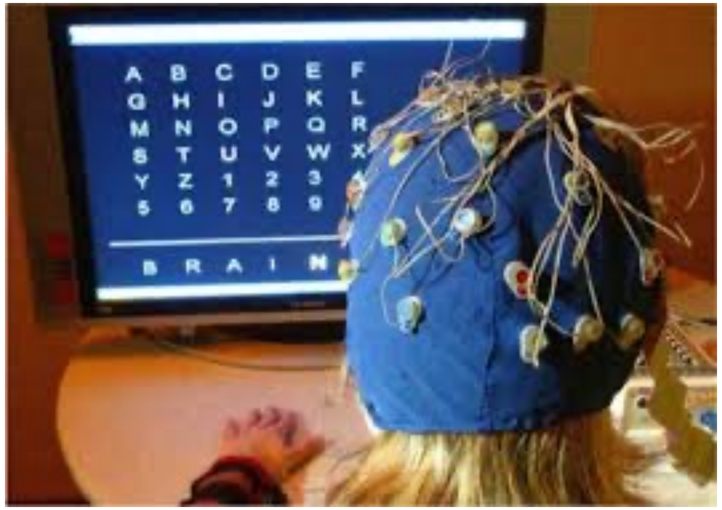
1D01

Speech Feature Analysis and Discrimination in Biological Information

Shogo Honda

BCI(Brain computer Interface)



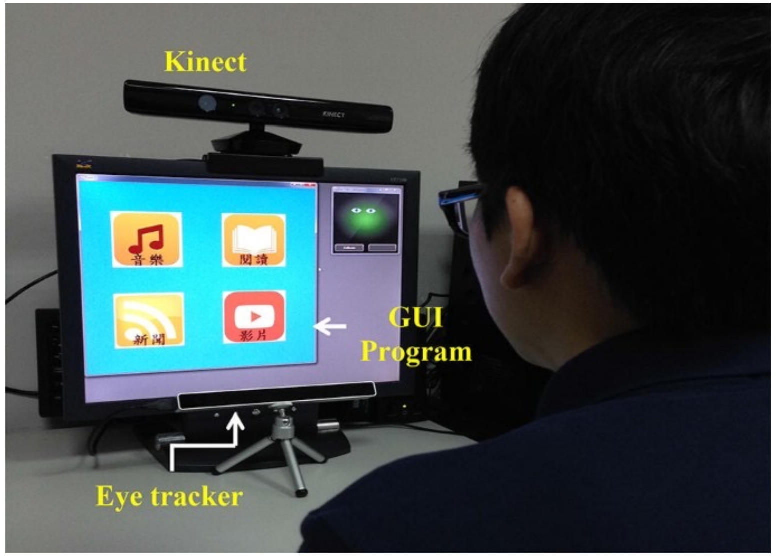
Lip reading

『Speech Interface』

for speech disorder

(EX) ALS, tongue cancer

Eye Tracking



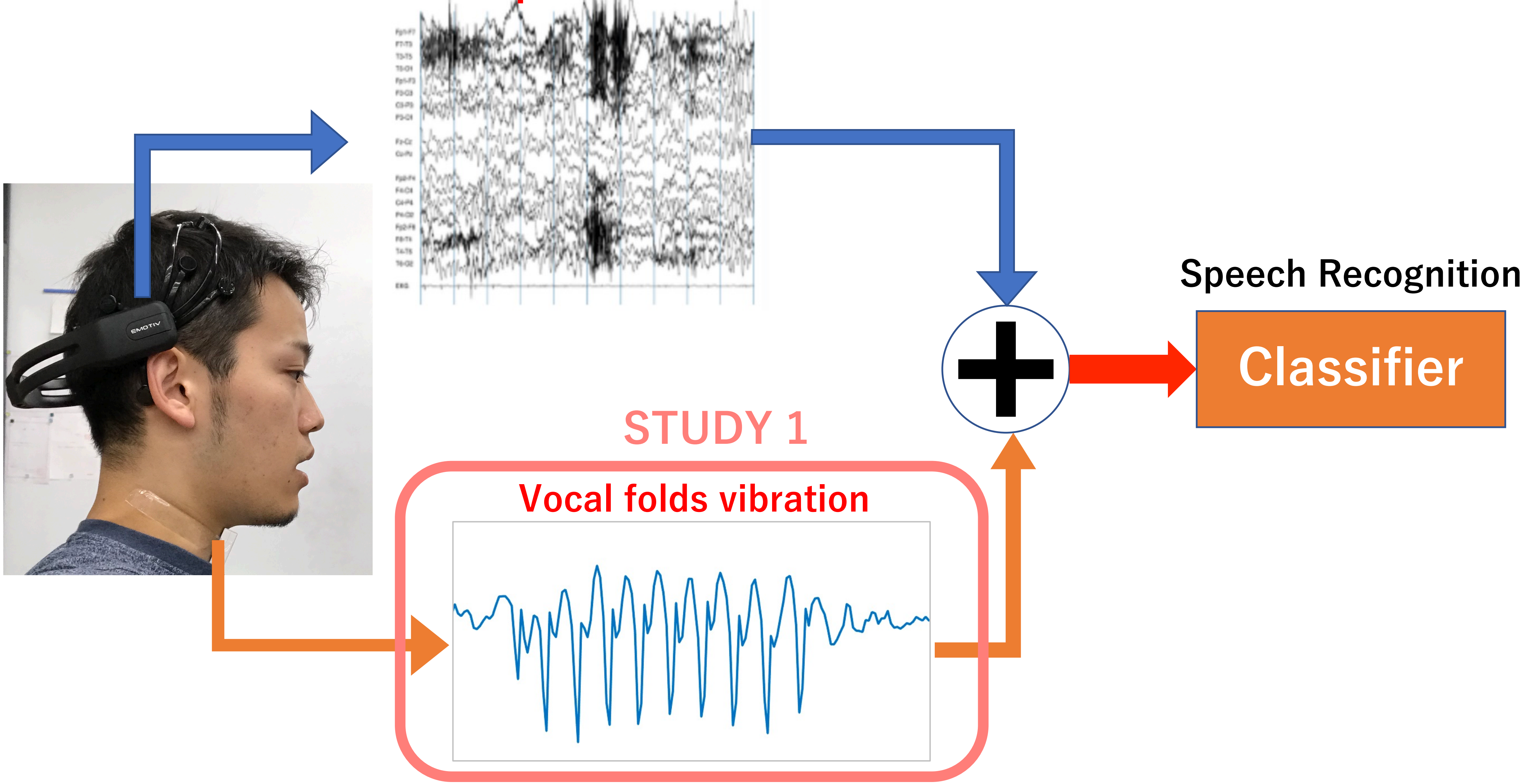
Nearly **40 years** of technological improvement using these specific pieces of information

→ Why don't we **rethink the biological signals** used?

In this study…

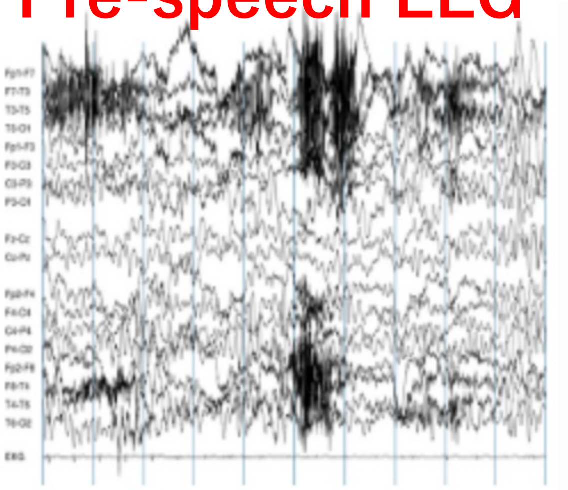
Finding new biological information for speech interface

Objectives


Pre-speech EEG
Speech Recognition
Classifier
STUDY 1
Vocal folds vibration

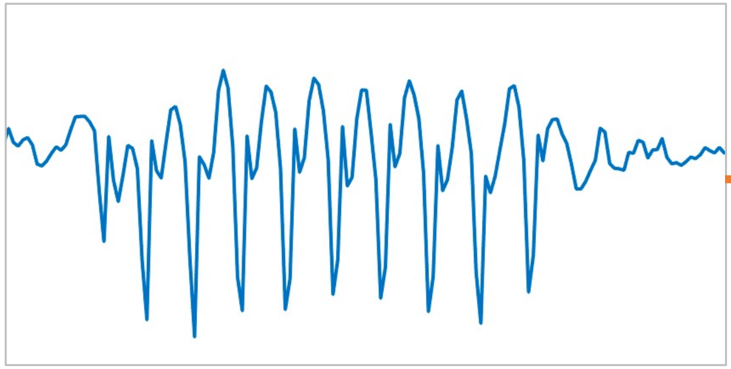
Objectives

Pre-speech EEG



STUDY 1

Vocal folds vibration



Classifier

Speech Recognition

Agenda

1. STUDY ONE (Vocal folds vibration)

"Japanese Vowel Discrimination by Throat Vibration"

2. STUDY TWO (Pre-speech EEG)

"Unvoiced Consonant Prediction from Pre-Speech EEG Data"

3. Conclusion

4. Future Work

Related studies on vocal folds vibration

1. Electroglottograph(EGG)

- Measures the degree of contact between the vocal folds
- Able to distinguish between natural voice and back voice[1]

[1] A. Mayr, "Parameters of Flow Glottogram and EGG for Vocal Registers-Modal, Falsetto and voce faringea."

2. Electromyography(EMG)

- Measure muscle cell movement
- Used in studies to assess muscle condition during swallowing[2]

[2] Cagla Kantarcigil et al. "Validation of a Novel Wearable Electromyography Patch for Monitoring Submental Muscle Activity During Swallowing"

No research has focused on the use of vocal folds vibration for speech recognition.

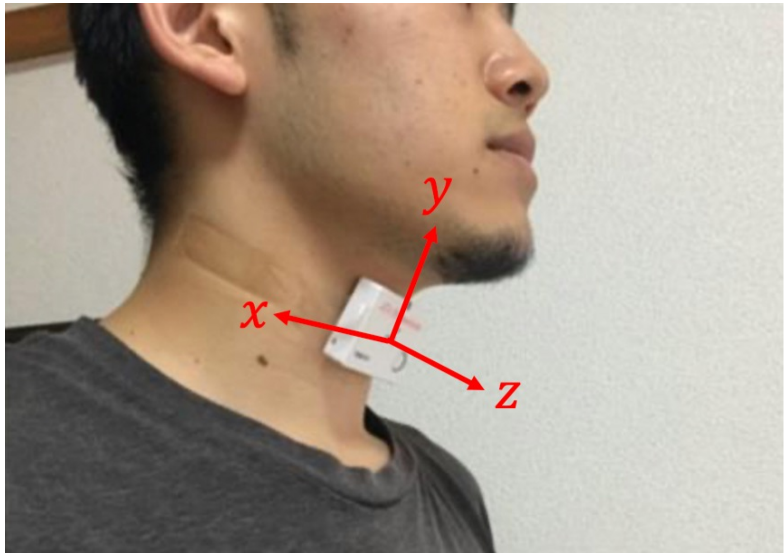
Measurement

Device:
Multifunctional sensor TSND121

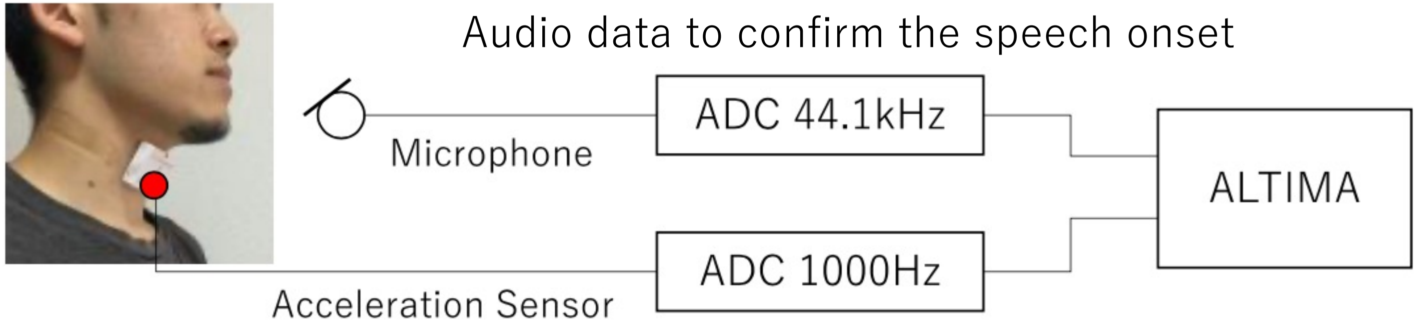
Attach to the throat

Use the built-in 3-axis acceleration sensor (Z-axis)

Attach to the position of the larynx where the vocal cords are located.



Measurement setup



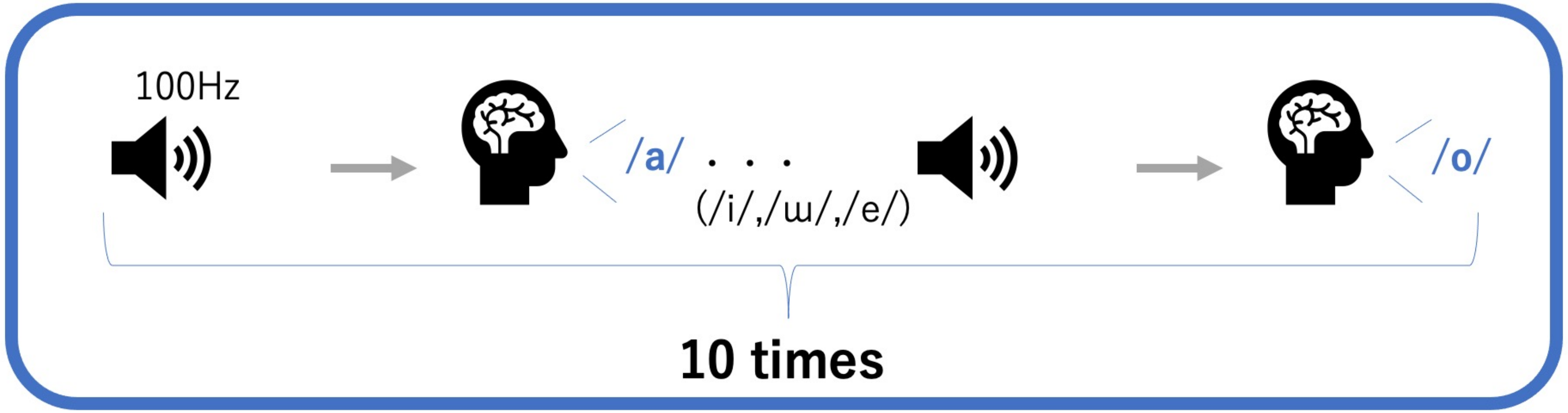
*ALTIMA:
Dedicated software for TSND121
- Collects acceleration data and audio data

Measurement Procedure

Number of subjects: 1
Voice contents: Japanese vowels
Number of repetitions: 10 times

Japanese vowels

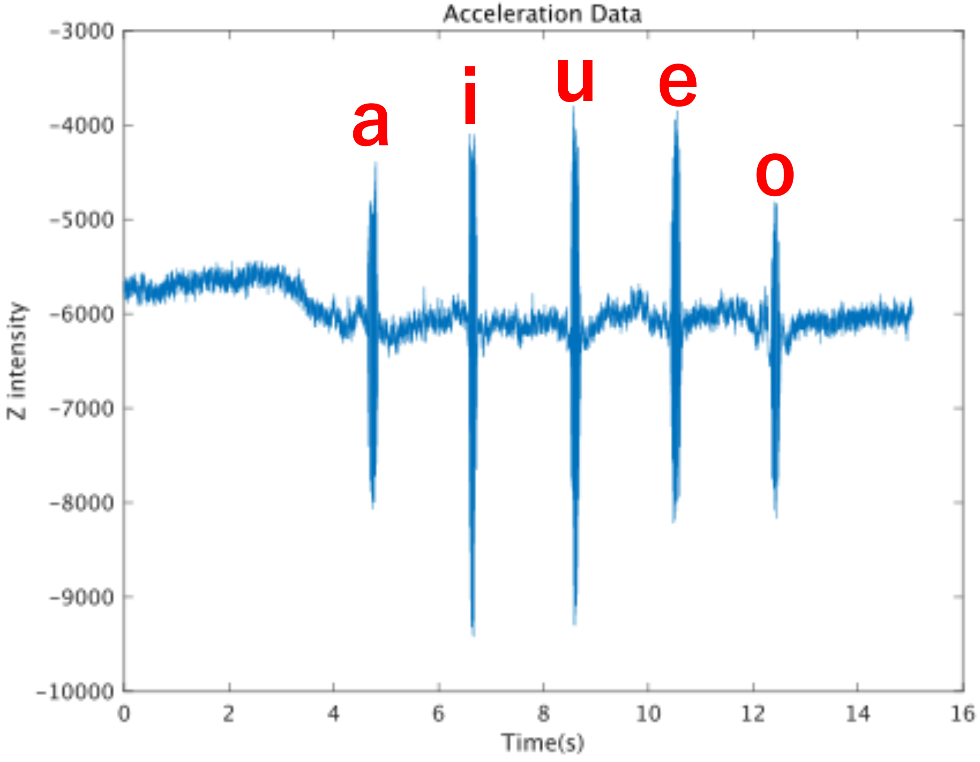
あ　い　う　え　お
/a/, /i/, /u/,/e/, /o/



The subject listened to the tone at 100 Hz before speaking.

Measured vibration data

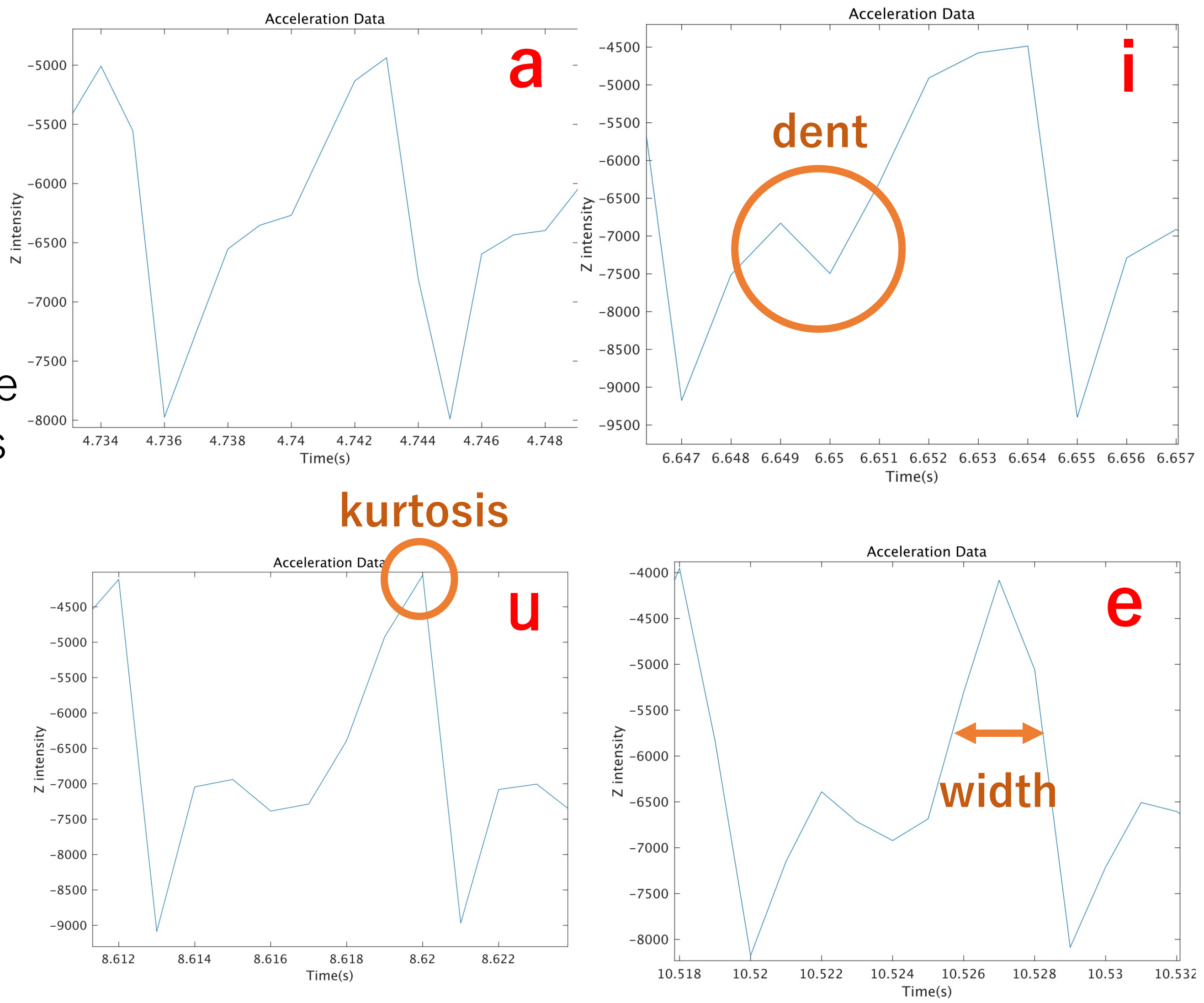
Left: acceleration data for one cycle
Right: magnified waveform of each period



Acceleration data (1 cycle)

Range 0.05s

No clear differences/characteristics of each vowel…



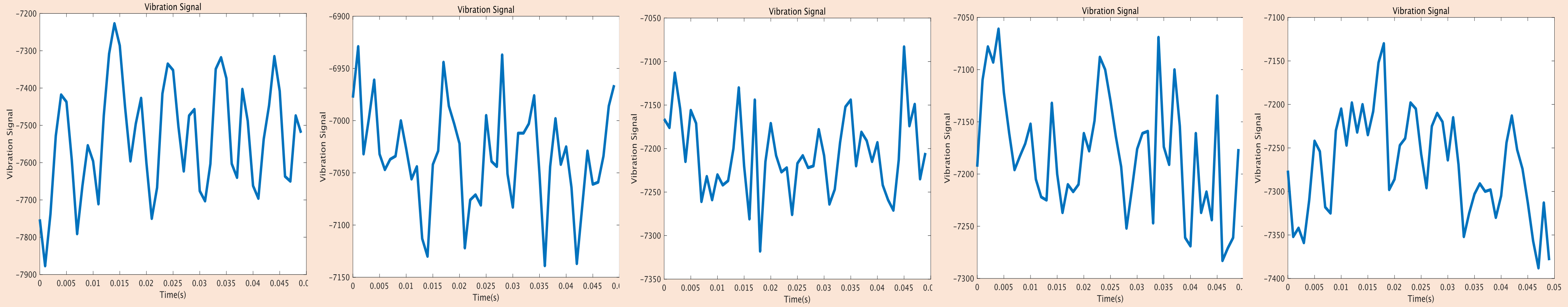
Magnified acceleration data (1 period)

Feature Extraction .. Converting each vowel into data to represent its characteristics

- **Time-series → Frequency**

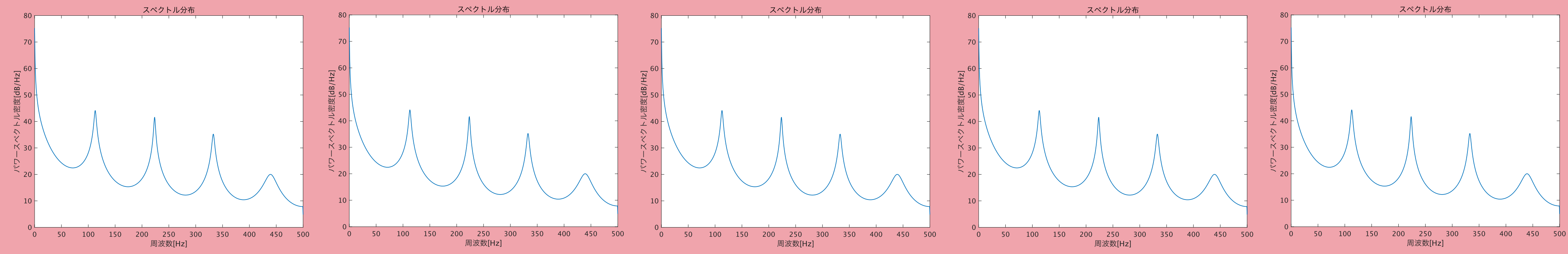
1. Cut off 0.3s from each vibration
2. Apply Hamming window with 300 samples
3. Yule-Walker method (order=10,nfft=2048)

/a/ /i/ /u/ /e/ /o/



Spectral Analysis
$(\times 10 \log 10)$

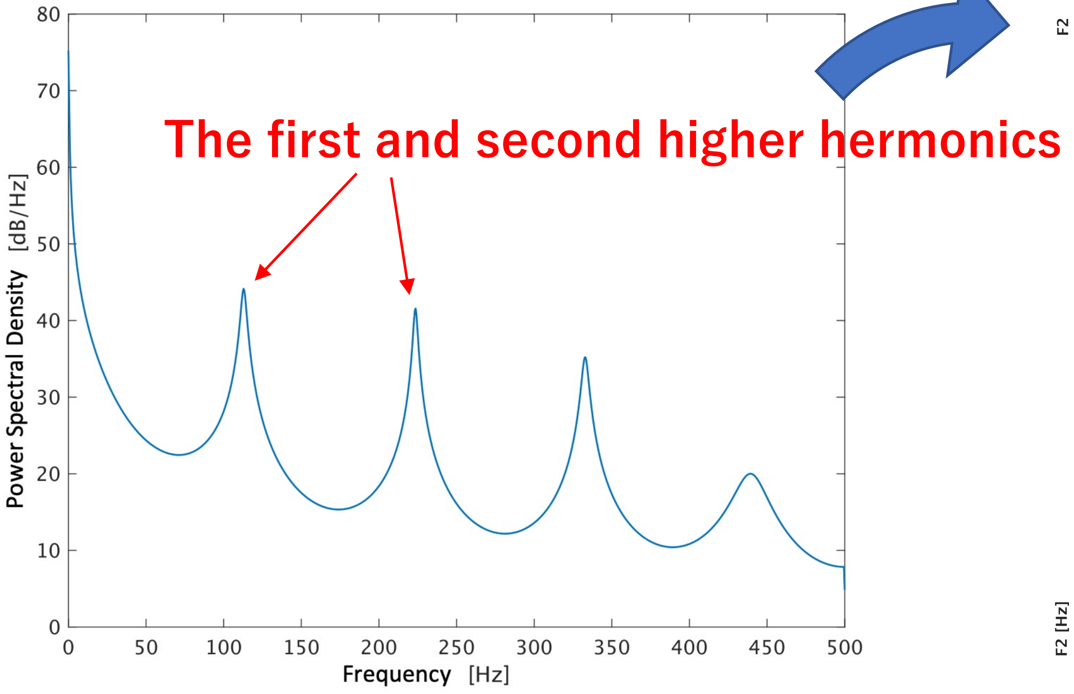
Spectral density distribution



Feature Extraction .. Converting each vowel into data to represent its characteristics

- Time-series → Frequency
- Feature Selection
- Feature plot

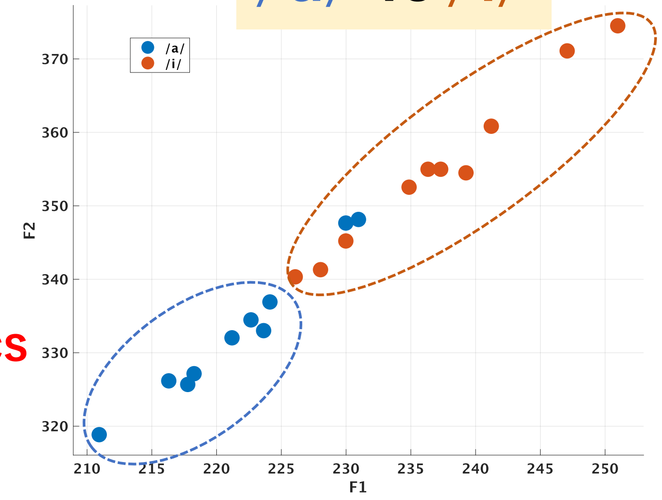
Spectral density distribution



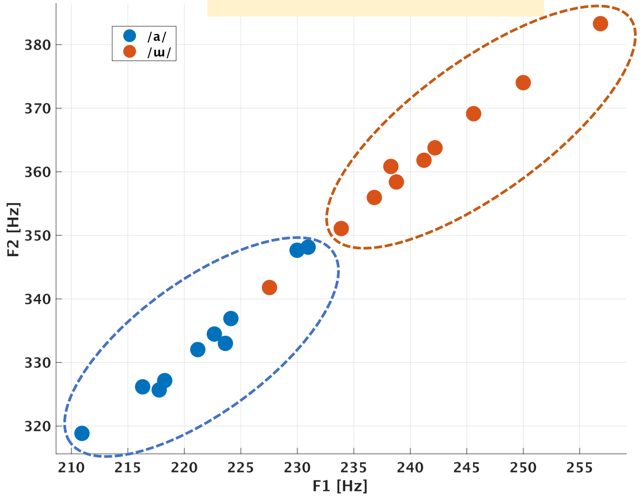
Feature Plots

Vowel Discrimination by Vocal Cord Vibration Plotted in Two Classes

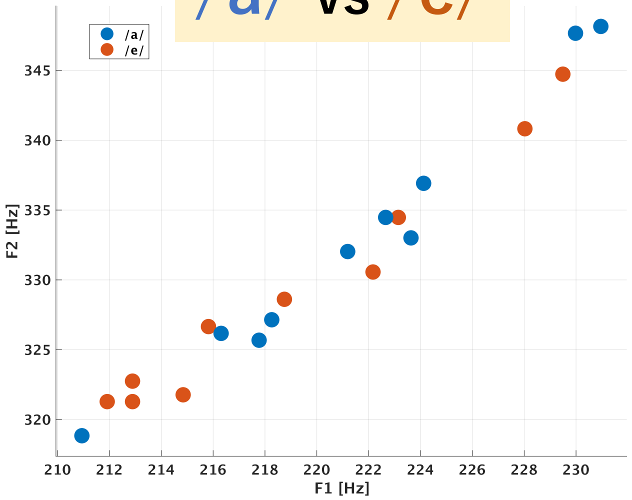
/a/ vs /i/



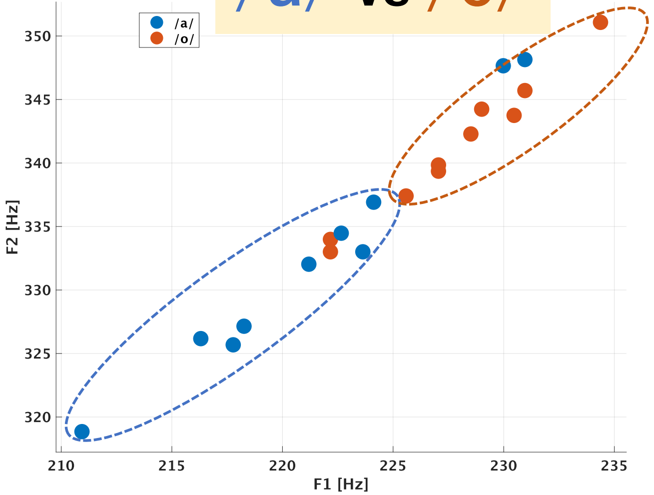
/a/ vs /u/



/a/ vs /e/

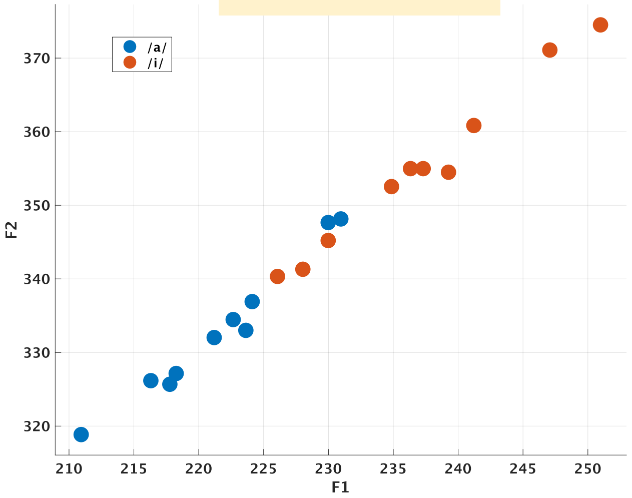


/a/ vs /o/

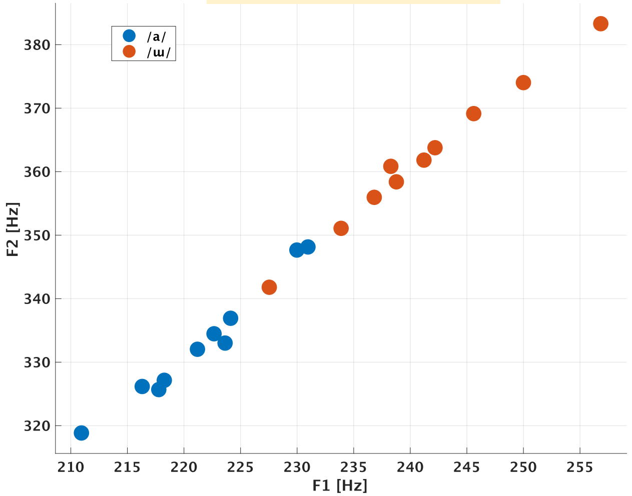


Vowel Classification

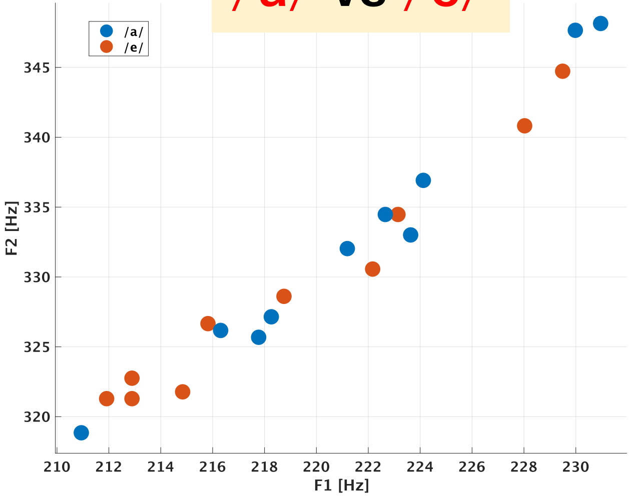
/a/ vs /i/



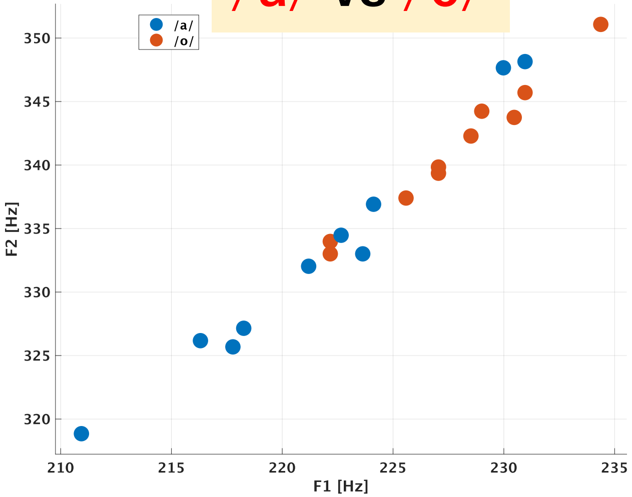
/a/ vs /u/



/a/ vs /e/



/a/ vs /o/



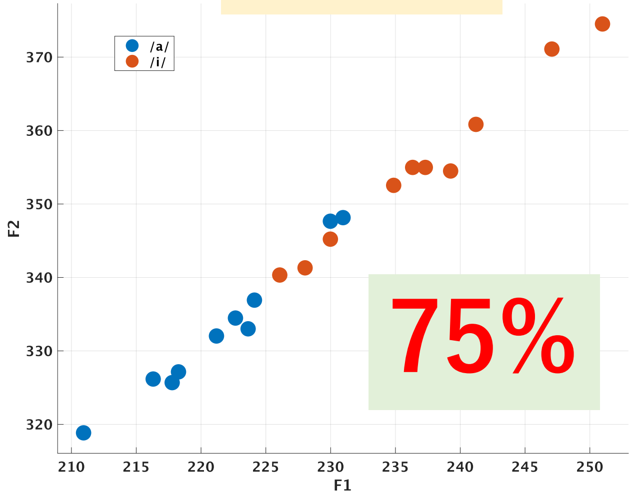
2-class linear SVM
MATLAB classification Learner App

Number of samples：50
Number of class：2
Cross validation：5 folds cross

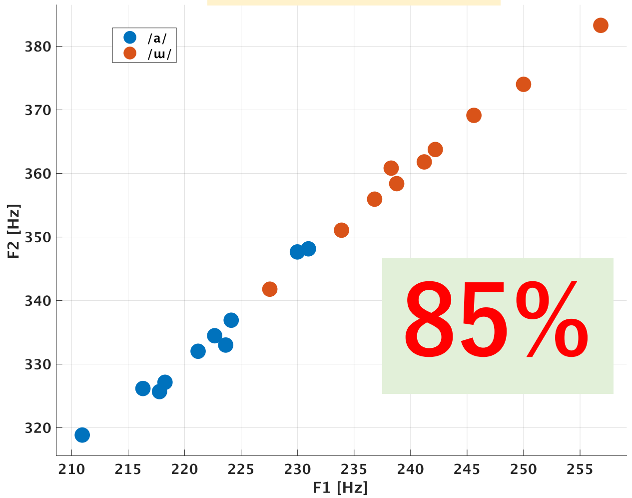
Vowel Classification

Discrimination accuracy：71% on average

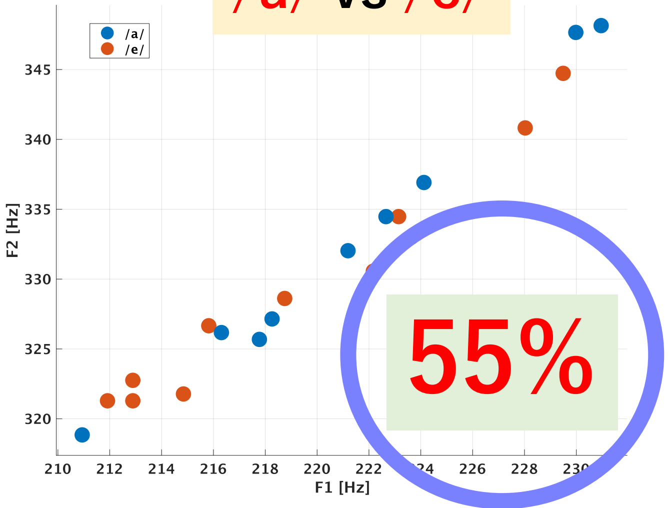
/a/ vs /i/



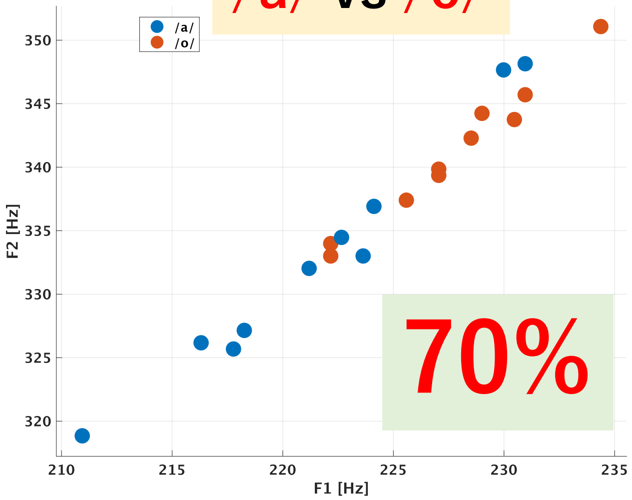
/a/ vs /u/



/a/ vs /e/



/a/ vs /o/



2-class linear SVM
MATLAB classification Learner App

Number of samples：50
Number of class：2
Cross validation：5 folds cross

Discussion

Achievement

No speech recognition by vocal folds vibration

- Recorded 71% vowel classification accuracy
- Indicated the possibility as new biological signals

Improvement

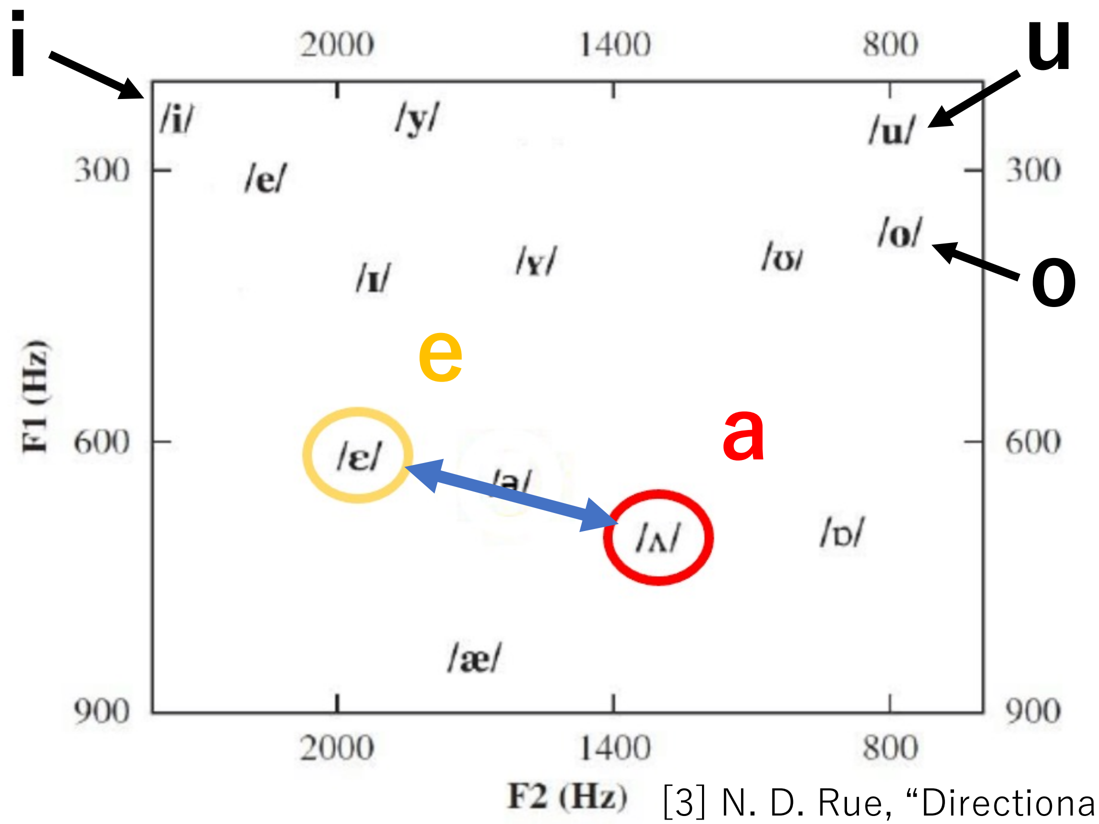
Low discrimination accuracy between /a/ and /e/

- Similarity of frequencies → Another feature value
- Small number of samples used for training

Discussion

Similarity of the first and second formants of vowels [3]

→ No significant difference in frequency [Solution]: Another feature



[3] N. D. Rue, “Directional asymmetries in vowel perception.”

Discussion

Achievement

No speech recognition by vocal folds vibration

- Recorded 71% vowel classification accuracy
- Indicated the possibility as new biological signals

Improvement

Low discrimination accuracy between /a/ and /e/

- Similarity of frequencies → Another feature value
- Small number of samples (1 subject) → More data collection

Discussion

Words are a combination of vowels and consonants

Study /ˈstʌdi/

Problem

Consonant recognition by vocal folds vibration is challenging.

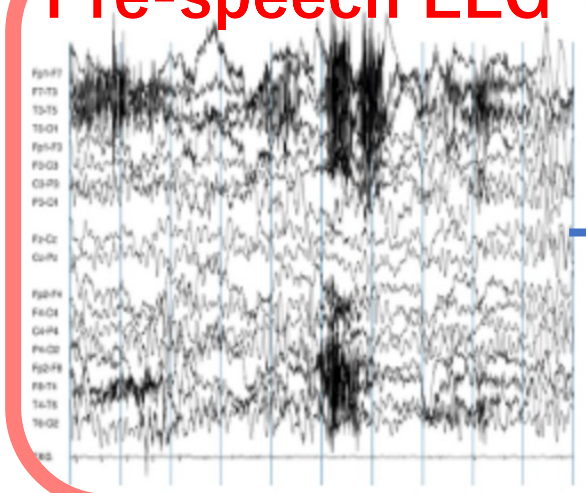
Next step

Need to find other biological signals that can classify consonants

Pre-speech EEG data

Objective

Pre-speech EEG



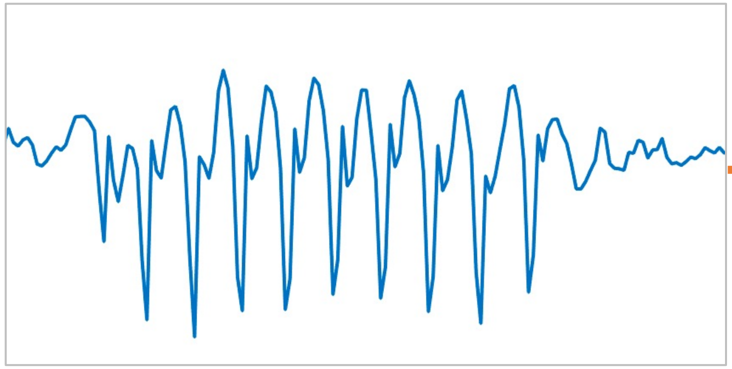
Classifier

Consonant recognition

STUDY 2

STUDY 1 ✅

Vocal folds vibration



Classifier

Vowel recognition

Speech-related studies on EEG

Ghane et al. [4]

- Measured EEG while the subject is **imaging** vowels
- Classified imaged **vowels** by SVM
- Classification accuracy was **76.7%**

[4] Ghane et al. “Learning Patterns in Imaginary Vowels for an Intelligent Brain Computer Interface (BCI) Design “

Moses et al. [5]

- Measuring **invasive EEG** during **vocalization**
- Classified the uttered words
- Classification accuracy was **47.1%**

[5] Moses et al. "Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria”

<Literature>	<Our study>
No vocalize	Vocalize
Vowel	Consonant
Invasive	Non-invasive
EEG during speech	Pre-speech EEG

Capable of capturing speech features by EEG

Measurement

Measured data and devices

Data [sampling rate]	Device/software
EEG signal [256Hz]	EPOC X (Emotiv Inc.)
Audio signal [44.1kHz]	USB microphone (Sanwa Supply Co.)
Trigger signal	PsychoPy 3

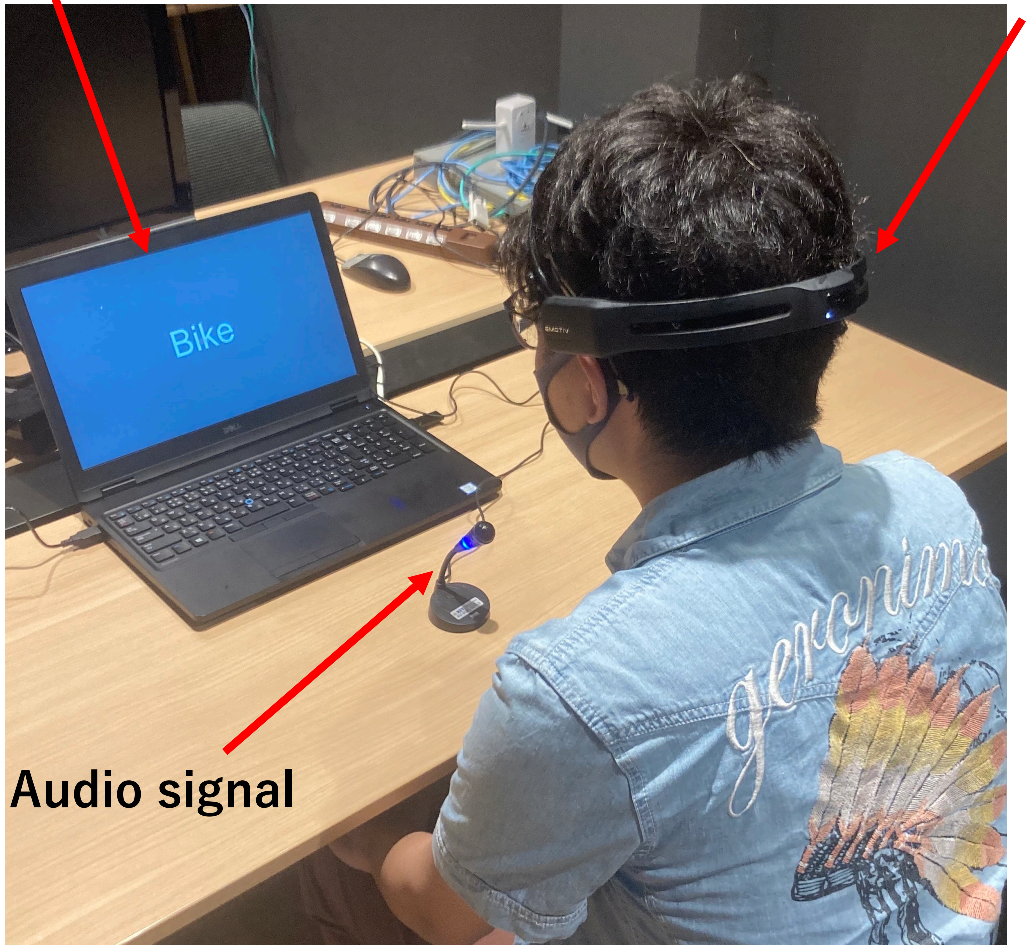
* These signals were measured simultaneously by LabRecorder

List of word prompts

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

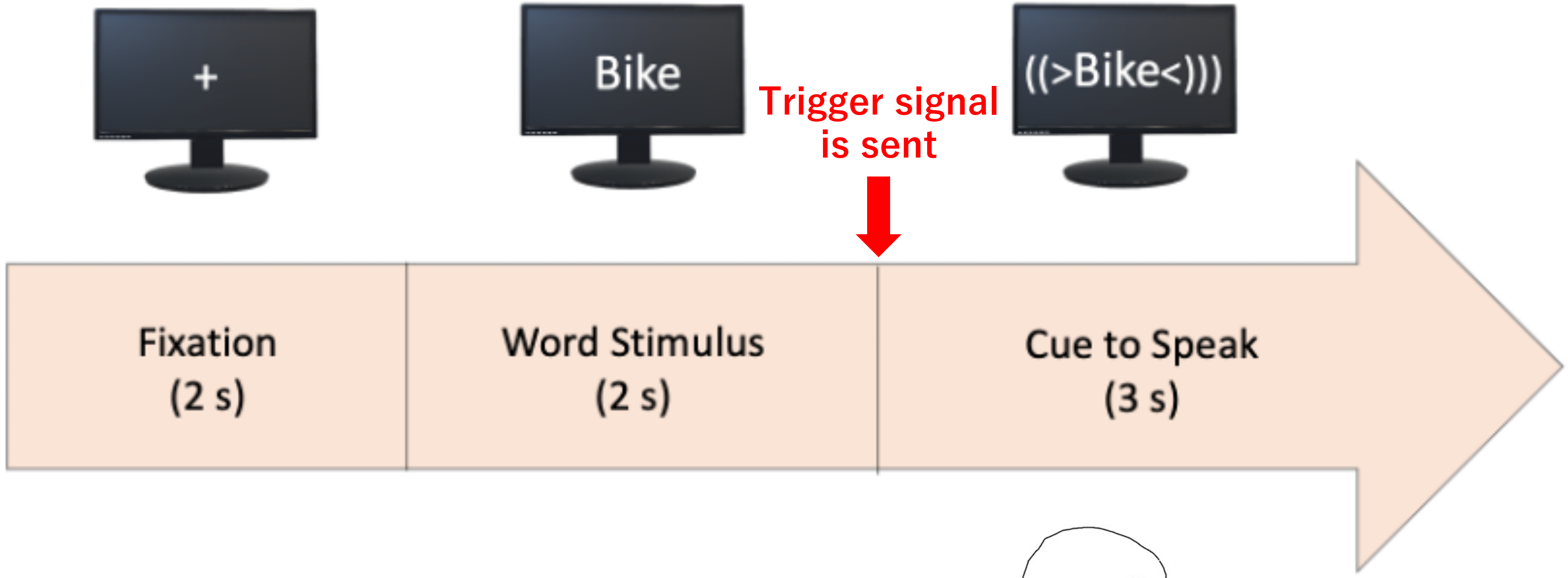
Word prompt
(+trigger signal)

EEG signal

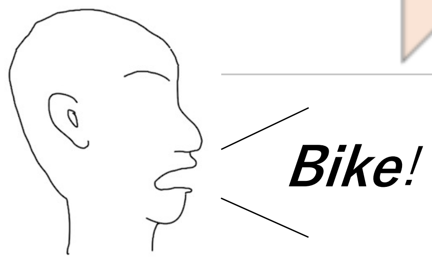


Audio signal

Measurement Procedure



Subject　7 people
Word content　25 words
Repeat　250 times x2

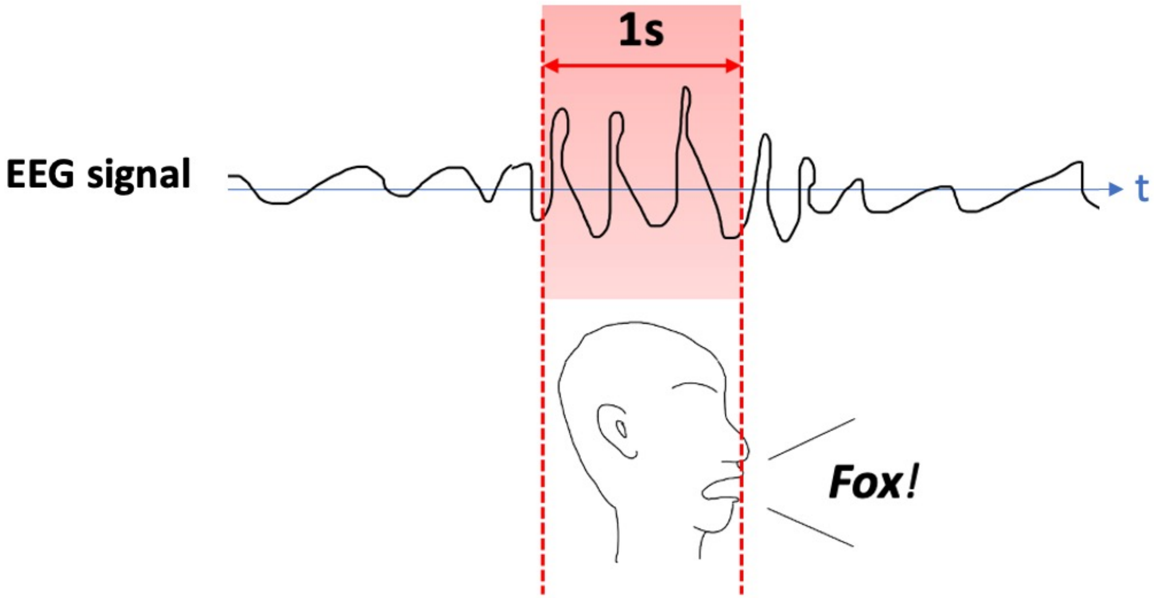


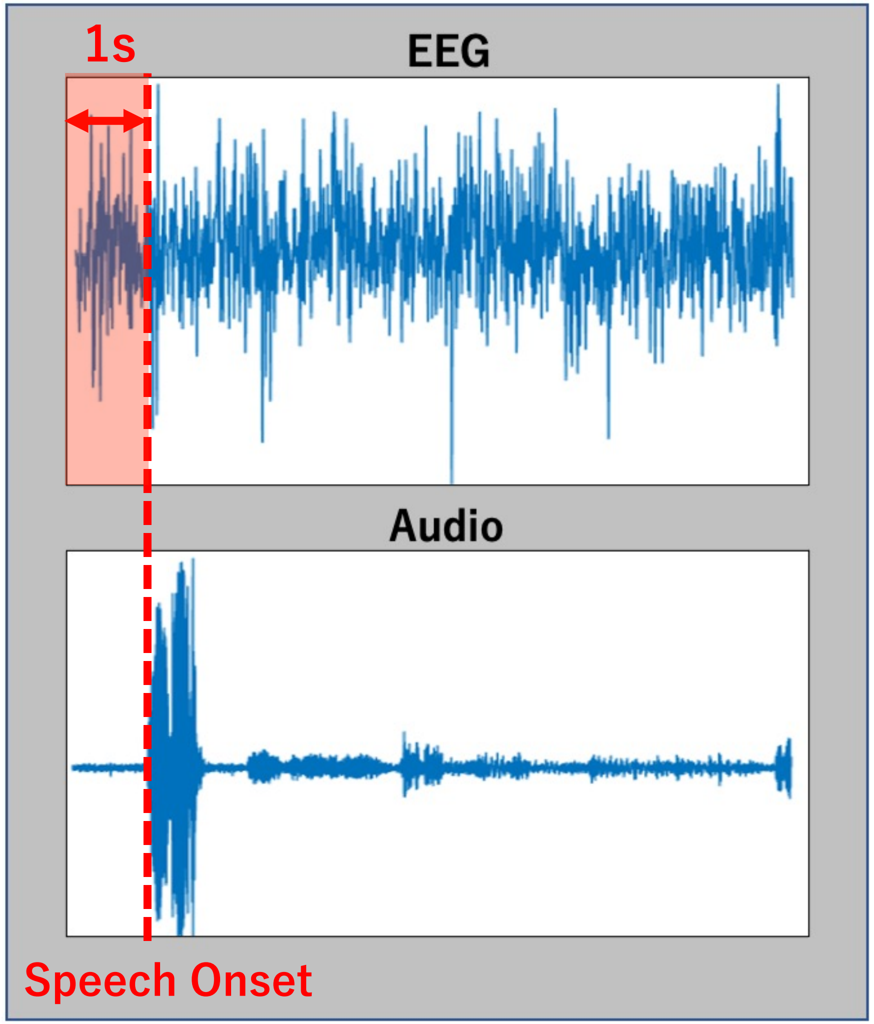
*1 To check the quality of the EEG measurement, calibration was performed for each experiment.
*2 The subjects were asked to practice pronunciation with a native speaker before the experiment.

Preprocessing

MATLAB and EEGLAB were used for preprocessing

- **Epoch**･･･Take pre-speech EEG (-1s~0s)





Preprocessing

MATLAB and EEGLAB were used for preprocessing

- **Epoch**…Take pre-speech EEG (-1s~0s)
- **High-Pass filter** (2Hz)…Remove low-frequency noise
  - In Ghane et al.[4], they took Band-Pass filter at 2~40Hz
  - Gamma waves (35Hz~) were observed in Moses et al.[5]

[4] Ghane et al. "Learning Patterns in Imaginary Vowels for an Intelligent Brain Computer Interface (BCI) Design "
[5] Moses et al. "Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria"

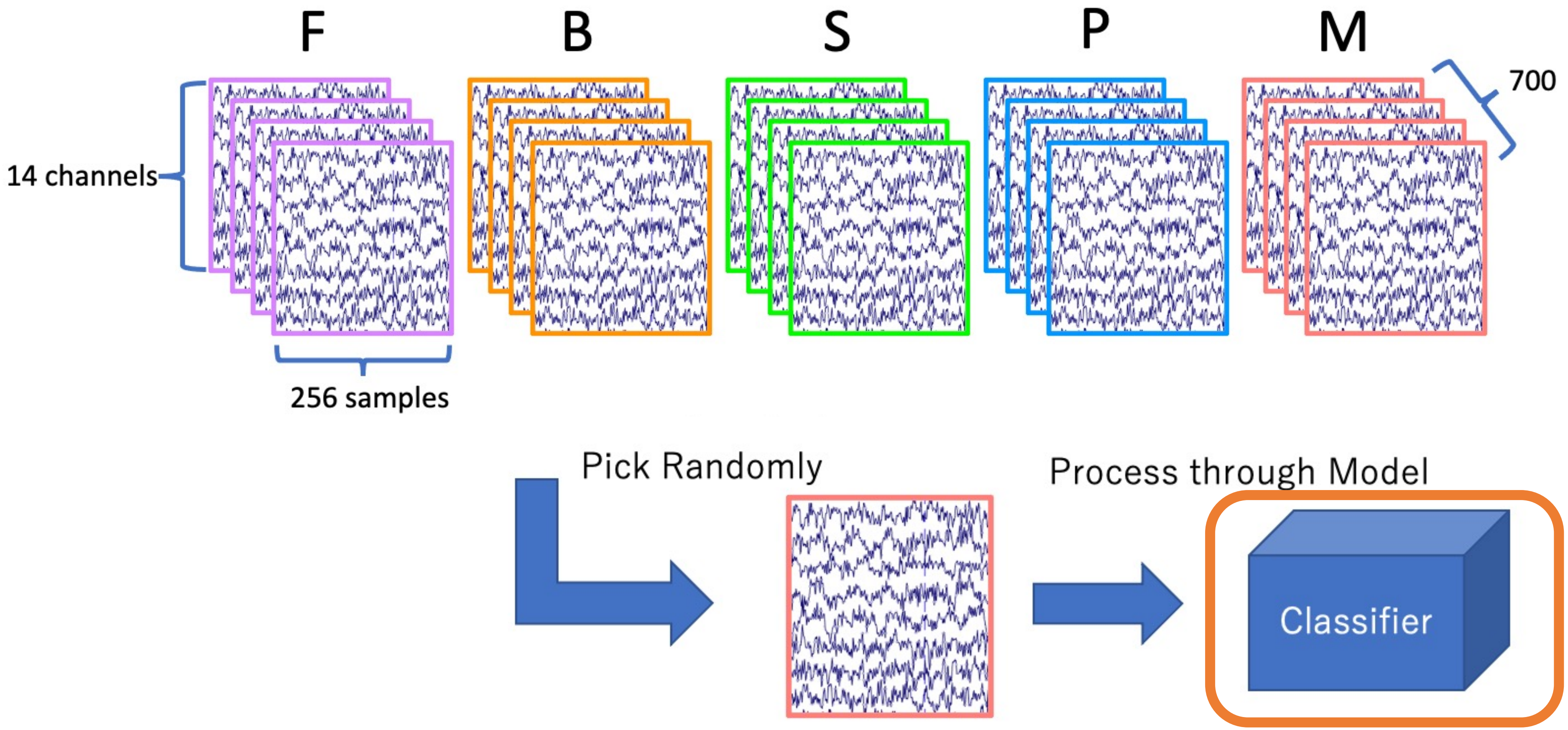
Preprocessing

MATLAB and EEGLAB were used for preprocessing

- **Epoch**…Take pre-speech EEG (-1s~0s)
- **High-Pass filter** (2Hz)…Remove low-frequency noise
  - In Ghane et al.[4], they took Band-Pass filter at 2~40Hz
  - Gamma waves (35Hz~) were observed in Moses et al.[5]
- **Min-max scaling**(-1~+1)… Keep the noise and brain wave differences between each subject within a certain range
- **Baseline** (-500ms~0ms)… EEG voltage offset adjustment

[4] Ghane et al. "Learning Patterns in Imaginary Vowels for an Intelligent Brain Computer Interface (BCI) Design "
[5] Moses et al. "Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria"

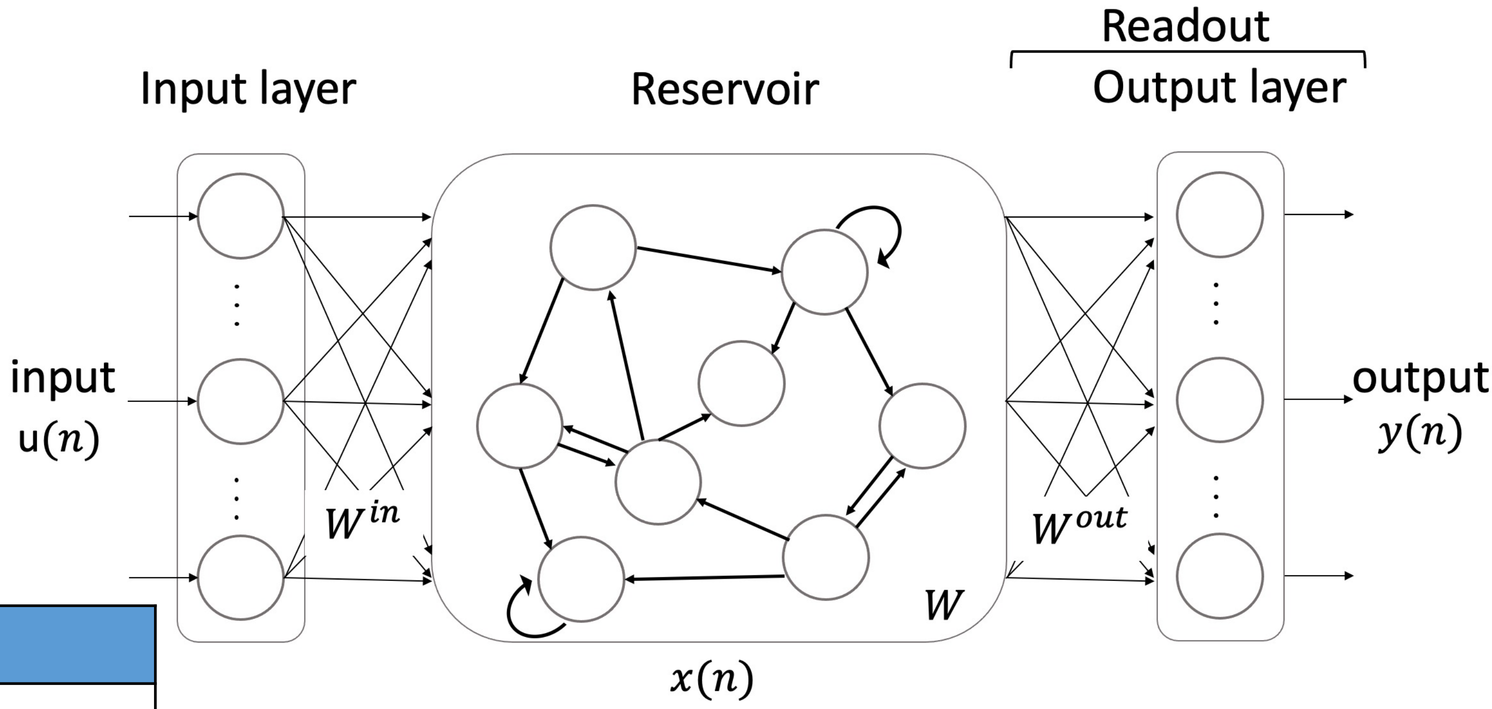
Data structure after preprocessing



Model for Consonant Classification

Echo State Network (ESN)

1. The kind of RNN model
2. Process time-series data
3. Reduce computational complexity
4. Many fixed parameter settings



Parameter	Meaning
N_u	Number of input layer nodes
N_x	Number of reservoir layer nodes
N_y	Number of output layer nodes
W^{in}	Input connectivity weight matrix
W	Recurrent connectivity weight matrix in the reservoir
α	Leaky rate

$$x(n+1) = f(W^{in}u(n+1) + Wx(n))$$

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

*f denotes the activation function.
In this study, the $\tanh$ function is used

ESN model for Consonant Classification

ESN parameter settings

Parameter	Meaning	Set
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
ρ	Spectral radius of W	0.9
W^{in}	Input connectivity weight matrix	
α	Leaky rate	

Training model

Linear regression model

Sample usage ratio

90% (train), **10%** (test)

ESN model for Consonant Classification

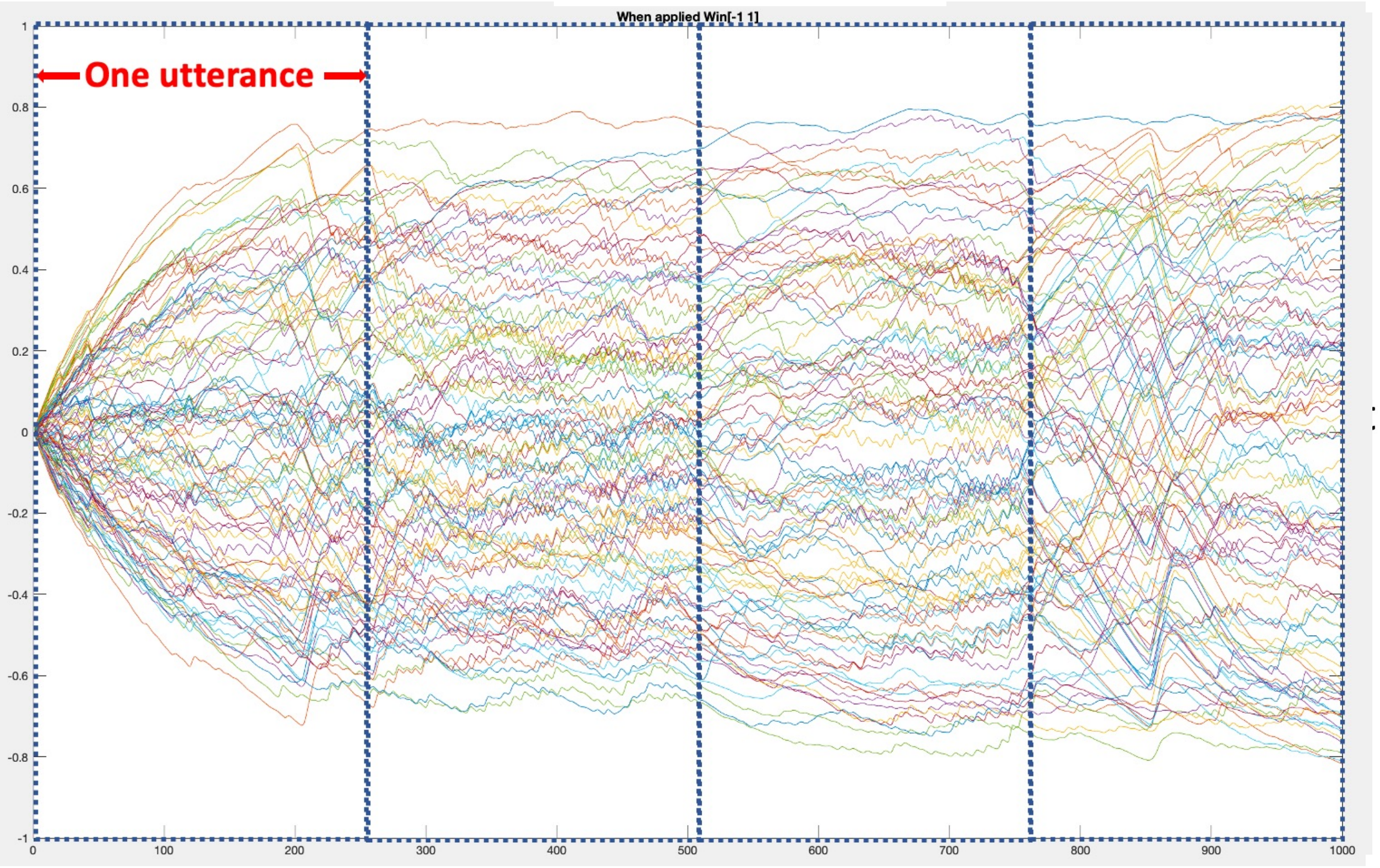
[-1 1]

W^{in}	Input connectivity weight matrix

$$x(n+1) = f\left(\boxed{W^{in}u(n+1)} + Wx(n)\right)$$

- Uniformly distributed random numbers
- It determines the performance power of the output.

➢ Set to [-1 1]



ESN model for Consonant Classification

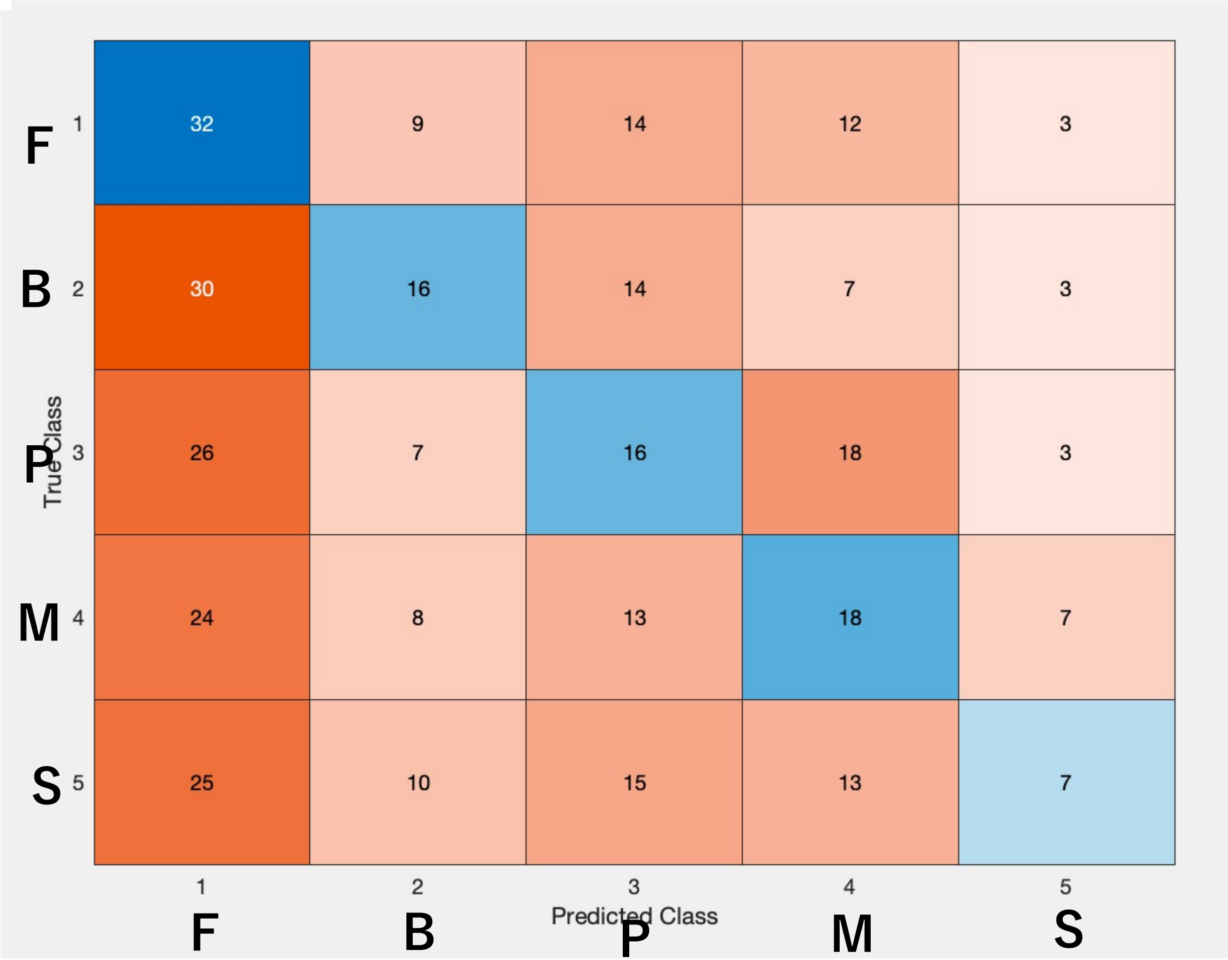
α	Leaky rate

$$y(n+1) = (1-\alpha)x(n) + \alpha f(W^{out}x(n+1))$$
$$\alpha \in (0,1]$$

- Control the speed of the time change of the reservoir state
- When $\alpha < 0.001$ $\rightarrow$ Prediction scattered
- When $\alpha > 0.1$ $\rightarrow$ Heavy concentrated

$\rightarrow \alpha = 0.009$

$\alpha > 0.1$



ESN Parameters and Settings for Consonant Classification

Parameter	Meaning	Set
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 W	0.9

Training model

Linear regression model

Sample usage ratio

90% (train), **10%** (test)

Discussion for Consonant Classification

Average classification accuracy 28.3%

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

F, B, P: Relatively high accuracy
S: Lowest accuracy

1. Consonant **B** features are more likely to appear in brain activity, while consonant **S** features may be relatively less likely to appear.

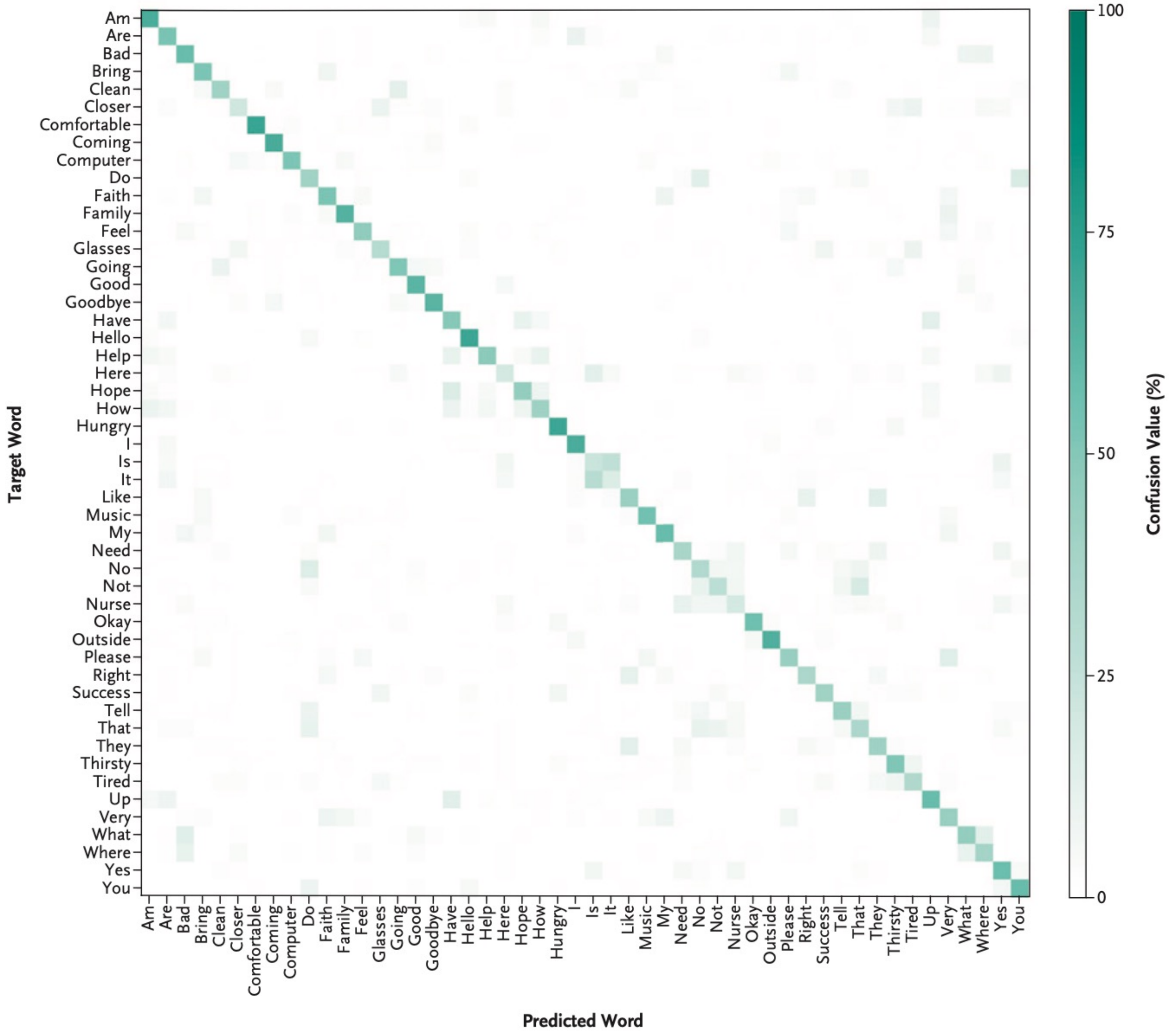
Discussion for Consonant Classification

F, B, P: Relatively high accuracy
S: Lowest accuracy

Similar tendency in Moses et al. [5]

Use words that start with the five consonants as this study

- **High** recognition accuracy for words starting with the consonants **B and F**
- **Low** recognition accuracy for words starting with the consonant **S**



Confusion matrix for 50 words classification

Result of consonant classification

Average classification accuracy 28.3%

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

F, B, P: Relatively high accuracy
S: Lowest accuracy

1. Consonant **B** features are more likely to appear in brain activity, while consonant **S** features may be relatively less likely to appear.
2. Differences in the **movement of the articulators** depending on the sound

Result of consonant classification

Differences in the movement of the articulators depending on the sound

Consonant

Plosive… A sound produced by completely blocking the path of the breath and then suddenly opening it

P and **B**

The articulators tense → Included in EEG data as features

Fricative…Made by narrowing the path of the breath_and pushing the sound out of it.

F and **S**

The articulators do not tense than plosive → Less affects EEG data

Discussion

Achievement

1. **Analyzed the pre-speech EEG**
2. **Verified speech discrimination with 28.3%**

Improvement

ESN training algorithm

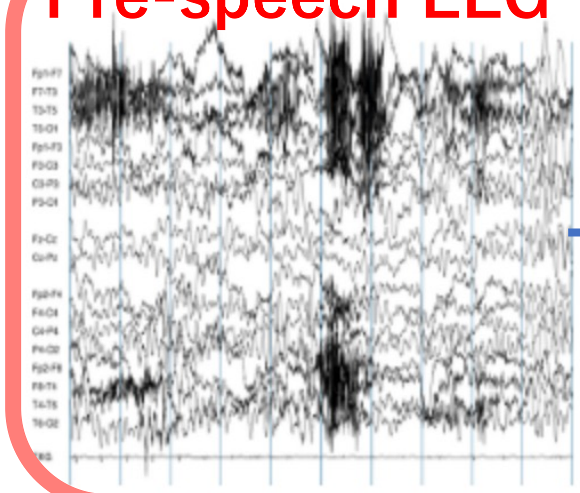
Linear regression → Gradient-based model

Subjects for EEG measurement

Non-native English speakers → Native English speakers

Conclusion

Pre-speech EEG

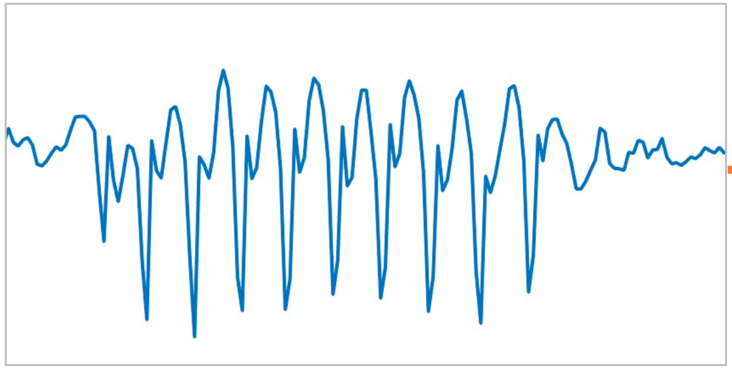


Classifier → 28.3% Consonant recognition

STUDY 2

STUDY 1

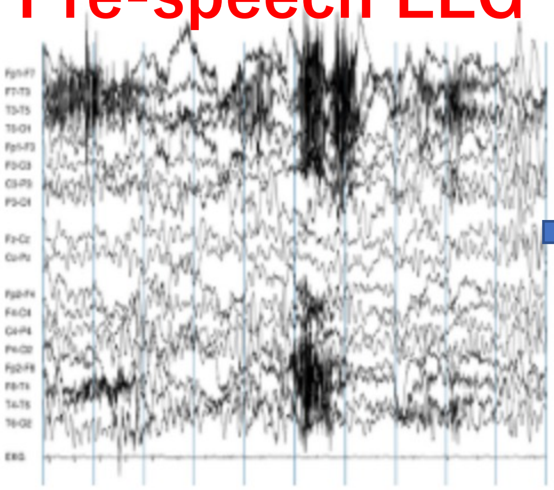
Vocal folds vibration



Classifier → 71% Vowel recognition

Future Work

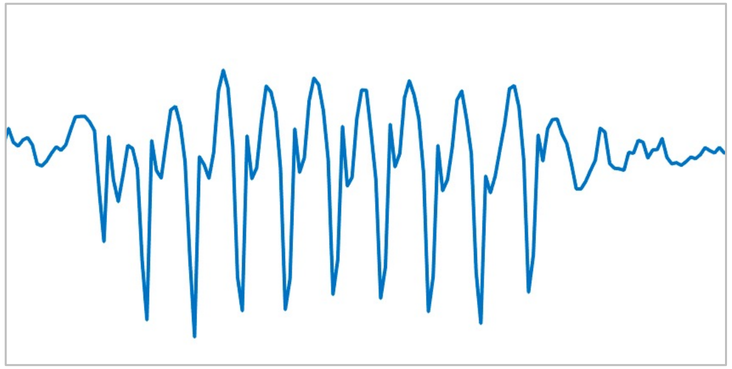
Pre-speech EEG



Enhanced
Speech Recognition

Classifier

Vocal folds vibration



Supporting Materials

Another training model: Least Mean Square (LMS)

[6] Wen et al., “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, 9 2019.

The steps of the LMS algorithm are presented as follows:

step 1 Define variables and parameters. In order to facilitate the processing, bias is combined with weights:

$$\mathbf{w}(n) = [\mathbf{b}(n), \mathbf{w}_1(n), \mathbf{w}_2(n), ..., \mathbf{w}_N(n)]^T, \quad (6)$$

where $b(n)$ is bias, n is iteration number.
The corresponding training sample is

$$\mathbf{x}(n) = [1, \mathbf{x}_1(n), \mathbf{x}_2(n), ..., \mathbf{x}_N(n)]^T. \quad (7)$$

step 2 The initialization. Assign small random initial values to the weights $\mathbf{w}(n), n = 0$.

step 3 Input the sample, calculate actual output $\mathbf{y}(n)$ and error $\mathbf{e}(n)$. According to the given expected output $\mathbf{d}(n)$, we can calculate

$$\mathbf{y}(n) = \mathbf{x}^T(n)\mathbf{w}(n). \quad (8)$$

$$\mathbf{e}(n) = \mathbf{d}(n) - \mathbf{y}(n). \quad (9)$$

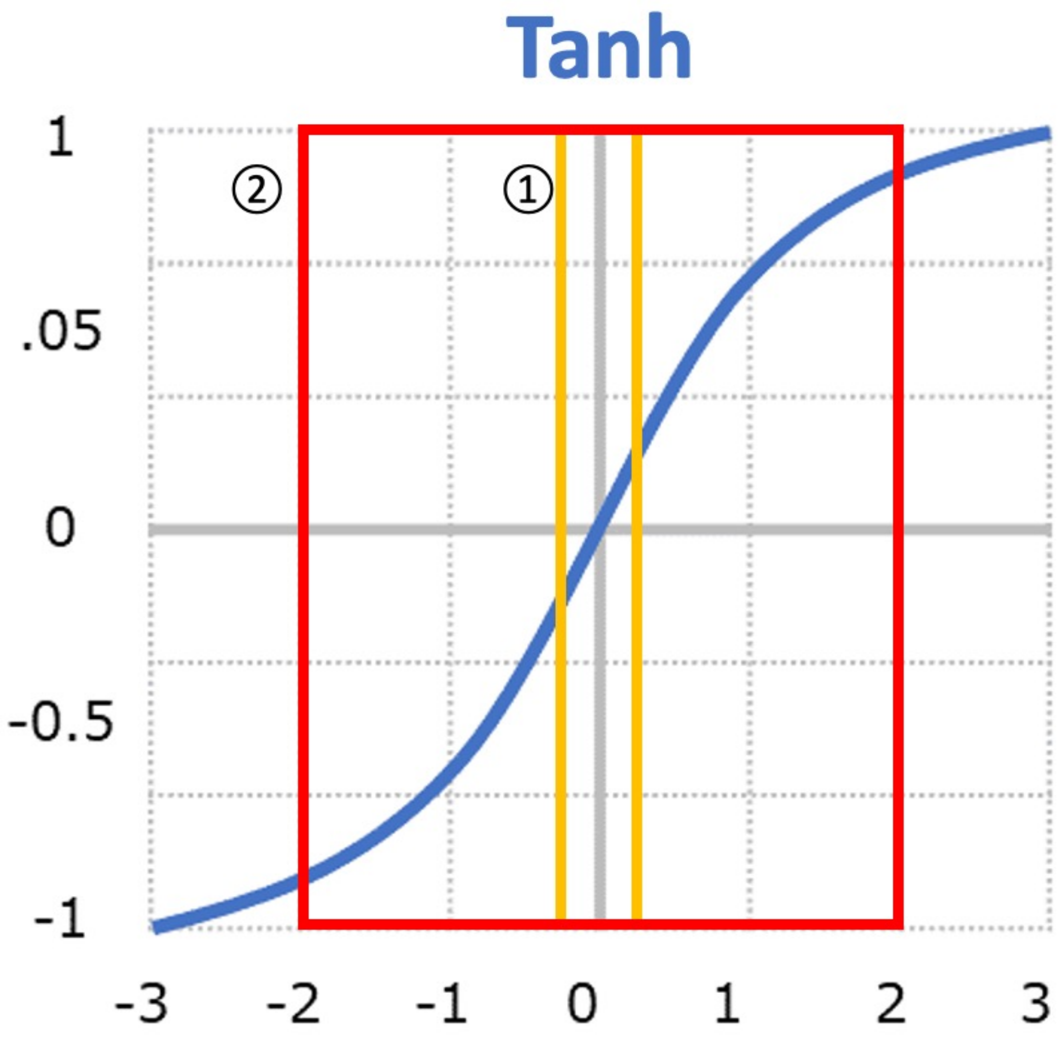
step 4 Adjust the weights vector. Set the learning rate η and calculate

$$\mathbf{w}(n+1) = \mathbf{w}(n) - \eta\mathbf{x}^T(n)\mathbf{e}(n). \quad (10)$$

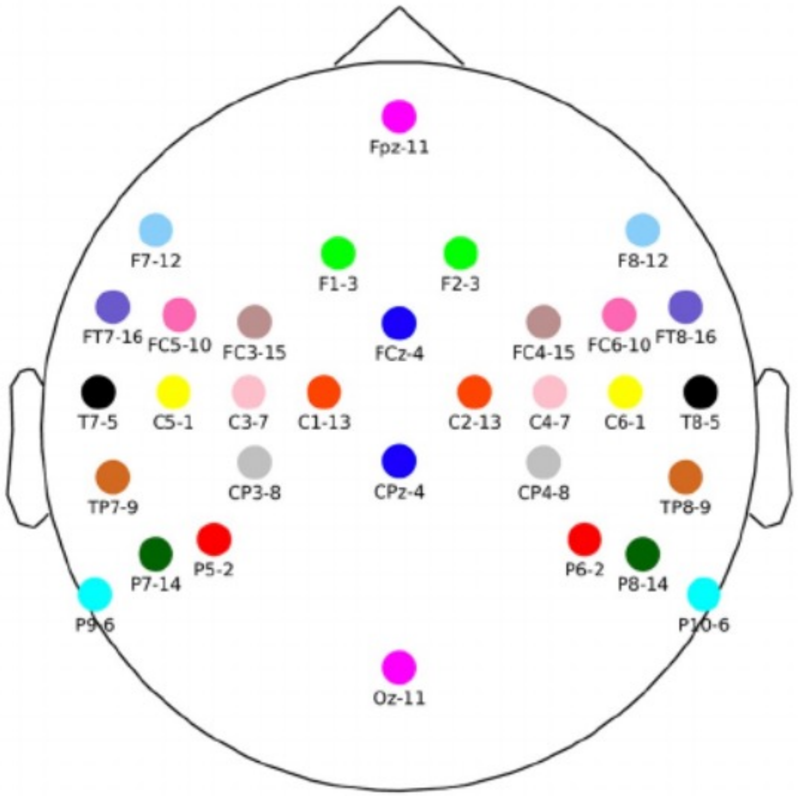
Calculates the error between the model output and the target output each time

Updates Wout sequentially to minimize the squared

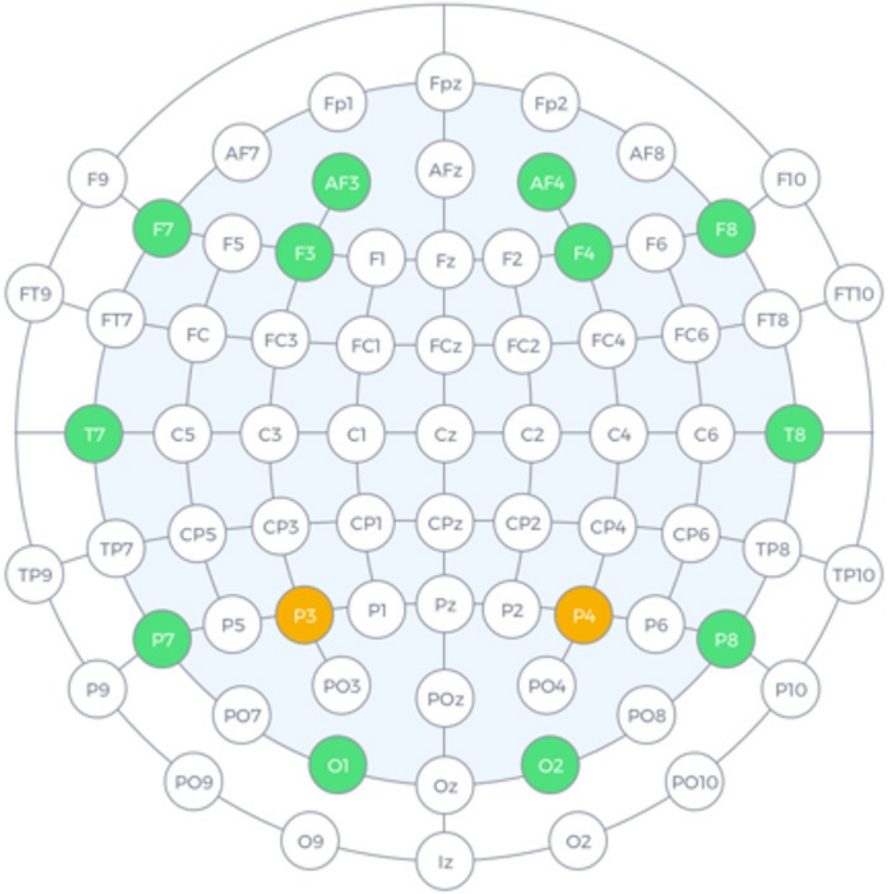
Activation Function: Tanh



Electrodes Position and Number of EEG



(d) Best 32 channels.



[7] J. Montoya-Mart ´ınez, J. Vanthornhout, A. Bertrand, and T. Francart, “Effect of number and placement of EEG electrodes on measurement of neural tracking of speech,” PLoS ONE, vol. 16, no. 2, 2 2021.

Academic Achievements

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