks words displayed on the screen following the instruction. As shown in Figure 8, each trial starts with a target fixation cross in the middle of the screen that is presented for 2 s. The stimulus word then appears for 2 s. A speech cue follows in the form of "(((>", and "<)))" for 3 s. Participants are to speak the stimulus word following the speech cue. The cue is followed by the fixation cross of the next trial. The experiment runs in 250 trials per block, for two blocks.

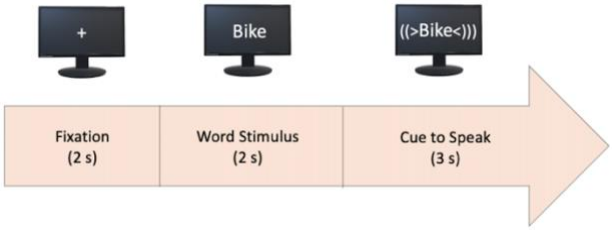

Fig.8 Experiment Procedure: each trial starts with a target fixation cross in the middle of the screen that is presented for 2 s. The stimulus word then appears for 2 s. A speech cue follows in the form of "(((>", and "<)))" for 3 s. Participants are to speak the stimulus word following the speech cue. The cue is followed by the fixation cross of the next trial. The experiment runs in 250 trials per block, for two blocks.

Before each recording session, the EEG measurement quality was calibrated by asking participants to look at up, down, left, and right and checking whether each corresponding EEG signal appeared on the real-time stream of EMOTIV PRO.

4.3 Preprocessing

In general, the raw EEG data is not used directly for analysis. Rather, the raw data is preprocessed to get the more efficient features for analysis. In this study, as we focus on the pre-speech EEG data, the EEG data were segmented into a certain time range and its amplitude was transformed into different amplitude range.

For the segmentation of EEG data, we used MATLAB (The MathWorks, Inc., Natick, MA) and the EEGLAB toolbox [22]. Before the segmentation, each speech onset of utterance was calculated and found by using the voice data and trigger signals. Using the function built-in EEGLAB, EEG data of each utterance was segmented into the range from -1000ms to 0ms respective to the speech onset. After the data were epoched, for removing the offset Baseline correction technique was applied. The baseline interval was selected from -1000ms to -500ms which is the first speech onset.

The features appearing in EEG data vary from person to person [23]. Therefore, the appropriate measurement location and analysis method should be chosen considering the fact. For this study, the amplitude of EEG data was scaled into the range from -1 to 1 by the Min-max scaling method. In addition, we applied High Pass Filter from 2 Hz to the EEG data. This is because the high-frequency EEG data have been observed for speech analysis [8,24] rather than the low-frequency EEG data, and the low-frequency EEG data is removed [7].

4.4 Classification

In this study, we aim to verify that pre-speech EEG data can be used to recognize the consonants using a machine learning model called Echo State Network (ESN) model. Here, we highlight the structure of ESN model we used for the classification of the consonants and how the hyper-parameters were adjusted based on the algorithms. Also, the result of the classification accuracy is observed.

4.4.1 Data Structure

In total, we collected 3500 samples of the pre-speech EEG data. Because of the five classes of consonant, there are 700 samples in each class. For this classification, we randomly assigned 90 % of samples as the training sample, and 10 % as the test sample. With the samples, we examined the consonant classification.

4.4.2 Echo State Network Model

For the consonant classifier, 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, ESN model dramatically reduces the computational complexity for training with a different training algorithm than that of a gradient-based learning algorithm that repeatedly updates the joint weights sequentially. However, in order to achieve high computational performance with such a method, the reservoir needs to be set appropriately. One of the most important settings is the hyperparameter setting.

4.4.3 Hyperparameter

In this study, we have tuned three main hyperparameters that have a significant impact on ESN. The first parameter is the input weight. Since the expressive ability of the activation function depends on the size of this parameter, it is necessary to choose a parameter that is appropriate for the input data. In general, the optimal value of Win is determined empirically. In this study, the values of the parameters were chosen by referring to the plots of the output of the activation function calculated by each node, in order to see if there were more features unique to each EEG of each consonant. This plot when the input weight is set as 0.01 is shown in Figure 9. Also, because another study [25] set the size of Win as the same as that of the recurrent weights, we set Win to [-1 1]. The plot with Win [-1 1] is shown in Figure 10.

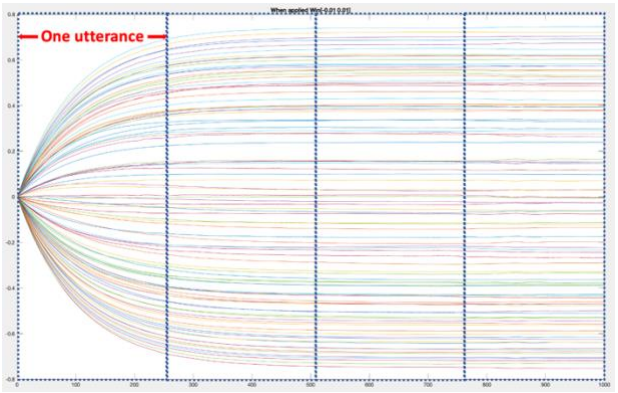

Fig.9 Part of the combined output of the EEG computed by the activation function of nodes during vocalization when the input weight is [-0.01 0.01] is plotted. The x-axis shows the number of samples, and the y-axis shows the output of the activation function. The area surrounded by dashes represents the EEG when a single word is uttered.

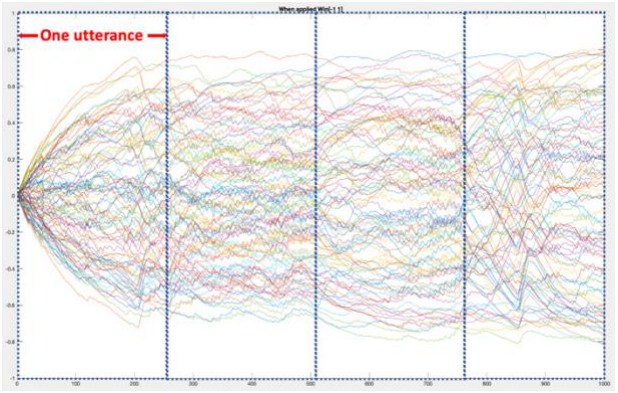

Fig.10 Part of the combined output of the EEG computed by the activation function of nodes during vocalization when the input weight is [-1 1] is plotted. The x-axis shows the number of samples, and the y-axis shows the output of the activation function. The area surrounded by dashes represents the EEG when a single word is uttered.

From the observation of these plots, we found that the output of the nodes when the weight is set to [-0.01 0.01] (Fig.9)

expresses no difference between the sections. On the other hand, the output of the node with the weight set to [-1 1] (fig.10) shows the characteristics of the waveform in each section. Furthermore, we learned that when the weight is set to more than 1, the waveform becomes too expressive and no difference can be seen in the vocal EEG of each word.

The second hyperparameter is the scaling parameter of the recurrent weight matrix W. It was set to 0.9 because the larger its value, the more information the node retains from the previous tense. The third parameter, the Leaky rate, was set to 0.009 by referring to the output plots of the nodes as well as Win and selecting a value that consistently produced high discrimination accuracy.

4.4.4 Evaluation Method

In a time-series classification task, a class label is given to a part or the whole of the time-series input data. In this case, one class label is 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. Since the target output is given as a one-hot vector, only the c-th output is expected to be close to 1 after training, while the other nodes are expected to be close to 0. Therefore, we first selected the element (numbers) that takes the maximum value among the elements of the output vector at each time, and then found the most frequent value of the element among the time range. If this element matches the class label then the classification is successful. Such a procedure was performed for each time range with the class label.

4.5 Result

For verifying whether the pre-speech EEG data can the consonant, we measured the pre-speech EEG data and verified the classification using ESN model. The confusion matrix that illustrates the classification accuracy is shown in Figure 9. From 1 to 5, the numbers represent the following consonant: F, B, P, M, S. The classification accuracy of 5 classes (consonants) was 30 %, which is a little above the random chance. However, as we can see in Figure 9, the predictions are scattered. The possible reason is the training algorithm we used. For aiming to reduce the computational complexity, we chose the simple linear method. By using a gradient-based learning algorithm with ESN model, the classification accuracy can be higher. In addition, though all channels of the EEG cap were used to measure EEG data in this study, to measure EEG data from specific areas where process speech may lead to even better accuracy. Moreover, additional preprocessing methods should be applied to EEG data for removing the noise.

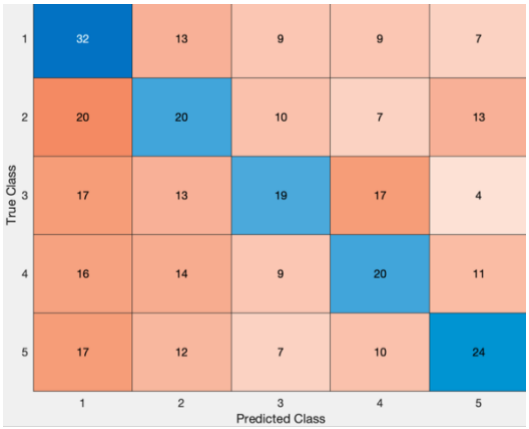

Fig.9 Confusion matrix: The x-axis is the true class, and the y-axis us the prediction classes. From 1 to 5, the numbers represent the following consonant, F, B, P, M, S. The predictions are scattered.

5. Conclusion

In this study, we have conducted two studies to find new biological signals that can be used for speech interface. First of them was the Japanese vowel classification using the vocal folds vibration data. We measured the acceleration data of the vocal folds vibration from the sensor attached to the throat and applied the spectral analysis to the measured data for extracting the feature values. We selected the first and second higher harmonic waves as the feature values that train the machine learning model, SVM. Because of the two-class classification and the small number of samples, we used SVM for the classification of Japanese vowels. The classifier outputted 71% accuracy on average. All classification got fairly high accuracy, except that of /a/ and /e/. This can be because both frequency values are similar, so that it is hard to discriminate them in the frequency domain. Therefore, the other feature values besides frequency 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 letter (F, B, P, M, S). The EEG data was epoched and its amplitude was transformed to a certain range. As we wanted to use this biological signal as speech interface, for the classifier we used ESN model, which processes fast and handles time-series data. The classification accuracy of 5 classes was 30 %, which is a little above the random chance. 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. By using a training algorithm of ESN such as a gradient-based learning algorithm, the accuracy can be better.

Through these two studies, we verified and found the possibility that the biological signals (the vocal folds vibration and the pre-speech EEG data) discriminate speech though there are still parts that should be improved. For future work, it may be a good method to use both two biological signals for speech recognition since the vocal folds vibration discriminates vowels and the EEG data discriminates consonants.

Bibliography

[1] J. J. Magee, M. Betke, J. Gips, M. R. Scott, and B. 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. [Online]. Available: https://www.researchgate.net/publication/224338425 A Human-Computer Interface Using Symmetry Between Eyes to Detect Gaze Direction

[2] M. J. Fagan, S. R. Ell, J. M. Gilbert, E. Sarrazin, and P. M. Chapman, “Develop- ment of a (silent) speech recognition system for patients following laryngectomy,” Medical Engineering and Physics, vol. 30, no. 4, pp. 419–425, 5 2008.

[3] R. J. Baken, “Electroglottography,” Journal of Voice, vol. 6, no. 2, pp. 98–110, 1 1992.

[4] R. T. Sataloff, S. Mandel, Y. D. Heman-Ackah, and M. A. p. o. o. Abaza, “Laryngeal electromyography.” [Online]. Available: https://books.google.com/books/about/Laryngeal Electromyography Third Edition.html?hl=ja&id=j-x6DwAAQBAJ

[5] H. Cai, J. Han, Y. Chen, X. Sha, Z. Wang, B. Hu, J. Yang, L. Feng, Z. Ding, Y. Chen, and J. Gutknecht, “A Pervasive Approach to EEG-Based Depression Detection,” Complexity, vol. 2018, 2018.

[6] A. Flinker, A. Korzeniewska, A. Y. Shestyuk, P. J. Franaszczuk, N. F. Dronkers, R. T. Knight, and N. 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, 3 2015. [Online].

Available: https://pubmed.ncbi.nlm.nih.gov/25730850/
[7] P. Ghane and G. Hossain, “Learning Patterns in Imaginary Vowels for an Intelligent Brain Computer Interface (BCI) Design,” 2020. [Online]. Available: http://arxiv.org/abs/2010.12066
[8] D. A. Moses, S. L. Metzger, J. R. Liu, G. K. Anumanchipalli, J. G. Makin, P. F. Sun, J. Chartier, M. E. Dougherty, P. M. Liu, G. M. Abrams, A. Tu-Chan, K. Ganguly, and E. F. Chang, “Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria,” New England Journal of Medicine, vol. 385, no. 3, pp. 217–227, 7 2021. [Online]. Available: https://www.nejm.org/doi/full/10.1056/NEJMoa2027540
[9] “ATRAdvanced Telecommunications Research Institute International.” [Online]. Available: https://www.atr.jp/about/atre.html
[10] 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, 4 2010.
[11] G. Toh and J. Park, “Review of vibration-based structural health monitoring using deep learning,” 3 2020.
[12] J. A. Gonzalez, L. A. Cheah, A. M. Gomez, P. D. Green, J. M. Gilbert, S. R. Ell, R. K. Moore, and E. 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, 12 2017.
[13] {Toledo-Perez2020} D. C. Toledo-Perez, J. Rodriguez-Resendiz, and R. A. GomezLoenzo, “A study of computing zero crossing methods and an improved proposal for EMG signals,” IEEE Access, vol. 8, pp. 8783–8790, 2020.
[14] {Rabiner1977} L. 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.
[15] {Noll2005} A. M. Noll, “Cepstrum Pitch Determination,” The Journal of the Acoustical Society of America, vol. 41, no. 2, p. 293, 7 2005. [Online]. Available: https://asa.scitation.org/doi/abs/10.1121/1.1910339
[16] {Chevalier1985} 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.
[17] {Kallas2012ModelingMachines} M. Kallas, P. Honeine, C. Richard, C. Francis, and H. 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.[Online]. Available: https://www.researchgate.net/publication/317381593 PREDICTION OF TIME SERIES USING YULE-WALKER EQUATIONS WITH KERNELS
[18] “(PDF) Directional asymmetries in vowel perception.” [Online]. Available: https://www.researchgate.net/publication/305422406 Directional asymmetries in vowel perception
[19] {Peirce2019PsychoPy2:Easy} J. Peirce, J. R. Gray, S. Simpson, M. MacAskill, R. H¨ochenberger, H. Sogo, E. Kastman, and J. K. Lindeløv, “PsychoPy2: Experiments in behavior made easy,” Behavior Research Methods, vol. 51, no. 1, pp. 195–203, 2 2019. [Online]. Available: https://link.springer.com/article/10.3758/s13428-018-01193-y
[20]{Labstreaminglayer/App-LabRecorder:Format.} “labstreaminglayer/App-LabRecorder: An application for streaming one or more LSL streams to disk in XDF file format.” [Online]. Available: https://github.com/labstreaminglayer/App-LabRecorder
[21]{Labstreaminglayer/App-AudioCapture:LabStreamingLayer} “labstreaminglayer/App-AudioCapture: Capture audio and stream it over LabStreamingLayer.” [Online]. Available: https://github.com/labstreaminglayer/ App-AudioCapture
[22]{Delorme2004EEGLAB:Analysis} A. Delorme and S. 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. [Online]. Available: http://www.sccn.ucsd.edu/eeglab/
[23] {ItoProposalEEG}S.-i. Ito, Y. Mitsukura, M. Fukumi, and N. Akamtsu, “Proposal of the EEG Analysis Method Using the Individual Characteristic of the EEG.”
[24] {Mersov2016a} A. M. Mersov, C. Jobst, D. O. Cheyne, and L. De Nil, “Sensorimotor Oscillations Prior to Speech Onset Reflect Altered Motor Networks in Adults Who Stutter,” Frontiers in Human Neuroscience, vol. 10, no. SEP2016, 9 2016. [Online]. Available: /pmc/articles/PMC5009120//pmc/articles/PMC5009120/ ?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5009120/
[25] G. Tanaka, , R. Nakane, A. Hirose, , and , “Rizab¯a konpyu¯tingu : jikeiretsu pata¯n ninshiki no tame no ko¯soku kikai gakushu¯ no riron to ha¯dowea.” [Online]. Available: https://www.morikita.co.jp

本研究に関する研究業績

1) The Best Poster Award, Distributed Processing System Society Workshop (DPSWS), November 2020