Doctoral Dissertation

Neural decoding of sentences using synchronization between EEG and speech rhythm

Hiroki Watanabe

September 15, 2019

Graduate School of Information Science
Nara Institute of Science and Technology

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

Hiroki Watanabe

Thesis Committee:

Professor. Satoshi Nakamura	(Supervisor)
Professor. Yuji Matsumoto	(Co-supervisor)
Assistant professor. Hiroki Tanaka	(Co-supervisor)
Associate professor. Sakriani Sakti	(Co-supervisor)
Dr. Lars Meyer	(MPI for Human Cognitive and Brain Sciences)

Neural decoding of sentences using synchronization between EEG and speech rhythm*

Hiroki Watanabe

Abstract

Recent research has attempted to realize an electroencephalogram (EEG)-based speech recognition (neural decoding of speech) during speech perception or imagined speech for providing a means of communication for patients with severe motor disabilities. The goal of the thesis is to propose novel features based on knowledge of neurophysiology for the neural decoding because the previous research is performance-oriented without the understanding of the underlying neural mechanism. For the purpose, the thesis focused on neural phase synchronization with speech. The previous magnetoencephalography research had shown that synchronization during speech perception enables speech to be classified. The thesis investigated whether the synchronization enables both perceived and imagined speech to be classified using EEG.

Experiment 1 investigated the performances in EEG-based decoding of three Japanese spoken sentences using different classifiers from the previous research (template matching: baseline, logistic regression, support vector machine (SVM), and random forest) and using phase information in multiple frequency bands relevant to linguistic processing. The trained models were evaluated by subject-dependent, -inclusive, and -independent manners. Results showed the best accuracies achieved 50.0% in SVM trained by theta, 51.9% in template matching trained by multiple frequency band (delta, theta, alpha, beta, and gamma),

*Doctoral Dissertation, Graduate School of Information Science, Nara Institute of Science and Technology, September 15, 2019.

and 50.5% in SVM trained by multiple frequency bands in subject-dependent, -inclusive and -independent classification, respectively.

Experiment 2 investigated whether EEGs during imagined speech synchronize with the rhythm of the imagery. Because of the unobservable nature of the speech imagery, the imagined speech was replaced with the overt counterpart. I regressed three types of overt nonsense speech envelopes from EEG during the speech imagery and calculate the correlation between the regressed envelope and the overt speech envelope. The template matching classified the speech to clarify whether EEG phases during the imagined speech are modulated depending on the speech imagery. The variability of the duration of the imagined speech across trials was corrected using dynamic time warping (DTW). Results showed a significant correlation between the EEG-based regressed envelope and the overt speech one. The average classification accuracy achieved 38.5%, which significantly outperformed a chance of rate (33.3%). In Experiment 3, the synchronization during the imagined speech in the linguistic relevant frequency bands (from delta to gamma) and classification performances was investigated using meaningful sentences. As results, the theta band showed the marginally significant PLVs. The DTW-based template matching trained by theta phase patterns achieved marginally significant performances (43.1%) against the chance rate (33.3%) in the subject-dependent classification.

The thesis succeeded in demonstrating the neural phase information induced by neural phase synchronization is the effective feature for neural decoding of speech by clarifying that EEG phases in both speech perception and imagined speech task enabled speech to be classified. It was also shown that the neural decoding models in speech perception are possible to be generalized to unknown users and use of phase patterns in the multiple linguistic-relevant frequency bands improve accuracies in the subject-independent classification.

Keywords:

neural oscillations, neural decoding, phase synchronization, speech rhythms

Contents

List of Figures

List of Tables

List of Abbreviations

ALS Amyotrophic lateral sclerosis.

ANOVA Analysis of variance.

AST Asymmetric Sampling in Time.

BCI Brain-computer interface.

BOLD Blood-oxygen-level-dependent.

CNS Central nervous systems.

Cphase Crosstrial phase coherence.

CSP Common spatial pattern.

DIVA Directions Into Velocities of Articulators.

Dphase Phase dissimilarity index.

DTW Dynamic time warping.

ECoG Electrocorticogram.

EEG Electroencephalogram.

ERD Event-related desynchronization.

ERMF Event-related magnetic field.

ERP Event-related potential.

FBCSP Filter bank common spatial pattern.

FIR Finite impulus response.

fMRI Functional magnetic resonance imaging.

fNIRS Functional near-infrared spectroscopy.

GLMM Generalized linear mixed model.

IC Independent component.

ICA Independent comoponent analysis.

IIR Infinite impulse response.

ITPC Inter-trial phase clustering.

LOO cv Leave-one-out cross-validation.

LOSO cv Leave-one-subject-out cross-validation.

MEG Magnetoencephalographm.

MI Mutual information.

PLV Phase-locking value.

PSP Postsynaptic potential.

RVM Relevance vector machine.

S/N Signal-to-noise.

SCP Slow cortical potential.

SSVEP Steady-state visual evoked potential.

STFT Short-time Fourier transform.

STG Superior temporal gyrus.

SVM Support vector machine.

SWDA Stepwise linear discriminant anlysis..

1. Overview of the Thesis

1.1 Brain-computer interface for communication aid

Recent technologies to measure electroencephalogram (EEG) enables users to connect their brains and a computer directly to operate machines or computers such as robot arms [36], wheelchairs [47], and spelling systems [24]. Such systems, which are called brain-computer interface (BCI), are mainly motivated to compensate for the lost physical functions of people [84]. A BCI system enables to provide the patients with the locked-in syndrome, who cannot move their muscles voluntarily with having a normal consciousness and perceptual abilities, means of expressing their intentions without making body movements. One of the most widespread BCI for communication aid is an EEG-based spelling system. The spelling system enables the users to select letters or icons on a monitor using their EEG responses, for example, a P300 event-related potential (ERP) [24].

On the other hand, for approximate the BCI-based speech communication to daily life conversation (i.e., speech communication without a spelling system), recent research has attempted brain-based speech recognition (neural decoding of speech). This decoding is positioned on one of the BCI for communication aid but aims to realize more natural communication using brain signals (Fig. 1).

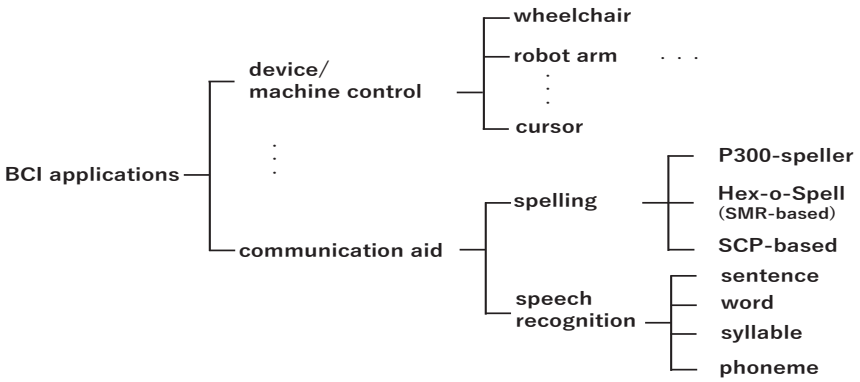

Figure 1. A position of the neural decoding of speech among the existing BCI systems for device control and communication aids. See also [53] and Chapter 2.

In the decoding, participants were required to listen to the speech (hereafter,

speech perception task) and imagine the articulatory movements of speech without any moving their articulators (hereafter, imagined speech task). The former task aims to recognize what speech the participants heard and later one does to recognize what speech the participants imagined. The first attempt of the decoding was performed by Suppes et al. (1997) [74], demonstrating that English words were discriminated using EEG in both speech perception and imagined speech tasks. One of the most recent studies of neural decoding of speech reported the imagined two words were classified with accuracy above 95% in the best case [59].

1.2 Focus of the thesis

Regardless of such successful mapping from EEG signals to speech with high accuracy, one perspective is still missing: what are neurophysiological mechanisms underpinning the neural decoding of speech? Recent performance-oriented classification algorithms (e.g., deep neural network) trained by a large dataset have been updated classification performances from conventional shallow learning algorithms in exchange for its interpretability of features contributed to classification. Such state-of-the-art classifiers also prosper the performances of the neural decoding of speech (see [48] for review of recent algorithms for BCI). However, understanding the neural mechanisms (i.e., how neural systems are modulated depending on speech stimuli) and extracting features based on the mechanism is still of importance in the neural decoding of speech. This importance is emphasized especially in an EEG-based classification, which is the most suitable for BCI systems in terms of compactness and running cost, in the following reasons. First, EEG data acquisition has not been pervasive in our daily life so that we collect EEG data to build a large dataset from scratch. Second, EEG data is prone to be contaminated by many kinds of artifacts: external line noise, muscle artifact, eye-movements, unrelated brain activities, which are greatly larger than EEG signals in many cases. To deal with the above-mentioned situations, feature engineering is considered as one of the solutions because it improves signal-to-noise (S/N) ratio of EEG data and enables classifiers to work with the limited amount of data.

Considering the above-mentioned things, the thesis aims to propose a novel feature for EEG-based neural decoding of speech based on cognitive neurophysio-

logical research. One suitable neurophysiological mechanism enabling the neural decoding of speech is neural phase synchronization with speech acoustics [49]. The neural oscillations in the auditory cortices match their phase to rhythms of speech to extract language relevant units from speech acoustics during speech perception. This means that the synchronization induces stimulus-specific, replicable EEG phase patterns leading to successful classification [49]. Luo and Poeppel (2007) [49]'s magnetoencephalography (MEG) study demonstrated that neural phase information in theta (4–8 Hz) frequency band discriminated speech utilizing the neural synchronization between theta oscillations between syllabic rhythms included in the speech. Besides, the neural phase matching to the external stimuli shows less individual variability across the users [41]. The less individual differences can be an advantage for a practical application because pre-training is not necessary for a new user once a classification model was constructed.

To utilize the neural phase synchronization for the neural decoding of speech, several points need to be overcome. (1) It is necessary to investigate phase synchronization-based classification performances using EEG, not MEG because EEG has considered as the best choice for BCI applications thanks to its portability of measuring devices with a relatively low running cost. However, given that EEG has worse S/N ratio and spatial resolution than MEG, it is still unclear whether the previous MEG-based neural decoding allows direct application to the EEG-based classification. (2) It also remains unclear whether this classification can be applied to the imagined speech task because modulations of neural dynamics during imagined speech has not yet fully understood. While neural decoding of speech requires the two tasks (speech perception and imagined speech), the decoding of imagined speech is practically utilized for the communication aid BCI that a content of the imagined speech is outputted via a speaker. If a similar neural modulation to speech perception is induced during the imagined speech, it is suggested that phase patterns of neural oscillations during imagined speech enables to discriminate the imagined speech.

Thus, the primary goal of the thesis is to reveal that the EEG phase information induced by neural synchronization successfully classify speech. This goal aims to fill a missing gap between neural engineering and cognitive neurophysiology field: The former is performance-oriented research and the latter aims

to understanding neural mechanisms in cognition. More specifically, to achieve the goal of the thesis, totally three experiments were conducted: (Experiment 1) EEG-based neural decoding of perceived speech, (Experiment 2) investigation of EEG phase synchronization with imagined speech and (Experiment 3) EEG-based neural decoding of imagined speech.

Experiment 1 investigated whether the previous MEG research [49] enables to apply EEG-based neural decoding of speech. While the decoding of imagined speech is required for the communication aid BCI, the decoding of perceived speech was treated as a first step for demonstrating the effectiveness of neural phase synchronization by extending the previous research. For the purpose, classification of three spoken sentences using EEG during a speech perception task was adopted, which was the same task to the MEG research. Another purpose of Experiment 1 is to obtain better classification results compared to the method proposed in the previous research. To this end, several different classifiers were constructed (logistic regression, support vector machine (SVM), random forest) in addition to the previous classifier, template matching. Furthermore, models were trained by other frequency bands relevant to linguistic processing (delta, alpha, beta, and gamma) in addition to theta (previous research) and combination of these frequency bands for performance improvement. These frequency bands were chosen because it has suggested that phases in these frequency bands also synchronized with speech rhythms [54].

Experiment 2 investigated whether EEGs during imagined speech synchronize with the rhythm of the imagery. This research question was set to apply the classification method to the imagined speech classification. Because of the unobservable nature of the speech imagery, the imagined speech was replaced with the overt counterpart. I regressed three types of overt nonsense speech envelopes from EEG during the speech imagery and calculate the correlation between the regressed envelope and the overt speech envelope. The template matching classified the speech to clarify whether EEG phases during the imagined speech are modulated depending on the speech imagery, which is further evidence for neural phase synchronization. In Experiment 3, the synchronization during the imagined speech in the linguistic relevant frequency bands (from delta to gamma) and classification performances was investigated using meaningful sentences.

1.3 Contributions of the thesis

The thesis contributes to a field of neural engineering by proposing novel features based on neurophysiological mechanisms for neural decoding of speech. The neurophysiological validity of features has received less attention in the previous neural decoding of speech. Besides, investigation of neural phase synchronization during the imagined speech contributes to cognitive neurophysiology. While the neural correlates of the motor imagery [27] or a time course model of the speech imagery [76] have investigated so far, the neural dynamics of speech imagery have been less attended. Such insights lead to further understanding of brain dynamics during linguistic processing. The contributions of the thesis are summarized as follows.

Thesis contributions to neural engineering and neurophysiology

1. Proposing novel features for EEG-based neural decoding of speech based on the neurophysiological mechanisms.

2. Investigating neural phase synchronization during the imagined speech for further understanding of neural dynamics during linguistic processing.

1.4 Organization of the thesis

I briefly state the organization of the following sections in the thesis.

Chapter 2. An overview of BCI is introduced: a definition of BCI, its architecture, brain signals used for it, and the introduction of BCI for communication aid and its disadvantages. An introduction and a brief history of neural decoding of speech are also stated.

Chapter 3. An introduction of neural phase synchronization with acoustic information is stated: general introduction of the synchronization, its neural generators, and quantification methods, methods to classify speech using the synchronization.

Chapter 4. Experiment of EEG-based neural decoding of perveived speech is reported and discussed.

Chapter 5. Experiment of EEG phase synchronization with the imagined speech are reported and discussed.

Chapter 6. Experiment of EEG-based neural decoding of imagined speech is reported and discussed.

Chapter 7. The summary of the thesis and future directions toward practical usage are discussed.

2. Overview of brain-computer interface

2.1 Definition and architecture of BCI

Since Hans Berger reported that electrical activity of the brain can be recorded from the human scalp [8], many research has investigated the relationships between brain activity and our sensory, motor and cognitive process. In parallel, the rapid development of brain measuring devices enables the recording of brain activity outside of the experimental room [85]. A combination of such neurophysiological research and recent rapid advances in hardware for measuring brain activity realized a direct connection between central nervous systems (CNS) such as cerebral cortex activity and computer or machine to control electrical devices in daily life. Such a brain-based system is called BCI, which can be defined as "a system that measures CNS activity and converts it into the artificial output that replaces, restores, enhances supplements or improves natural CNS output and thereby changes the ongoing interactions between the CNS and its external or internal environment (p.3)" [84]. So far, BCI applications have mainly utilized in medical and welfare areas for helping people with motor diseases such as amyotrophic lateral sclerosis (ALS) or physically handicapped people. BCI provides means of replacing their lost function of motor movements, for example, by controlling a wheelchair [47], moving a robot arm [36, 40] and moving a cursor on the computer screen [2, 9] using the brain activity (see [56] for a review of the BCIs for such an assistant technology).

An architecture of the traditional BCIs is divided into three components: data acquisition, feature extraction, feature translation algorithm and device control part [84, 82] (Fig. 2). In a data acquisition part, several types of brain activity are used for the system. The measured data are translated into a signal processing division. The division is composed of a feature extraction and translation algorithm part: The former extracts feature relevant to the task or stimulus and in the later part, the machine learning algorithm is trained beforehand and the algorithm translates the current input into the prediction output. The prediction output is converted into actual movement or control of devices such as moving a wheelchair forward or opening a robot hand.

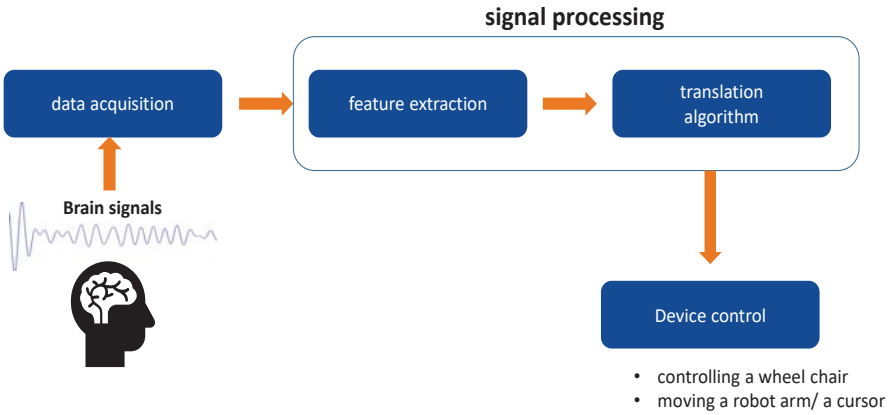

Figure 2. A figure of a BCI architecture. This figure is modified from Wolpow et al. (2002) [82], p.771. In the BCI system, features are extracted from the measured brain signals. The features are translated into command to control devices by the translation algorithm.

2.2 Brain activities for BCI

Several types of brain signals are widely used for BCI systems (Fig. 3). Each brain signal has pros and cons. The signals used for the BCI are divided in terms of types of neuronal activity (i.e., electrical activity and metabolic activity) and whether a signal acquisition system requires to injure the user (e.g., surgical operation).

The electrical activity of the brain is generated by a synaptic transmitter connecting to a pyramidal neuron. The synaptic transmitter causes a flow of an ion into a neuron (postsynaptic potential; PSP) and it creates an electrical dipole. When an electrical field of the dipole is produced synchronously by a large neuronal population heading in the same direction, these electrical activities are measurable from electrodes attached on the human scalp. These electrical signals are called EEG. In the case where the electrical field is measured from subdural electrodes placed on a surface of the brain, the signal is called electrocorticography (ECoG). While ECoG provides a good spatial (but not a whole brain, though) and its temporal resolution that enables a BCI user to achieve better performances than EEG-based one (for example, in a spelling system [15]), a purpose of implanting electrodes is limited to a medical reason such as treatment

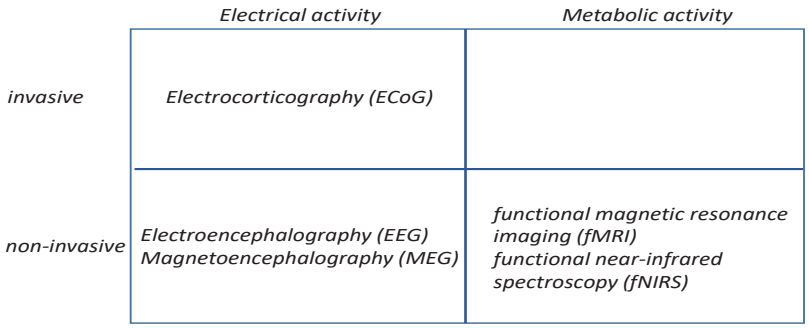

Figure 3. Types of brain signals used in BCI. The signals are divided in terms of types of brain activity (i.e., electrical activity and metabolic activity) and invasiveness.

of epilepsy. MEG is a measured magnetic field generated by the electrical activation of a neuronal population, which is captured by using superconducting quantum interference devices (SQUIDs). MEG is also measurable from outside of a human scalp non-invasively.

The metabolic activity does not measure neurons' activity directly, but measures changes in an amount of hemoglobin, which functions to supply oxygen, in the blood flow associated with the electrical activity by neurons (called blood-oxygen-level-dependent; BOLD). Among methods to measure BOLD signals, functional near-infrared spectroscopy (fNIRS) measures a ratio between oxyhemoglobin and deoxyhemoglobin. fNIRS relies on the differences in the wavelength of lights that each hemoglobin absorbs to measure the ratio by using a pair of sensors (source and detector sensors) placed on the scalp. One of the other methods to measure the BOLD signals is functional magnetic resonance imaging (fMRI). fMRI also measures a ratio of the two hemoglobins based on magneticity of deoxyhemoglobin using a scanner.

One point of view to design the BCI application is compactness and running cost of a measurement device. Among non-invasive methods, fMRI can provide a good spatial resolution (i.e., the unit of a millimeter), but the running cost is too high to retain it for personal use due to the expensiveness of helium gas which is necessary to keep superconducting materials cool. The same thing can be said

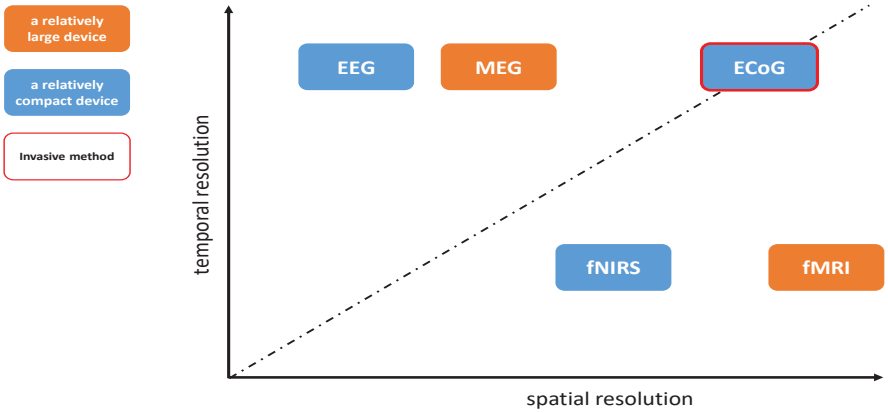

Figure 4. Each brain signal is summarized in terms of spatial and temporal resolution, compactness of measuring device and invasiveness. Compactness is expressed by a color of the box (orange: relatively large, blue: relatively compact).

to MEG. The SQUID sensors, which measure MEG signals, also need helium to keep it cool. Besides, both measuring apparatus was too bulky to place them in a user's home.

EEG and fNIRS are more attractive as a daily use BCI. They are non-invasive, compact and a relatively low running cost. However, there are pros and cons: spatial and temporal resolution. EEG has a precise temporal resolution (i.e., a unit of milliseconds) but it has a difficulty to identify neuronal sources of the electrical activity because EEG signals propagate through resistances such as skull and skin. In contrast, fNIRS has a better spatial resolution than EEG (fMRI > fNIRS > EEG) because infrared light is unaffected by resistance such skull. However, cerebral blood flows take several seconds to increase after the event onset, thus, the temporal resolutioin is not good. Considering these things, at present, EEG is the most widely used for a daily use BCI application. Types of brain signals were summarized in terms of their spatial resolution, temporal resolution, and invasiveness in Fig. 4.

2.3 Major brain responses used for a BCI application

Once brain signals are measured, features are extracted for decoding brain signals to operate a BCI application. Many existing EEG-based BCI relies on the

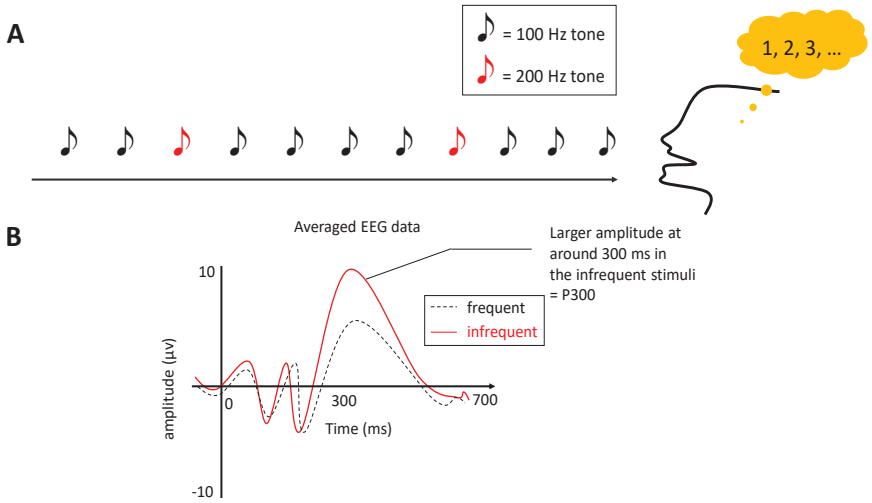

Figure 5. (A) An example of the oddball paradigm. The participant counts a number of occurrences of infrequent stimuli. (B) A schematic figure of P300. Infrequent stimuli evokes larger amplitude (i.e., P300) compared to frequent stimuli at around 300 ms from the onset (0 ms).

following brain responses: P300, steady-state visual evoked potential (SSVEP), sensorimotor rhythm (SMR), and slow cortical potential (SCP) [84].

P300. A P300 is an event-related potential measured by M/EEG (in case of MEG, referred as an event-related magnetic field; ERMF). It is elicited by an attended (e.g. by counting a number of the occurrences) infrequent stimulus among a sequence of frequent stimuli [75]. This paradigm is called an oddball paradigm (Fig. 5). This component distributes over a centro-parietal region on the scalp with a positive deflection after around 300 ms relative to stimulus onset. Farwell and Donchin [24] first applied a P300 to spell letters displayed in a monitor, which is called P300-speller.

Steady-state visual evoked potential (SSVEP). An SSVEP is a sequence of evoked potentials to a visual stimulus such as a flush. When a visual stimulus flushes repeatedly at a certain frequency, visual evoked responses are also generated at the same frequency to the visual stimuli. Utilizing this mechanism, an SSVEP-based BCI differentiates which visual stimuli a user is attending among

multiple visual stimuli flushing at different frequencies. The user can execute the function or select icons via SSVEP-based BCI because each visual stimulus is linked to function or an icon. For example, a user can decide cursor direction on a display by attending a visual stimulus linked to a direction [2].

Sensorimotor rhythm (SMR). An SMR is neural oscillations generated from a sensorimotor cortex. This oscillatory rhythm is consist of two sub-frequency bands: mu (around 8–12 Hz) and beta rhythm (around 18–30 Hz). Power of both rhythms decreases (event-related desynchronization: ERD) before actual motor movements or motor imagery with different topographical patterns depending on the movement (e.g., a contralateral hand region for hand movements; left hemisphere for right-hand movement) [58]. The differences in movement imagery can be classified based on the EEG patterns [65]. The first BCI utilizing SMR is constructed by Wolpow et al. (1991) to control a cursor based on EEG in the upward and downward directions [83]. More recently, the SMR-based BCI have achieved three-dimensional movements to control a cursor [52].

Slow cortical potential (SCP). An SCP is an event-related potential associated with motor movement and imagery. This potential lasts from several hundred milliseconds to several seconds. After the user's massive training to self-regulate an SCP for several weeks or several months, a user can control, for example, a spelling system based on SCP amplitudes [43, 44].

2.4 BCI for verbal communication

2.4.1 EEG-based spelling systems

Among the existing BCIs, verbal communication using BCI is mainly realized by a spelling system which the user can select a letter on the display using their brain activities. The first spelling system was developed by Farwell and Donchin in 1988 [24] based on P300. In the first P300-speller system, a 6×6 matrix is presented in a monitor. By a column and row flushing randomly and a user attending one letter in the matrix (participants counted the total number of the flush of the target), the same situation to an oddball paradigm were created to elicit a P300 (Fig. 6): (1) flushing of the row/columns evokes P300 to the target row/columns,

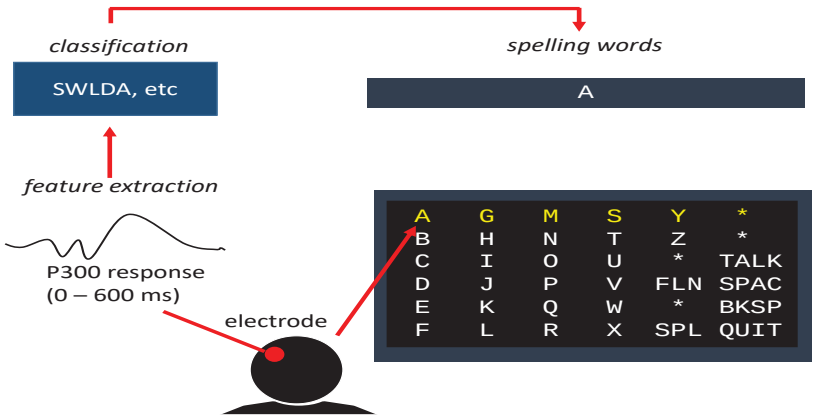

Figure 6. A schematic figure of the P300 speller. The figure of the spelling matrix is modified from [24]. The user attends one letter on the display. Each row and column flush randomly (colored in yellow). After the P300 extracted by averaging data across trials, a classifier detects the letter that the user attended.

(2) the algorithms (e.g., stepwise linear discriminant analysis; SWDA) calculates scores for identifying which column and row evoked the P300, (3) the scores of each row and column are summed to assign the scores in each letter and the summed scores are further summed across trials, (4) the letter showing the highest score is identified. In their first system, the user could produce 2.3 characters per minute with 95% classification accuracy on average with the best cases. Since then, many different types of EEG-based spelling system have proposed such as SMR-based (e.g., Hex-o-Spell; [11, 57]) and SCP-based [10]. At present, a P300-speller is the most widespread BCI spelling systems because the P300-speller took several advantages compared to other EEG-based spelling systems: (1) relatively speedy to select letters, (2) almost all people utilize a P300 speller (more than 98% among 100 people [34] and (3) a massive prior training is not necessary.

Recent EEG-based spelling system further improved speediness to select a letter. In one of the state-of-the-art BCI spelling system [72], a combination of predictive spelling, where the system shows the words candidates before selecting all characters of words, and use of a language model achieved 12.72 characters per minute. This improvement of selection speed is approximately six times faster

than the first P300 spellers (2.3 characters per minute). The P300 speller has already used for patients with ALS in their home [35].

2.4.2 Major limitations of an EEG-based speller

While the EEG-based spelling system is the representative case of BCI for communication aid, there are some limitations.

Speed of communication. Even in the state-of-the-art P300 speller, the speed of communication is still not fast compared to spontaneous speech. Whereas 2.0–3.6 words per *second* are produced in spontaneous speech [77], one of state-of-the-art P300 speller [72] produces 2.53 words per *minutes*. That is, the speed of communication is slower about 60 times than conversation in daily life. Thus, even using the state-of-the-art P300 speller, it is difficult to realize speech communication at the level of conversation speed.

Attention-based spelling. While procedures in speech conversation (e.g., moving articulators) and a skilled typing need not assign much attention on the task, the EEG-based speller requires that the users direct their attention to a monitor display. Such attention-based spelling might exhaust the user.

2.5 Neural decoding of speech

2.5.1 Toward EEG-based speech recognition

In order to overcome the above-mentioned limitations, recent research has focused on brain-based speech recognition (neural decoding of speech; see Chapter 1) using EEG during perception and imagined speech tasks: The former task is to listen to speech and the latter is to imagine articulatory movements in the mind without any actual movements (Fig. 7). Compared to one-by-one selection in the EEG-based spelling system, a time lag until the users' intention is conveyed is expected to be shorter because the neural decoding predicts speech directly from single-trial EEG. Besides, such decoding does not require users to keeping their attention to the display monitor.

The first attempt of the decoding was performed by Suppes et al. (1997) [74]. They tried to classify seven words during both speech perception and imagined

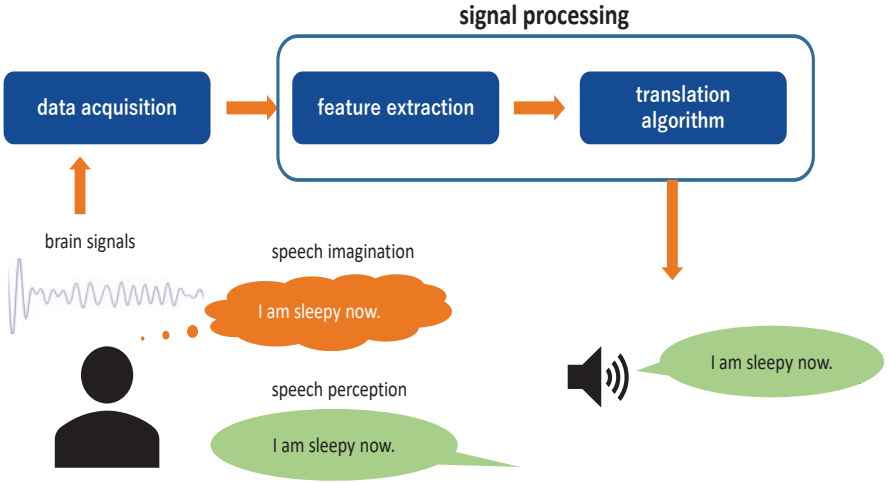

Figure 7. A schematic architecture of a speech neural decoding system. Brain signals are acquired while the user is imaging or perceiving speech. The relevant features to the linguistic processing are extracted from the signals and the decoder predicts the content of the speech.

speech using EEG signals based on distances between a prototype data (an averaged data across training set) and test data, achieving 39–46 % accuracies in single-trial classification. Although the accuracies are not enough in single-trial classification, their research, for the first time, demonstrated the feasibility of EEG-based neural decoding of speech. In the next year, they extended their method into the classification of simple sentences (subject-object-verb; e.g., *John loves Mary.*) during a speech perception task. They achieved the best 46.5% accuracy in EEG-based single-trial classification with the application of a 2.5–9 Hz bandpass filter.

In 2009, D'zmura and colleagues progressed the Suppes et al's research to classify imagined speech of 2 syllables with 3 different imagination timing (totally, 6 classes) [23]. Their classification relied on a distance between the matched filter, which is pseudoinverse of an averaged envelope across training data belonging to each class (i.e., 6 filters), and an envelope of test data. Their method provided classification accuracy from 0.38 to 0.87 showing that the beta frequency band is the most informative for classification. Since seminal works by Suppes et al. and D'zmura, many researchers have attempted the EEG-based neural decoding

of speech in both speech perception and imagined speech tasks for phoneme, syllable, word and sentence level (Table 1).

Table 1. List of single-trial, EEG-based neural decoding of speech research

Authors	Materials	Task	Accuracy
Suppes et al. (1998) [73]	12 sentences	Perception	47%
Nguyen et al. (2018) [59]	2/3 words	Imagined	70/97%
Correia et al. (2015) [19]	2 words	Perception	53.7%
Chan et al. (2011) [16]	10 words	Perception	70%
Suppes et al. (1997) [74]	7 words	Perception	46%
Deng et al. (2010)[21]	2 syllable rhythms	Imagined	72.7%
Brigham et al. (2010)[14]	2 syllables	Imagined	88%
D'zmura et al. (2009) [23]	6 syllables	Imagined	87%
Zhao and Rudzicz (2015)[87]	2 phonemes	Imagined	77.5%
Chi et al. (2011)[17]	2 phonemes	Imagined	75%
DaSalla et al. (2009) [20]	2 phonemes	Imagined	72%

To my best knowledge, the best promising classification performances in single-trial, EEG-based neural decoding of speech was achieved by Nguyen et al. (2018) [59]. They demonstrated that using a relevance vector machine (RVM) trained by tangent vectors on the Riemannian manifold, the two-word classification achieved the maximumly 96.9% accuracy in the best case and the average accuracy across 6 participants was 80.1%±8.0.

2.5.2 Limitations in neural decoding of speech

Realizing the neural decoding of speech as BCI application seems to be very attractive for patients with severe motor disabilities because it enables the user to communicate in a real-time without making body movements. To this end, some limitations need to be solved. First, the number of classes for classification is very limited. The average number of classes listed in Table 1 is only 4.33 and the number of classes in imagined speech classification is fewer than perception (Imagined: 3.17, Perception: 7.75), which it might be related to the fact that the imagined speech varied across trials temporally causes more difficulty than the

speech perception task. Second, classification performances in almost all research do not match ones of the existing spelling system. Given that the P300-speller achieved around 95% accuracy to select one character, it is considered that a similar accuracy level is required to use the decoding as the substitute for the speller.

Another limitation is that a neurophysiological mechanism underpinning the decoding of speech is still not widely understood, especially in imagined speech. Recent learning algorithms such as deep neural networks do not require feature engineering because they can learn effective feature representation automatically. However, EEG data is prone to noise contamination such as external electricity (line noise), sweat, tiredness, unrelated brain activities (background EEG), muscle artifact and so on. Besides, it is not easy to construct large dataset (i.e., hundreds or tens of thousands) for a neural decoding task. In such cases, extracting relevant features to speech information and improve S/N ratio is still one of the effective ways (see Chapter 1).

Thus, given that feature extraction is still of importance in EEG-based classification and it might lead to further improvement of classification performances, the thesis focused on proposing a novel features for neural decoding of speech based on cognitive neuroscience (see Chapter 1). One of the promising neural mechanisms enabling neural decoding of speech is neural phase synchronization with speech rhythm. In the previous research, the neurophysiological mechanism enabling a sentence classification have already clarified, which is phase synchronization between neural oscillations and speech envelopes. This is a phenomenon that timing of phase in neural oscillation match synchronizes with the timing of phase in speech envelopes during speech perception (see Chapter 3 for the details). Given that the neural phase synchronization induces speech-specific phase patterns in neural oscillations, phase information enables speech to be classified[49, 38]. In the next chapter, I explain the details of neural phase synchronization and decoding methods using the synchronization.

2.6 Summary of Chapter 2

In this chapter, a summary of the current BCIs is stated focusing mainly BCI for communicatio aid and neural decoding of speech. In sum,

- BCI applications have utilized to replace the lost function of physically handicapped people.
- EEG is the best choice for BCI applications among brain activities because it is measurable non-invasively with relatively low cost.
- One of the most spreading BCI for communication aid is the EEG-based spelling such as P300-speller. These limitations are the slow speed of communication rate and attention-based spelling where the users keep directing their attention to the display during spelling.
- EEG-based neural decoding of speech aims to recognize speech directly from EEG signals to overcome these limitations.
- The performances of EEG-based neural decoding does not achieve the level of the existing speller. Besides, neural mechanisms underpinning the decoding is not fully understood.

3. Neural Phase synchronization

3.1 Phase synchronization during acoustic processing

In this chapter, I describe details of phase synchronization phenomenon based on neurophysiological research. Besides, phase synchronization-based method of neural decoding is discussed.

Speech is consists of a spectral modulation (e.g., formant transition) and temporal modulations. Among the temporal modulations, speech envelope (Fig. 8) shows a quasi-periodic fluctuation at approximately from 4 to 8 Hz. The speech envelope is mainly dominant in syllable information.

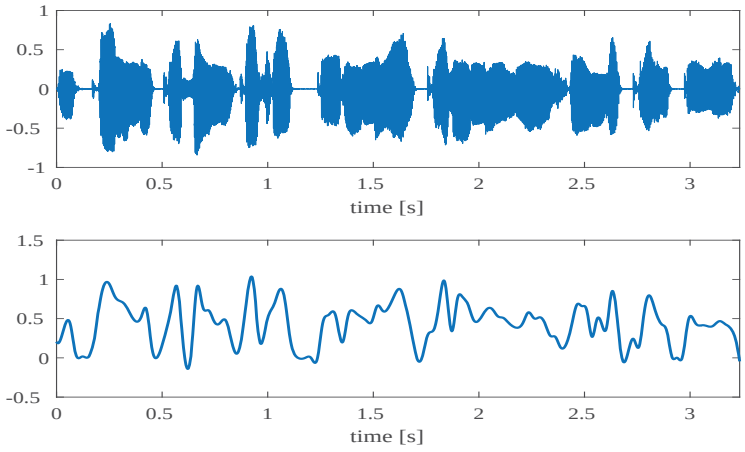

Figure 8. An example of speech (Upper) and low frequency modulation, i.e., speech envelope (Bottom).

Recent studies have revealed a role of this slow temporal modulation on speech comprehension. Ghitza and Greenberg (2009) [29] showed that the time-compressed speech by a factor of 3 relative to the original speech decreased the intelligibility, but the insertion of 80 ms silent intervals after 40 ms speech segments most improved intelligibility of the compressed speech. The key point of this study lies on a point where regardless of that spectro-temporal information of speech is the same between the time-compressed speech and the one with silent insertion, insertion of the pauses constrains the timing of decoding speech in the brain (i.e., decoding of speech at every 120 ms points; $\sim$ 8 Hz) and the timing

affects the intelligibility of the compressed speech. Thus, this result suggests that the brain prefers to a 4–8 Hz temporal window to decode speech information for comprehension.

Neurophysiological research explained that this preferred temporal window (∼4–8 Hz) is derived from endogenous neural oscillations in the auditory cortex [30]: neural oscillations in the auditory cortex fluctuates endogenously at theta frequency range, and have a functional role on the extraction of syllabic information from continuous speech [66]. This extraction is performed via phase synchronization between low-frequency modulation in speech (i.e. speech envelope) and theta oscillations (see [63, 54] for review). Because the timing of neural oscillations is associated with a change in excitability of neuronal populations [46], phase-locking between acoustic information and neural oscillations enables acoustic processing during the high excitability of neuronal populations and, thus, leads to efficient processing of input speech [63].

Acoustic information in speech includes not only syllabic level slow fluctuations but more rapid temporal information as segmental features (e.g., formant transition, >30 Hz) and slower temporal information (<4 Hz) such as intonation boundaries. For processing such rapid and slow information, parallel and concurrent information processing were postulated in the delta (0.5–4 Hz), theta and the distinct higher frequency bands (>30 Hz) of neural oscillations. The low-gamma oscillations track the rapid fluctuation and it is demonstrated that low-gamma frequency band (38–42 Hz) shows phase synchronization with acoustic fluctuation of the corresponding rhythm [50]. The delta oscillations track intonation boundaries [13]. The tracking in this slow oscillations is also observed to an abstract syntactic knowledge such as phrasal or sentence boundaries [22, 55]. A schematic figure of neural synchronization with a hierarchy of speech rhythm is described in Fig. 9.

The Asymmetric Sampling in Time (AST) hypothesis [66] argued the hemispheric lateralization in the neural oscillation tracking to acoustic information: syllable tracking via theta oscillations is performed in the right hemisphere dominantly and phoneme tracking via low-gamma oscillations dominantly in the left hemisphere. The previous simultaneous recordings of EEG and fMRI [30] and MEG research [50] supports the hemispheric lateralization proposed in the AST

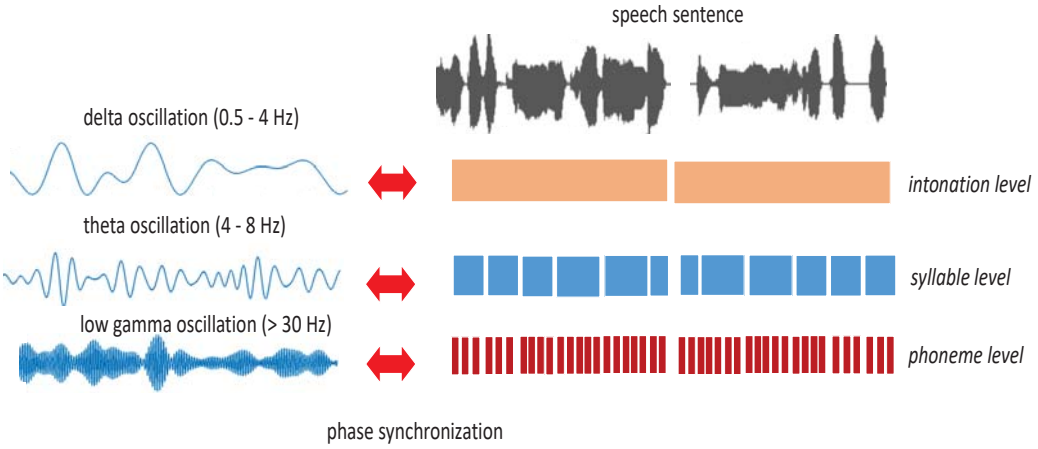

Figure 9. A schematic figure of phase synchronization with speech rhythms. The speech is consists of multiple rhythms: intonations, syllables, and phonemes. Neural oscillations in the corresponding frequency band tracks these rhythms for efficient acoustic processing.

hypothesis. Besides, Kubanek et al. (2013) investigated more detail region in the theta tracking using ECoG: the theta tracking is performed in a belt area which is adjacent to the primary auditory cortex [42]. As for intonation tracking, source localization by MEG signals revealed that the tracking in delta oscillations is observed in the posterior superior temporal gyrus (STG) in the right hemisphere [13].

3.2 Mathematical quantification of phase synchronization

Through the synchronization phenomenon, neural oscillations match their phases to the timing of the speech envelope. There are several equations to quantify a degree of this synchronization phenomenon. First, in many cases, phase-locking value (PLV) was used as the index:

$$PLV_{ch} = |T^{-1}\Sigma_{t=1}^{T} e^{i(\phi_{EEG_t} - \phi_{speech_t})}|. \quad (1)$$

T is the number of time points, ϕ_{EEG_t} is an instantaneous phase of bandpass-filtered EEG data and ϕ_{speech_t} is an instantaneous phase of a speech envelope at a time point t.

Another equation is the inter-trial phase clustering (ITPC), which calculates the consistency of phase patterns across trials:

$$ITPC_{tf} = |N^{-1}\Sigma_{trial=1}^{N} e^{i\phi_{trial}^{tf}}| \tag{2}$$

where N is the number of trials, ϕ_{trial}^{tf} is a phase angle of a time-frequency point tf. Poeppel and colleagues used a similar index to ITPC [49, 38], which is called crosstrial phase coherence (Cphase):

$$Cphase_{kit} = \left[\frac{\sum_{n=1}^{N}\cos(\theta_{knit})}{N}\right]^2 + \left[\frac{\sum_{n=1}^{N}\sin(\theta_{knit})}{N}\right]^2 \tag{3}$$

where i is a frequency bin, t is a time bin, n is a trial, k is a type of speech stimuli. More recently, mutual information (MI), which is a degree of the interdependency of two probability distributions, have used:

$$MI(X,Y) = H(X) + H(Y) - H(X,Y), \tag{4}$$

$$H(X) = -\Sigma_{i=1}^{n} p(x_i) log_2 p(x_i), \tag{5}$$

$$H(X,Y) = -\Sigma_{j=1}^{m}\Sigma_{i=1}^{n} p(x_i,y_j) log_2 p(x_i,y_j) \tag{6}$$

where $H(X)$ and $H(Y)$ is entropy of time series X and Y, respectively. $H(X,Y)$ is a joint entropy between both time series. For calculating entropy, after time series data are divided into bins (a value of bin ith in X is expressed as x_i), the probability distribution is calculated. Mutual information can capture not only a linear relationship between the two time-series signals but the non-linear relationship.

3.3 Classification using neural phase synchronization

Previous research demonstrated that this phase synchronization phenomenon enables to classify spoken sentences because phase-locked responses to speech sentence induce a replicable and sentence specific phase patterns in neural oscillations generated in the auditory cortex [49, 38]. Luo and Poeppel (2007) [49] used a MEG-based template matching method to classify three spoken sentences during speech perception. First, they calculated phase dissimilarity index (Dphase) to identify discriminable MEG channels:

$$Dphase_i = \frac{\Sigma_{k=1}^{K}\left(\frac{\Sigma_{t=1}^{T} Cphase_{itk,within}}{T} - \frac{\Sigma_{t=1}^{T} Cphase_{itk,across}}{T}\right)}{K} \tag{7}$$

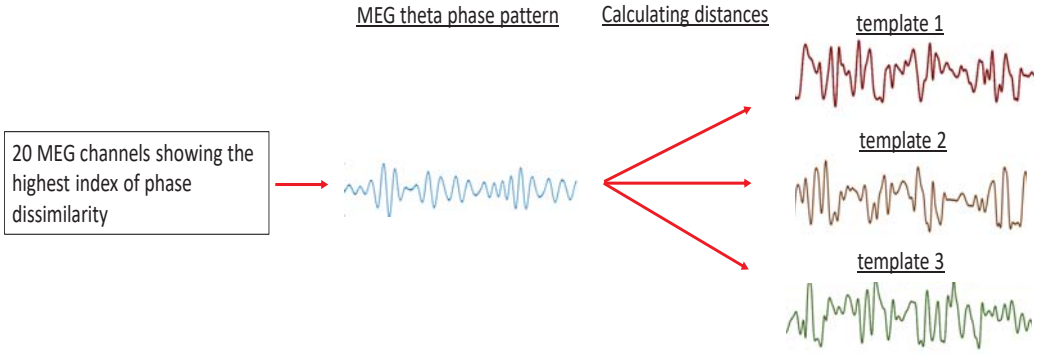

Figure 10. An schematic figure of the previous classification method. First, MEG phase patterns are extracted from selected 20 channels based on the phase dissimilarity index. The class of the template showing a minimum distance to test data is considered as a classification result.

where t is a temporal bin, i is a frequency bin, n is a trial and k is a type of sentence. $Cphase_{itk,within}$ is calculated by equation (3) using trials belonging to a sentence type k. $Cphase_{itk,across}$ is calculated using trials that is chose randomly from MEGs during listening to each sentence (7 trials from each sentence). They extracted a phase pattern of the 20 MEG responses showing the best Dphase in theta frequency bin (3–7 Hz). The phase pattern is a vector of phase values (phase patterns = $[\phi_1 ... \phi_t]$; ϕ_t is calculated in each 500 ms time window with 100 ms shift). In the following study, Howard and Poeppel (2010) performed a similar classification analysis of three spoken English sentences and reported that accuracies in subject-dependent models were from 0.402 to 0.571 (chance rate: 0.33) [38]. Their classification method is summarized in Fig. 10.

Zhang et al. (2012) [86] classified two overtly spoken sentences from ECoG signals of high gamma power. In their experiments, after the model speech was played, participants read aloud the speech overtly. For classification, they used a dynamic time warping (DTW) method, which is a non-linear manipulation to find an optimal path minimizing the distance between two signals, in order to correct a duration variability across trials. In the DTW, firstly, the distances between every possible combination of data points in two signals $X \in \mathcal{R}^{\mathcal{N}}$ and $Y \in \mathcal{R}^{\mathcal{M}}$ were calculated. This procedure constructs a distance matrix. When

using the Euclidean distance, the distance matrix is as follows:

$$d(n,m) = \sqrt{\Sigma(x_n - y_m)^2}, \tag{8}$$

$$d_\phi(x,y) = \Sigma_{l=1}^{L} \frac{d(\phi_x(k), \phi_y(k))}{C_\phi} \tag{9}$$

where $\phi_x(k)$ and $\phi_y(k)$ are warping functions. The optimal path was obtained by minimizing:

$$D(X,T) = \min_{\phi} d_\phi(X,Y). \tag{10}$$

First, overt speech data were realigned to model speech using DTW, and then, the realign matrix is used to realign ECoG data. The ECoG signal template of each class was constructed by averaging the realigned ECoG signals per class. The correlation coefficients between the realigned single-trial ECoG test data and templates were used as features. Test data was classified by linear discriminant analysis using the correlation coefficients as features. On average, their classification accuracy achieved 77.5% (chance rate: 50 %). Considering that the classification was performed by a template-based method of the speech waveforms, they state that the classification was based on differences in temporal modulation included in the speech. Thus, this result suggests that neural synchronization is also observed during generating overt speech and their classification relied on this neural synchronization. To my best knowledge, there is still no research investigating whether this neural phase synchronization can apply to single-trial, EEG-based classification. Thus, in Experiment 1, I investigated performances of EEG-based neural decoding of speech with single-trial data by extending Luo and Poeppel's research [49].

3.4 Summary of Chapter 3

In Chapter 3, the phase synchronization phenomenon between neural oscillations and speech rhythms in delta, theta, and low-gamma frequency band is explained including the neural source of this synchronization and how to quantify the synchronization. Besides, the previous method of classifying speech basd on phase synchronization are introduced. In sum,

- Speech consists of hierarchical speech rhythms: intonation (–4 Hz), syllables (4–8 Hz), and phoneme level (>30 Hz). Neural oscillations in the

corresponding frequency bands during speech perception synchronize with these rhythms to extract relevant linguistic information in parallel.

- Phase synchronization can be quantified in several methods: PLV, ITPC, Cphase, mutual information.
- Some studies are demonstrating that neural phase synchronization can discriminate speech using MEG or ECoG, but EEG-based classification using phase synchronization have less investigated.

4. EEG-based neural decoding of perceived speech

4.1 Purposes of Experiment 1

The purpose of Experiment 1 is to investigate whether neural phase synchronization can classify speech using single-trial EEG signals. While the decoding of imagined speech is required for the communication aid BCI, the decoding of perceived speech was treated as a first step for demonstrating the effectiveness of neural phase synchronization. This research is extended from the previous MEG research [49, 38]. Thus, I utilized a similar method to the previous research: classification of three spoken sentences using phase information of neural responses. The reason why Experiment 1 focused on the sentence classification is that the previous MEG research used sentences to induce the phase synchronization [49, 38]. The sentences were arbitrarily chosen as well as the previous MEG experiments in terms of not having the same words and having similar duration across sentences. Except for the above-mentioned reasons, there were no intentions in the sentence selection procedure.

The experiment also aims to obtain a better classification accuracy than a method proposed in the previous research because better classification performances are attractive for BCI and, it is expected that the EEG-based classification performances are worse than MEG-based one due to poor S/N ratio and spatial resolution [16]. First, I proposed a novel features from the previous research utilizing only theta phase information. Specifically, phase information in the lower and higher frequency bands than theta were extracted as features: delta (–2 Hz), theta (4–8 Hz), alpha (10–14 Hz), beta (16–20 Hz), gamma (38–42 Hz) because these frequency bands also shows phase synchronization with speech rhythms (As for the synchronization in the beta band, see [28]).

After classification performance of the single-frequency band was tested to investigate whether each single frequency band includes discriminable information for spoken sentence classification, features combined every frequency band was used for classification for better classification performance. Additionally, to utilize spatial information, I investigated the performances of features extracted by a filter bank common spatial pattern (FBCSP) method [3, 18]. This method is an extension of the common spatial pattern (CSP) [69], which is a method to

construct the optimal spatial filter to classify EEG data with two classes, to deal with multiple frequency bands. Second, I investigated the performances of four types of classifiers: (a) template matching (baseline), (b) logistic regression, (c) support vector machine (SVM), and (d) random forest because the previous research used only one classifier, template matching. These models were chosen because they can deal with small-scale and high-dimensional data such as EEGs.

Furthermore, I also tested the generalizability of the models to other unknown users. Although the previous methods only performed subject-dependent classification, this indicates that it is necessary to build each model specifically for each subject. This means that the number of models is equal to the number of subjects, so collecting data for model training from all users is always necessary. In contrast, when subject-independent models are trained, a model can deal with all possible subjects. This way, a model can be used for a new user that has never been seen in training data. This is crucial when constructing a BCI application without the need to retrain a model with a new subject. Given that Kerlin et al. (2010) [41] demonstrated that neural phase patterns induced by neural synchronization are replicable across different listeners, the subject-independent classification might be possible. Thus, to investigate the performances of the subject-independent classification, I compared the performances with the ones of the subject-dependent models. Besides, because of the large differences in the number of data between subject-dependent and -independent models hinder the direct comparisons of the performances, I also investigated the performances of the subject-inclusive models. In the subject-inclusive models of the current experiment, while classification accuracies of the participant data were evaluated by leave-one-out cross-validation, the training data set includes data from both the participant and other participants. By using the subject-inclusive models, the classifiers learn the participants' data with a similar number of the training data to the subject-independent classification.

In summary, research questions in the experiment are as follows:

(1) How accurately do EEG-based neural decoding models utilizing neural phase synchronization classify three Japanese spoken sentences in subject-dependent, subject-inclusive, and subject-independent manners?

(2) Which of the three classifiers improved classification accuracy over template

matching (the baseline classifier)?

(3) Do the proposed features including phase patterns in different frequency bands contribute to improving classification accuracy?

4.2 Methods of Experiment 1

4.2.1 Participants

Seventeen right-handed L1 Japanese speakers participated in data recordings. The average age (7 females, 10 males) was 24.3±1.9. All participants agreed to participate and gave informed consent in writing. They all reported no history of neurological illness and no hearing abnormalities. This experiment was approved by the ethical review board of the Nara Institute of Science and Technology.

4.2.2 Experimental materials

I constructed three Japanese sentences for the classification task (Table 2). All sentences had a similar duration and did not include the same word across sentences. The sentences were recorded by a female L1 Japanese speaker (16-bit and 44.1 kHz). She was instructed to utter these sentences at a normal speech rate and without any pauses in the middle of the sentences. The recording was conducted in a soundproof chamber. The duration range of the sentences was from 2,925 to 3,278 ms (average: 3,146 ms). Fig. 11 shows the duration of the moras, which is the rhythm unit of Japanese, in all of the spoken sentence stimuli. The peak was obtained around 6–8 Hz, which corresponds to the theta frequency range.

4.2.3 EEG recordings

EEG data were measured with an amplifier (BrainAmp DC, Brain Products GmbH., Germany) from 32 Ag/AgCl electrodes. EEGs were referenced to a right earlobe electrode. An additional AFz electrode was used as a ground. The measurement and ground electrodes were mounted on an elastic cap (EASY-CAP GmbH., Germany) according to the international 10% system. Electrode impedance was kept below 10 kΩ before the recording. Raw EEG data were

Table 2. Japanese sentences used in classification task

sentence 1
あなたが昨日夢中で読んでいた本は面白かった。
Anataga kinou muchuude yondeita honwa omoshirokatta.
(*The book that you were absorbed in yesterday was interesting.*)
sentence 2
ついさっき女の子が私に言ったことは本当の話。
Tsui sakki onnanokoga watashini ittakotowa hontouno hanashi.
(*What the girl said to me just now is true.*)
sentence 3
向こうの壁に飾っているのは彼のお兄さんが書いた絵。
Mukou no kabeni kazatteirunowa kareno oniisanga kaita e.
(*The picture on the other wall was drawn by his older brother.*)

filtered with a 0.016-Hz high-pass filter and a 250-Hz low-pass filter during the recording. The sampling rate was 1,000 Hz. Stimulus presentation was controlled by the Presentation software (Neurobehavioral Systems, Inc., U.S.A). Speech stimuli were presented via earphones (ER-1, Etymotic Research, Inc., U.S.A).

The sentences were presented to each participant aurally. A trial included one behavioral task based on previous research [38] to keep participants' attention on the stimuli. The procedure of the sentences are as follows:

1. Participants sat on a comfortable chair in a dimly lit sound-attenuating room. A monitor and keyboard were mounted on a desk in front of them.

2. They placed their right index finger on the J key and their left index finger on the F key, and they maintained this position during the data recording. The participants received instructions to remain motionless, to fixate their eyes on a fixation mark on the monitor display, to not to blink as much as possible during stimulus presentation, and to rest their eyes between trials if necessary.

3. The pairs of different or the same sentences, e.g., different: sentence 1 - sentence 2, same: sentence 1 - sentence 1, were constructed automatically. The order of pairs was randomized.

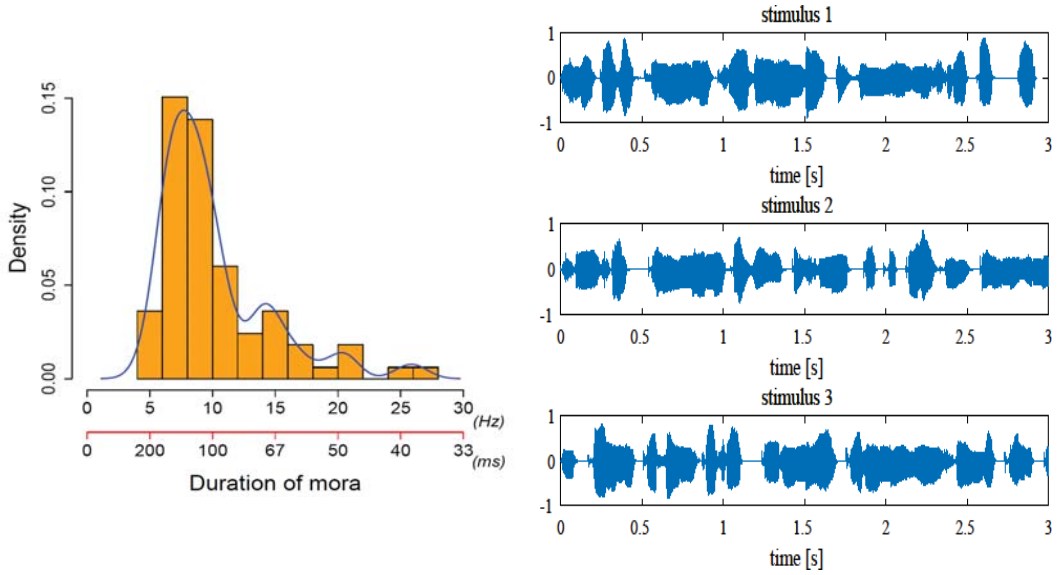

Figure 11. Histogram of average duration of moras (Left) and waveforms of spoken sentences used in experiment (Right).

4. A sentence, "Are you ready?", appeared on the display. After participants started a trial by pushing the space key, a fixation mark (+) appeared at the center of the display.

5. The first sentence in a pair was played at 1,500 ms from the trial onset, followed by the playback of the second sentence at 7,500 ms and a short tone sound at 12,000 ms.

6. Participants pushed the F key when both sentences were the same and pushed the J key when they were different. Trials finished automatically at 14,500 ms. I summarized the procedure of one trial in Fig. 12.

A brief rest was inserted after all pairs (different pairs: 6, same pairs: 3) were presented to participants. The next session was started by pushing the space key. Participants had four sessions in total with the same procedure. Each sentence was presented 24 times to each participant. The EEG data recording lasted approximately 1 to 1.5 hours including preparation.

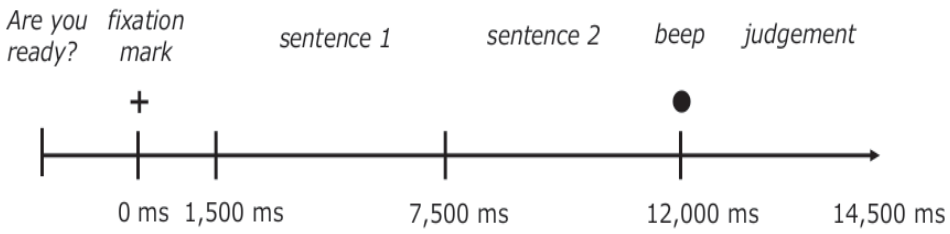

Figure 12. Procedure of one trial. Two sentences were presented to participants at a fixed timing. They judged whether the two sentence were same or not after the beep at 12,000 ms from the task onset.

4.2.4 Preprocessing of EEG

The FieldTrip toolbox for MATLAB (The MathWorks, Inc., U.S.A) was used for EEG data analysis [60]. First, one-pass zero-phase finite impulse response (FIR) high-pass filter at 0.1 Hz (filter order: 4,562-th, a window type: hamming) was applied to continuous EEG data. Line noise at 60 Hz and its second harmonics (120 Hz) were attenuated by using a discrete Fourier transform filter. And then, EEG data were re-referenced to average values of TP9 and TP10 electrodes. After EEG data were epoched from $-1{,}000$ to 4,000 ms relative to the onset of speech presentation, linear trends of EEG data were corrected.

Next, I rejected trials contaminated with large amplitude artifacts and muscle artifacts. For large amplitude artifacts, trials exceeding ± 200 μv were removed from further analysis. This rejection procedure was not applied to data from FP1 and FP2 electrodes because data from these electrodes were very often contaminated by eye movement-related artifact which was removed by independent component analysis (ICA) later. Trials including muscle artifacts were detected by a z-score-based method implemented in FieldTrip and by visual inspection. To calculate the z-score, each trial was bandpass-filtered in the range [110 Hz, 140 Hz] that is generally considered to reflect muscle activities. The filtered trials were converted to a Hilbert envelope per electrode. Data at each time point were z-normalized, and z-values were then averaged across electrodes. If any z-value in the time points in a trial exceeded a predefined threshold value, the trial was judged as having an artifact. The predefined threshold was set to 10. After this

automatic procedure, I judged whether a trial that was automatically judged as having an artifact included muscle artifacts visually. In average, 15.4%±8.9 of trials across all participants were removed.

EEG data were decomposed of independent components (ICs) by ICA. The ICs reflecting blinks, eye movements, electrocardiograms, electromyograms, and noise derived from electrodes were selected by inspecting the waveforms and topographies of the ICs visually. The selected ICs were removed from the EEG data. EEG data were low-pass filtered by a sixth-order two-pass infinite impulse response (IIR) Butterworth filter (60 Hz) to improve the S/N ratio.

4.3 Quantification of neural phase synchronization

For the purpose of determining whether EEG phases synchronized with speech, I quantified the degree of the synchronization in each frequency band (delta: 0–2 Hz, theta: 4–8 Hz, alpha: 10–14 Hz, beta: 16–20 Hz, low-gamma: 38–42 Hz) using PLV (range = [0, 1], 0: no phase synchronization, 1: perfect synchronization, see Equation (1) in Chapter 3) [45, 88]. PLV was calculated per electrode and frequency band. To calculate PLV in each frequency band, eighth order IIR Butterworth band-pass filter was applied to both EEG and speech stimuli (which was already downsampled to 1,000 Hz) at ± 2Hz at each frequency band (half-power frequency). In the case of the delta frequency band, eighth-order IIR Butterworth low-pass filter was applied at 4Hz (half-power frequency). To take consistency across trials and participants into accounts, before taking the norm of averaged PLVs in the time domain, the values were averaged across trials and participants in a complex domain. After the averaging, the norm of the values was calculated [88].

To determine whether the calculated PLV in each electrode and each frequency band was random or not, a permutation test was performed by constructing a null-hypothesis distribution. To construct the distribution, PLVs were calculated for the dataset that the correspondences between speech and EEG were randomly shuffled. The maximum and minimum null-hypothesis PLVs among electrodes and frequency bands were selected to construct the distribution to correct multiple comparisons [51]. The alpha level was set to 0.05. The number of iterations was 2,500. I counted the number of values of the null-hypothesis distribution

exceeding the observed PLVs for each electrode and each frequency band. The number was divided by the number of iterations for calculating p values.

4.4 Spoken sentence classification

4.4.1 Feature extraction

I extracted phase information from EEG trials by using short-time Fourier transform (STFT; FFT points: 500, shift points: 100, Hanning window tapering, 30 windows in total; duration of EEG trials: 0–2,900 ms). Here, i, j, k, and n represent a frequency bin (2–50 Hz; 2-Hz interval), a shifting window in STFT, a sentence type, and a trial. The phase angle vectors in a single-trial were concatenated across channels, thus, the phase angles in one trial had a vector $\phi_{fk} \in \mathcal{R}^{1\times(N_w\times N_c)}$ where N_c is a number of electrodes (i.e., 30), N_w is a number of time windows (i.e., 30) in a fth frequency bin at a kth trial . To construct features corresponding to each frequency band, I further concatenated the phase angle vector ϕ_{fk} across frequency bins (2 Hz for delta, 4–8 Hz for theta, 10–14 Hz for alpha, 16–20 Hz, and 38–42 Hz for gamma). The number of feature dimensions was 900 and 2,700 for delta and other frequency bands, respectively.

Besides, features were extracted using FBCSP method to utilize both spatial information and multiple frequency bands [3, 18]. In this method, a CSP method [69] is applied to band-pass filtered EEG signals in multiple frequency bands. The CSP is a spatial filter to maximize the difference in the variance across EEG signals in the two-class classification. These CSP-filtered features from multiple frequency bands were concatenated and utilized as an input of classifiers.

The CSP algorithm is as follows. First, EEG data matrix is represented by $E^{N\times T}$ where N is the number of channels and T is the number of time samples. The covariance of EEG data is calculated:

$$C = \frac{EE^T}{trace(EE^T)} \tag{11}$$

where T represents tranpose of a matrix. The values of the covariance matrix were averaged across trials belonging to a class. Composite covariance matrix is obtained by

$$C_c = C_1 + C_2. \tag{12}$$

C_1 and C_2 represets the averaged covariance matrix from trials belonging to class 1 and class 2, respectively. C_c is factored by eigenvalue decomposition:

$$C_c = V_c \lambda_c V_c^T \tag{13}$$

where V_c is a matrix of eigenvectors and λ_c is the matrix including eigenvalues in a diagonal elements. For the purpose of equalizing the variances in the V_c space, the whitening matrix is construted by

$$P = \sqrt{\lambda_c^{-1}} V_c^T. \tag{14}$$

And, then a spatial filter is constructed by

$$W = (V_c^T P)^T. \tag{15}$$

Finally, EEG data E was projected by the spatial filter W:

$$Z = WE. \tag{16}$$

Generally, in the CSP method, only first and last mth rows were used as features for classification. In the current experiment, $m = 2$ was used. The procedure of FBCSP used in the current experiment is summarized in Fig. 13. For classification, a classifier was trained using these concatenated signals. I used a one-vs-one strategy for three-class classification in the current experiment because the CSP method can deal with the two-class classification.

4.4.2 Classifiers and evaluation method

Template matching, logistic regression, SVM, and random forest were trained for classification. The task was to predict a sentence by using phase angle features in a single-trial EEG. The classifiers were trained by using phase angle features in each single frequency band and ones combined across all frequency bands (*multiple frequency bands feature*). I used a Python library, Scikit-learn [62], and custom Python scripts for training and evaluating classifiers.

The models were trained in subject-dependent, -inclusive and subject-independent manner, which were evaluated using leave-one-out cross-validation (LOO cv) and leave-one-subject-out cross-validation (LOSO cv). In LOO cv, classification

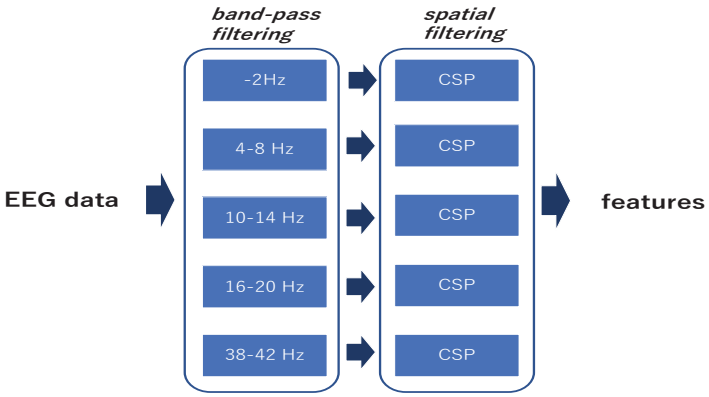

Figure 13. Feature extraction by FBCSP [3]. A CSP method is applied to band-pass filtering data in each frequnecy band. The spatial filtered data in each frequnecy band are concatenated as feaures.

models were constructed using single-participant data. In case of the subject-dependent model, one sample among trials in the dataset is used for test data, another sample is used for validation data to decide the optimal hyperparameters of models and the remaining data were used for model training. This procedure was repeated for all samples to be used as test data. In the case of subject-inclusive classification, data from other subjects were used for training data. Other procedures were the same as the subject-dependent classification. In LOSO cv, I used the data of one subject as a test set and the data of the next subject as a validation set for hyperparameter tuning. The remaining data were used for training. A validation dataset was not constructed in template matching because the model did not have a hyper-parameter.

Because EEG data were easily contaminated by artifact, there is a possibility that the classifier relied on information unrelated to phase synchronization. To confirm the possibility, feature importance of trained classifier was calculated per classifier and the topographical patterns of the feature importances were plotted. If an obvious mismatch between PLV topographies and feature importance topographies was observed, the classifier was expected not to rely on phase synchronization information. The descriptions of the classifiers used in the experiment are as follows.

(1) Template matching

I created a vector of the averaged features per class; each element of feature vectors was averaged across trials in a training dataset. The vector of the averaged features was considered as a template of each class. The Euclidean distance between the test data and each template was calculated. The class with the minimum distance was considered as a prediction result:

$$\hat{S} = \arg\min_{S} \Sigma_{t=1}^{T} (\Phi_{St} - \Phi_{xit})^2 \tag{17}$$

where $\hat{S}$ is an estimated sentence, Φ_{St} is a phase value of the template at time t and Φ_{xit} is a phase value of test data i at time t. T is the total number of time points. The variance of each feature among templates was utilized as the feature importance because a larger variance indicates that the distances among the templates in the feature are far apart from each other.

(2) Logistic regression

Logistic regression is a method to can be expressed as follows:

$$p(S_c = 1|\Phi_{xi}) = \frac{1}{1 + e^{-(w^{\mathrm{T}}\Phi_{xi} + b)}} \tag{18}$$

where S_c is a class of sentence, Φ_{xi} is a phase pattern of test data i, w is weights of features. Classifiers were trained by using L2 regularization. I used a one-vs-the-rest multiclass strategy. The best cost parameter was searched for in the range $[-4, 4]$ in log space. I used the average weight assigned to each feature among one-vs-the-rest classifiers as feature importance.

(3) SVM

A formula of SVM can be expressed as follows :

$$g(\Phi_x) = \begin{cases} 1 & w^{\mathrm{T}}\Phi_{xi} + b \geq 0 \\ -1 & w^{\mathrm{T}}\Phi_{xi} + b < 0. \end{cases} \tag{19}$$

SVM can decide a class boundary to maximize a margin by solving the following optimization problem:

$$\min_{w,b,x} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \xi_i \quad \text{s.t. } y_i(w^{\mathrm{T}}\Phi_{x_i} + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \tag{20}$$

where y_i is a class label, w is a weight vector, Φ_{x_i} is a phase pattern in trial i, C is a cost parameter. I used a linear kernel and the one-vs-the-rest multiclass strategy for classification. The tuning of a cost parameter, type of regularization, and procedure for calculating feature importance was the same as logistic regression.

(4) Random forest

Random forest is one of ensemble learning algorithms: performing random sampling, constructing decision trees and predict labels by voting among the decision trees. In a decision tree algorithm, each split at a node is performed to maximize information gain (IG):

$$IG(T) = I(S) - \sum_{j=1}^{m} \frac{|S_j|}{|S|} I(S_j) \tag{21}$$

where $I(S)$ and $I(S_j)$ are impurity of a whole data set and a subset, respectively. $|S|$ and $|S_j|$ represents a length of the two sets. In this classification, the impurity is defined by entropy:

$$Entropy = -\sum_{i=1}^{K} p_{mk} \log_2 p_{mk} \tag{2}$$

where p_{mk} is a proportion of the number of k-th class data. K denotes a total number of classes. The best number of trees and maximum tree depth were searched for in the ranges [10, 50, 100, 150] and [5, 10, 15], respectively.

4.5 Results of Experiment 1

4.5.1 EEG phase synchronization with speech

To determine whether phase synchronization to speech stimuli was observed, I plotted PLV topographies per frequency band (Fig. 14). The electrodes shown by stars represent that the PLVs of the electrodes statistically significantly differed from the null-distribution. As results, fronto-central electrodes in theta showed significant phase synchronization with speech (C3, C4, F7, T7, Fz, Cz, FC2, FC5, FC6, and CP5). While there were no other significant electrodes, a fronto-central

region in alpha, beta, and gamma showed a tendency of phase synchronization: the similar topographical patterns to significant electrodes in theta. Taking these results into accounts, these frequency bands also might contribute to the neural synchronization-based classification of spoken sentences.

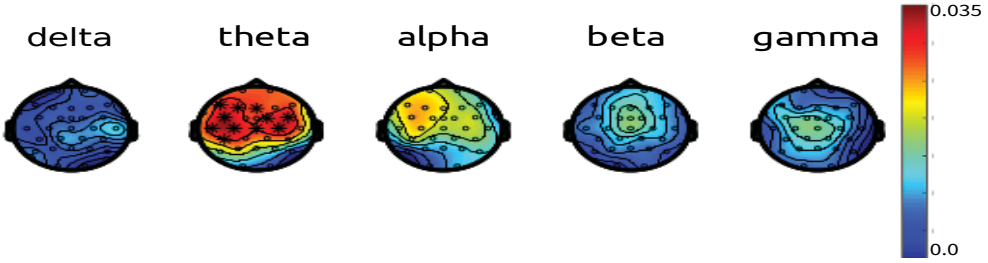

Figure 14. Topographies of PLV for each frequency band. Electrodes shown by stars represents statistically significant differences from the null-distribution.

4.5.2 Topographies of feature importance

I plotted topographies of feature importance per frequency band and classifier in subject-dependent (Fig. 15), subject-inclusive (Fig. 16) and subject-independent classification (Fig. 17). An overall visual inspection suggests that in the subject-dependent classification, all classifiers assigned the importance on the fronto-central region in the theta frequency band with a tendency of the right lateralization, but ones in SVM and logistic regression seems to be more local at a single electrode. Such local response at a single electrode is unlikely to be observed in EEG topographies because EEG tends to spread over the scalp due to volume conduction. Other fronto-central responses are observed in beta in SVM and logistic regression, but these are also local responses.

In the subject-inclusive and independent classification, fronto-central responses of theta in SVM and logistic regression diminished. The template matching and random forest showed similar topographical patterns to PLV.

4.5.3 Classification performances

Subject-dependent classification

The subject-dependent classification accuracy from each participant was summa-

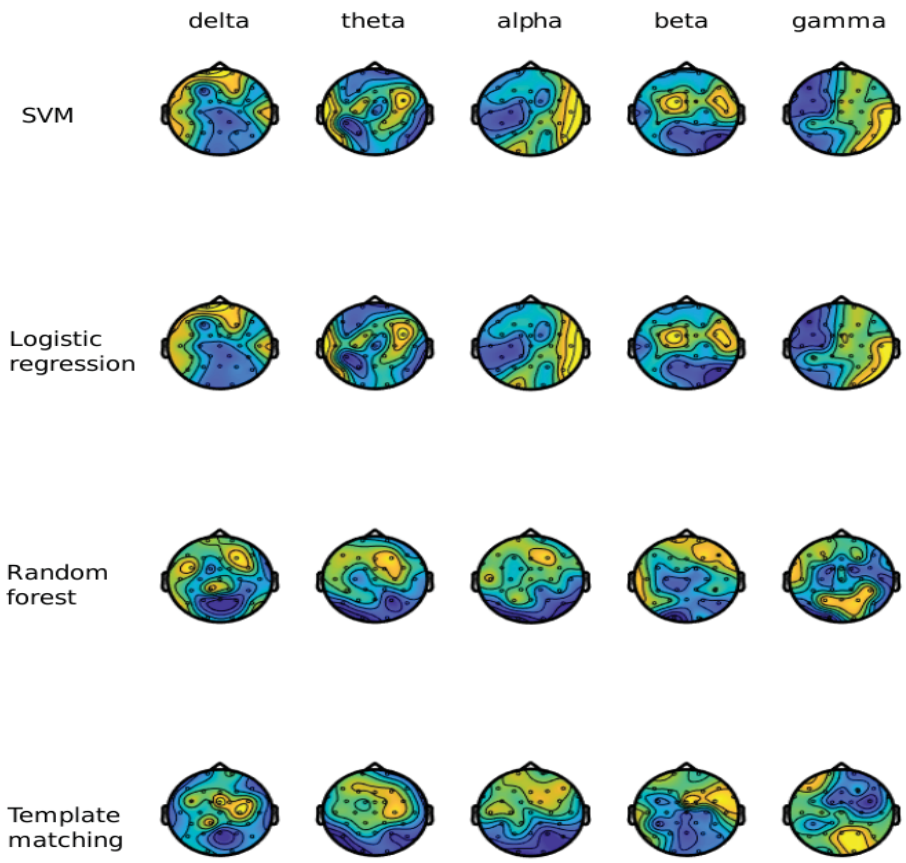

Figure 15. Feature importance topographies for each frequency band and each classifier. The feature importance was obtained from classifiers trained in a subject-dependent manner.

rized in Fig. 18 per combination of classifiers and features. The best mean accuracy across participants was 50.0% from SVM trained by theta features (Table 3). To judge whether performances from each model is random, I performed one-sample t-test on accuracy from each participant against a 33.3% chance level. As a result, SVM trained by delta, theta, alpha and *multiple frequency bands*, logistic regression trained by delta, theta, alpha and *multiple frequency bands*, random forest trained by theta, beta and multiple frequency bands, template matching trained by theta and multiple frequency bands achieved significant performances. However, the performance of random forest trained by beta was significantly

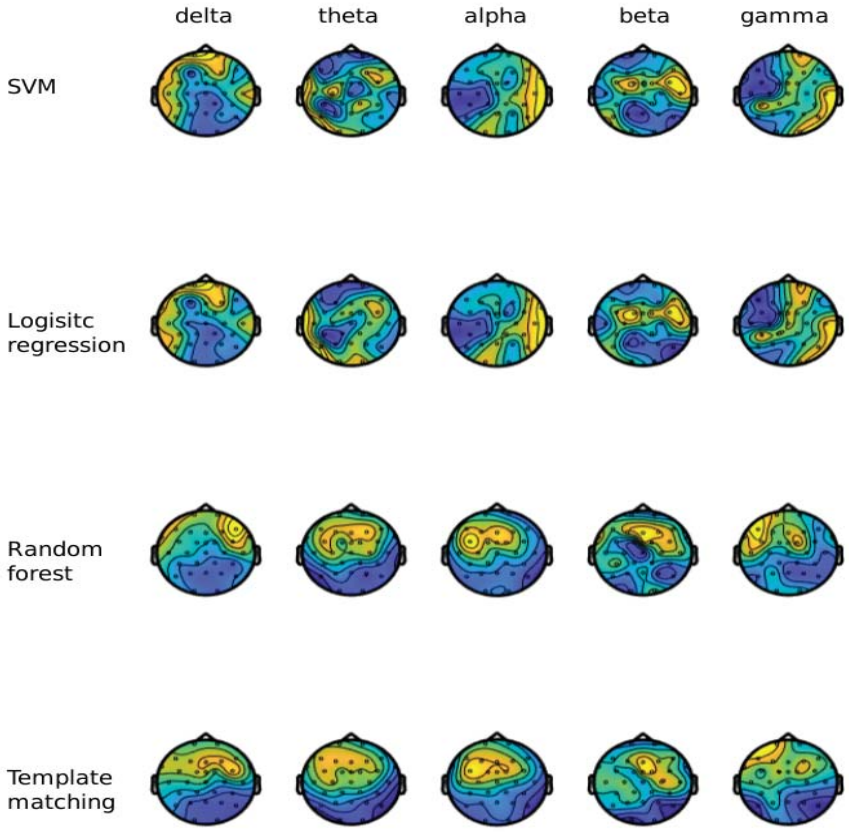

Figure 16. Feature importance topographies for each frequency band and each classifier. The feature importance was obtained from classifiers trained in a subject-inclusive manner.

worse than the chance rate. The classification accuracies which were significantly above the chance rate (33.3%) were marked with asterisks (*) in Table 3.

To evaluate the effect of classifiers and features on the accuracy, I constructed a generalized linear mixed model (GLMM) by using the number of correct classifications as response variables. I used R [68] and an *lme4* package [6] for model construction. The response variables were postulated to follow a binomial distribution because the variables can take only two values for each piece of test data: correct or not in N trials (N depends on each participant). A logit link function was used for the model. Types of classifiers and features were incorporated into

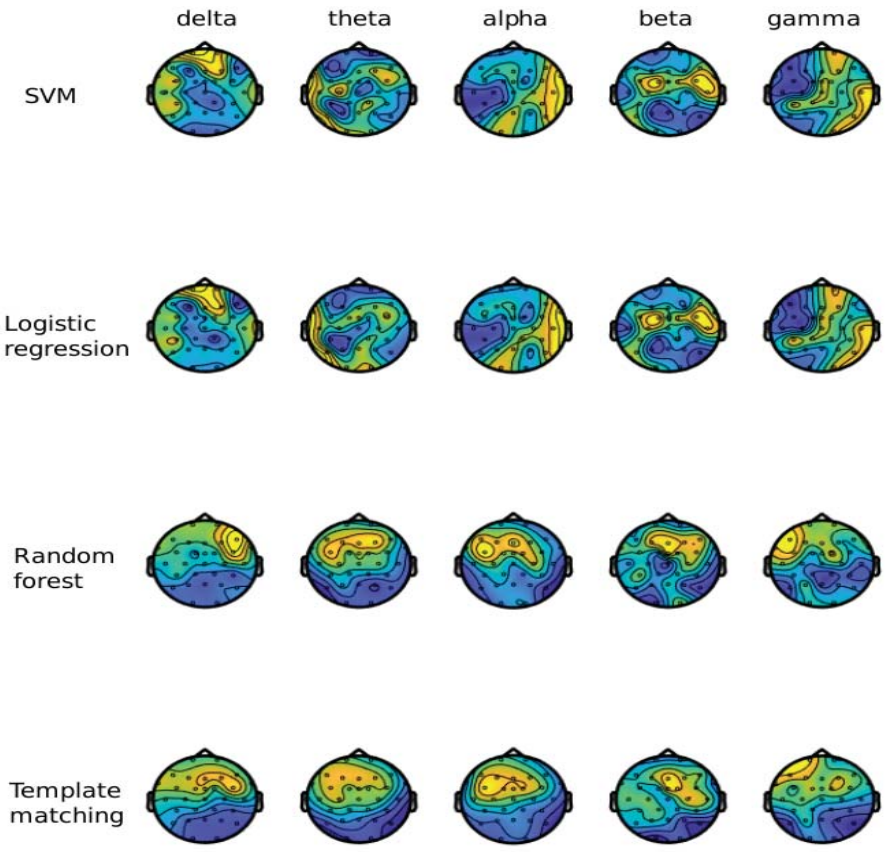

Figure 17. Feature importance topographies for each frequency band and each classifier. The feature importance was obtained from classifiers trained in a subject-independent manner.

a GLMM as fixed effects. The model had intercepts for each fold as random effects. The statistical significances of fixed effects were tested by using Type II Wald chi-square tests with a *car* R package [26]. As a result, I found a statistically significant effect of features [$\chi^2(5) = 265.4$, $p<0.01$] and classifiers[$\chi^2(3) = 44.1$, $p<0.01$].

I performed multiple comparisons for each fixed effect by using a *multcomp* package [37]. *P*-values were adjusted for multiple comparisons by using the Tukey-Kramer method. Multiple comparisons among classifier types revealed that random forest lowered accuracy compared with the other classifiers signifi-

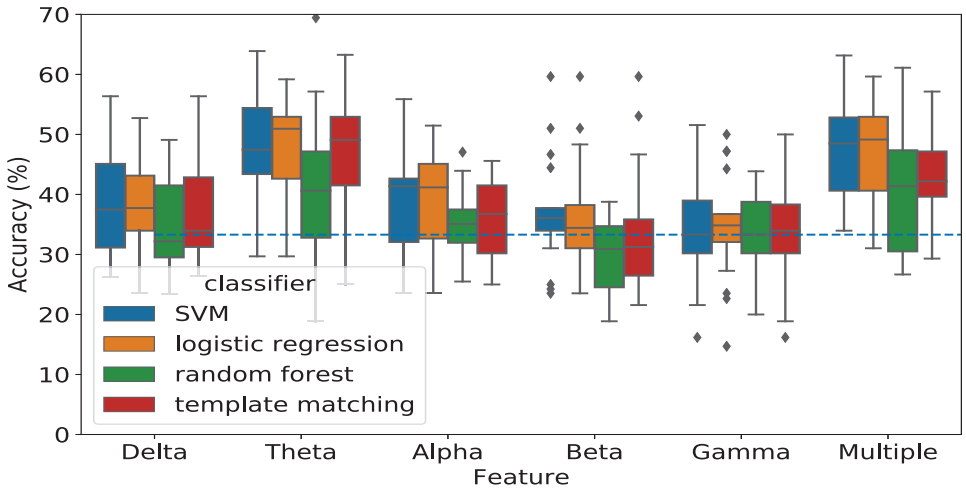

Figure 18. Classification accuracies per classifier and feature in subject-dependent classification. Each box represents accuracies from all folds. Horizontal line is 33.3% chance level.

cantly (Table 4). Besides, SVM outperformed marginally significantly template matching (p=0.09). The other classifiers did not differ from each other. The results of multiple comparisons among feature types were summarized in Table 5. More importantly, theta features marginally significantly outperformed multiple frequency bands feature.

To further explore the effects of classifiers on the accuracies, I performed multiple comparisons of classification accuracies of classifiers trained by features from which the best performance was obtained (i.e, theta) using the Tukey-Kramer method. As results, all comparisons did not show any statistically significant differences (p>0.5).

Subject-inclusive classification

The subject-inclusive classification accuracy from each participant was summarized in Fig. 19 per combination of classifiers and features. The best mean accuracy across participants achieved 51.9% from template matching trained by the multiple frequency bands feature (Table 6). To judge whether performances from each model is random, I performed one-sample t-test on accuracy from each

Table 3. Mean accuracies across participant per classifier and feature type in subject-dependent classification. Sample standard deviations are given in parentheses. Best accuracy is shown in bold.

	Delta	Theta	Alpha	Beta	Gamma	Multi
SVM	38.4* (8.7)	**50.0*** (10.6)	39.1* (8.5)	36.9 (9.1)	34.5 (9.1)	47.2* (8.4)
LR	37.9* (7.6)	49.6* (10.8)	38.6* (7.5)	36.7 (9.3)	34.4 (9.0)	47.5* (8.4)
RF	34.9 (7.8)	41.6* (12.0)	35.0 (5.4)	29.8 (6.2)	32.9 (6.9)	40.4* (10.3)
TM	37.9 (9.2)	49.1* (11.6)	36.3 (6.6)	34.1 (10.2)	33.1 (9.1)	43.2* (7.2)

Multi: Multiple frequency band feature, LR: Logistic regression, RF: Random forest, TM: Template matching, * $p<0.05$

Table 4. Multiple comparisons for classifier types in subject-dependent models

	Estimate (S.E)	p-value
Logistic regression -vs- Template matching	0.076 (0.376e-01)	0.18
SVM -vs- Template matching	-0.088 (0.381e-01).	0.09
Random forest -vs- Template matching	-0.136 (0.375e-01)**	<0.01
SVM -vs- Logistic regression	0.012 (0.375e-01)	0.99
Random forest -vs- Logistic regression	-0.212 (0.379e-01)**	<0.01
Random forest -vs- SVM	-0.224 (0.379e-01)**	<0.01

** $p<0.01$, . $p<0.1$. S.E denotes standard error.

participant against a 33.3% chance level. As a result, almost all models achieved significant performances above the chance level, except for template matching trained by beta. The classification accuracies which were significantly above the chance rate (33.3%) were marked with asterisks (*) in Table 6.

To evaluate the effect of classifiers and features on the accuracy, I constructed GLMMs again by using the number of correct classifications as response variables. As a result, while I found a statistically significant effect of features [$\chi^2(5) = 247.61$, $p<0.01$], there was no main effect of classifiers [$\chi^2(3) = 3.00$, $p=0.39$]. Multiple comparisons among classifiers and features were summarized in Table 7 and Table 8, respectively. The models trained by multiple frequency bands features did not show a significant difference from the performances than models

Table 5. Multiple comparisons for feature types in subject-dependent models

	Estimate (S.E)	p-value
Theta -vs- Delta	0.423 (0.457e-01)**	<0.01
Alpha -vs- Delta	0.014 (0.464e-01)**	0.99
Alpha -vs- Theta	-0.410 (0.457e-01)**	<0.01
Beta -vs- Delta	-0.132 (0.469e-01).	0.06
Beta -vs- Theta	-0.555 (0.461e-01)**	<0.01
Beta -vs- Alpha	-0.146 (0.468e-01)*	<0.01
Gamma -vs- Delta	-0.153 (0.469e-01)*	<0.05
Gamma -vs- Theta	-0.577 (0.462e-01)**	<0.01
Gamma -vs- Alpha	-0.167 (0.469e-01)**	<0.05
Gamma -vs- Beta	-0.021 (0.474e-01)	0.99
Multi -vs- Delta	0.307 (0.458e-01)**	<0.01
Multi -vs- Theta	-0.117 (0.451e-01).	<0.1
Multi -vs- Alpha	0.293 (0.458e-01)**	<0.01
Multi -vs- Beta	0.439 (0.462e-01)**	<0.01
Multi -vs- Gamma	0.460 (0.463e-01)**	<0.01

S.E denotes standard error. ** p<0.01, * p<0.05, . p<0.1.
Multi: Multiple frequency bands feature

trained by theta.

To further explore the effects of classifiers on the accuracies, I performed multiple comparisons of classification accuracies obtained from classifiers trained by features from which the best performance was obtained (i.e, multiple frequency bands feature) using the Tukey-Kramer method. As results, all comparisons did not show any statistically significant differences (p>0.5).

Subject-independent classification

Fig. 20 summarizes the classification accuracies from folds per combination of classifiers and features in subject-independent classification. The best mean accuracy across folds was 50.5% from SVM trained by multiple frequency bands feature (Table 9). To judge whether performances from each model is random

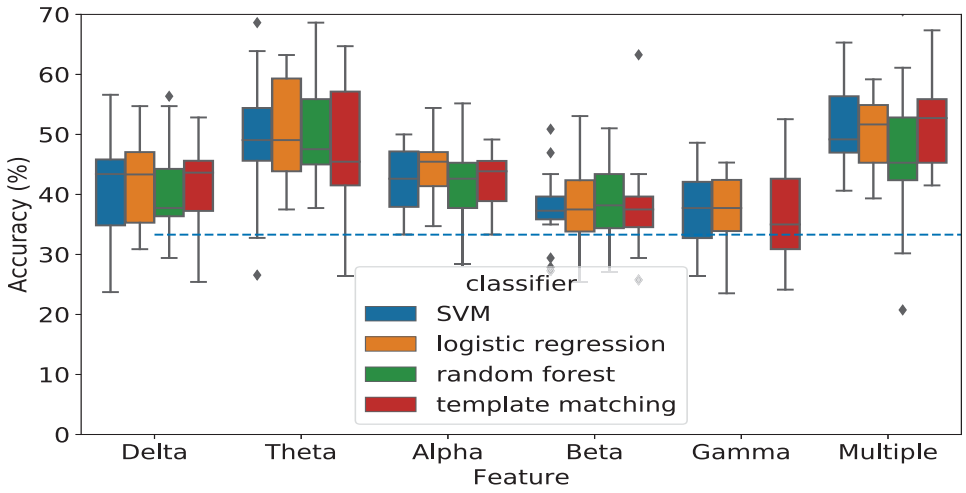

Figure 19. Classification accuracies per classifier and feature in subject-inclusive classification. Each box represents accuracies from all folds. Horizontal line is 33.3% chance level.

(i.e., chance level: 33.3), I performed one-sample t-test on accuracy from each participant. As a result, all models achieved significant results except for logistic regression trained by beta and template matching trained by gamma. The classification accuracies which were significantly above the chance rate (33.3%) were marked with asterisks (*) in Table 9.

To evaluate the effect of classifiers and features on the accuracy, I constructed GLMMs again by using the number of correct classifications as response variables. As a result, while I found a statistically significant effect of features [$\chi^2(5) = 265.4$, $p<0.01$], there was no main effect of classifiers [$\chi^2(3) = 44.1$, $p<0.01$] . Multiple comparisons among classifiers and features were summarized in Table 10 and Table 11, respectively. More importantly, models trained by multiple frequency bands feature showed better performances than models trained by theta.

To further explore the effects of classifiers on the accuracies, I performed multiple comparisons of classification accuracies obtained from classifiers trained by features from which the best performance was obtained (i.e, multiple frequency bands feature) using the Tukey-Kramer method. As results, all comparisons did not show any statistically significant differences ($p>0.5$).

Table 6. Mean accuracies across participant per classifier and feature type in subject-inclusive classification. Sample standard deviations are given in parentheses. Best accuracy is shown in bold.

	Delta	Theta	Alpha	Beta	Gamma	Multi
SVM	40.8* (8.0)	49.2* (9.7)	42.7* (5.2)	37.6* (5.9)	37.4* (6.2)	51.7* (7.1)
LR	41.8* (6.8)	50.6* (8.3)	44.3* (5.2)	38.7* (6.7)	37.6* (5.8)	50.6* (6.4)
RF	40.8* (7.4)	49.3* (8.4)	42.0* (6.2)	39.0* (6.0)	37.2* (6.1)	46.3* (11.0)
TM	41.5* (7.2)	48.0* (10.0)	42.5* (4.8)	37.8* (7.8)	37.0 (7.7)	**51.9* (7.5)**

Multi: Multiple frequency bands feature, LR: Logistic regression, RF: Random forest, TM: Template matching, * $p<0.05$

Table 7. Multiple comparisons for classifier types in subject-inclusive models

	Estimate (S.E)	p-value
Logistic regression -vs- Template matching	0.033 (0.037)	0.81
SVM -vs- Template matching	0.005 (0.037)	0.99
Random forest -vs- Template matching	-0.031 (0.037)	0.84
SVM -vs- Logistic regression	-0.028 (0.037)	0.87
Random forest -vs- Logistic regression	-0.064 (0.037)	0.31
Random forest -vs- SVM	-0.036 (0.037)	0.77

S.E denotes standard error.

Effect of use of spatial filter: FBCSP

To investigate the effects of FBCSP on accuracy improvement, SVM was trained using features extracted by the FBCSP procedure. As results, the average classification accuracies of the model were 25.8±15.1, 51.1±7.4, and 49.0±4.8 in subject-dependent, -inclusive, and -independent classification, respectively. Whereas the subject-dependent model did not reach the chance level in the mode trained by FBCSP features, the subject-inclusive and independent model significantly outperformed the chance rate ($t(16)=9.67$, $p<0.01$, $t(16)=13.02$, $p<0.01$; one-sample t-test against the chance rate).

In order to confirm whether use of spatial filter leads to performance gain in the classification task, I compared the classification accuracy to the best model

Table 8. Multiple comparisons for feature types in subject-inclusive models

	Estimate (S.E)	p-value
Theta -vs- Delta	0.338 (0.452e-01)**	<0.01
Alpha -vs- Delta	0.075 (0.454e-01)	0.56
Alpha -vs- Theta	-0.263 (0.450e-01)**	<0.01
Beta -vs- Delta	-0.125 (0.458e-01).	0.07
Beta -vs- Theta	-0.463 (0.455e-01)**	<0.01
Beta -vs- Alpha	-0.200 (0.457e-01)**	<0.01
Gamma -vs- Delta	-0.154 (0.459e-01)*	0.01
Gamma -vs- Theta	-0.492 (0.456e-01)**	<0.01
Gamma -vs- Alpha	-0.229 (0.458e-01)**	<0.01
Gamma -vs- Beta	-0.029 (0.462e-01)	0.99
Mult -vs- Delta	0.372 (0.452e-01)**	<0.01
Mult -vs- Theta	0.034 (0.448e-01)	0.97
Mult -vs- Alpha	0.297 (0.450e-01)**	<0.01
Mult -vs- Beta	0.497 (0.455e-01)**	<0.01
Mult -vs- Gamma	0.526 (0.456e-01)**	<0.01

S.E denotes standard error. ** p<0.01, * p<0.05, . p<0.1.
Multi: Multiple frequency bands feature

in subject-dependent, -inclusive, and -independent models (i.e., SVM trained by theta, template matching trained by multiple frequency bands, SVM trained by multiple frequency bands, respectively). Fig. 21 summaries the classification accuracies of the best models and models trained by FBCSP features in subject-dependent, -inclusive, and -independent models. As results of two-sample t-tests, in case of the subject-dependent model, the FBCSP showed significant worse performances (t(16)=-5.23, p<0.01) because the FBCSP model did not reach the chance rate. In the cases of the subject-inclusive and -independent model, the statistically significant difference was not found (t(16)=-0.30, p=0.77 and t(16)=-0.80, p=0.43, respectively).

Comparisons among subject-dependent, -inclusive, and -independent

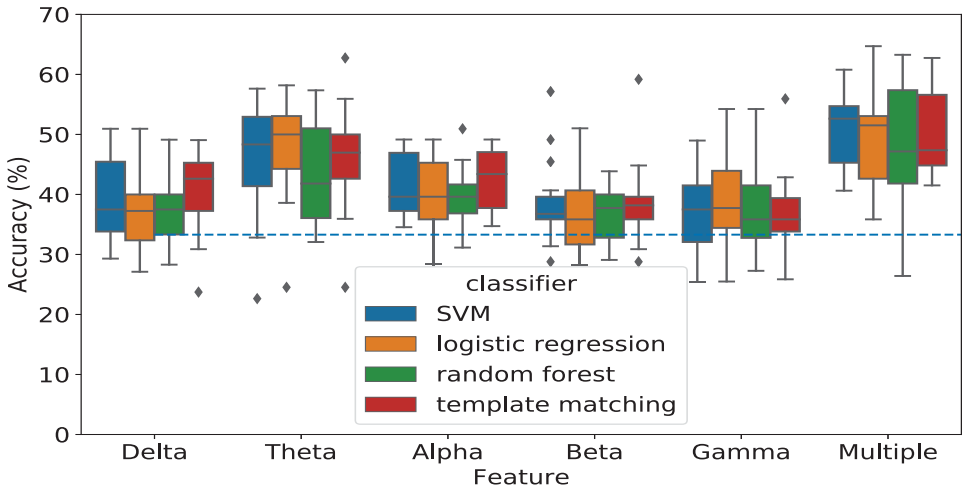

Figure 20. Classification accuracies per classifier and feature in subject-independent classification. Each box represents accuracies from all folds. Horizontal line is 33.3% chance level.

models

To investigate the effect of subject-dependency on classification performances, I performed multiple comparisons of classification accuracies obtained from subject-dependent, -inclusive, and -independent classification using the Tukey-Kramer method. As results, both subject-inclusive and -independent outperformed the performances of subject-dependent models (p<0.01). Besides, subject-inclusive models showed significantly better performances than the subject-independent models (p<0.05).

Next, I compared the best classification performances between subject-dependent, -inclusive and -independent models (SVM trained by theta, template matching trained by multiple frequency bands, and SVM trained by multiple frequency bands). As a result of multiple comparisons using the Tukey-Kramer method, a statistically significant difference was not found (p>0.05).

4.6 Discussion of Experiment 1

In Research 1, I investigated the performances of sentence classification based on single-trial EEG phase patterns during speech processing. The research aimed to

Table 9. Mean accuracies across folds per classifier and feature type in subject-independent classification. Sample standard deviations are given in parentheses. Best accuracy is shown in bold.

	Delta	Theta	Alpha	Beta	Gamma	Multi
SVM	38.8* (6.7)	46.0* (9.1)	41.2* (5.1)	38.3* (6.8)	36.7* (6.2)	**50.5*** (6.1)
LR	37.1* (6.7)	47.9* (7.9)	39.5* (6.2)	36.4 (6.5)	38.6* (8.2)	49.8* (7.5)
RF	37.1* (5.2)	44.5* (8.1)	39.9* (4.7)	36.8* (4.9)	37.4* (6.9)	47.8* (9.8)
TM	40.3* (6.7)	46.0* (8.5)	42.3* (5.0)	38.5* (6.5)	36.7 (6.6)	50.4* (6.8)

Multi: Multiple frequency bands feature, LR: Logistic regression, RF: Random forest, TM: Template matching, * $p<0.05$

Table 10. Multiple comparisons for classifier types in subject-independent models

	Estimate (S.E)	p-value
Logistic regression -vs- Template matching	-0.034 (0.372e-01)	0.798
SVM -vs- Template matching	-0.019 (0.372e-01)	0.959
Random forest -vs- Template matching	-0.074 (0.373e-01)	0.197
SVM -vs- Logistic regression	0.015 (0.373e-01)	0.977
Random forest -vs- Logistic regression	-0.040 (0.374e-01)	0.711
Random forest -vs- SVM	-0.055 (0.373e-01)	0.453

S.E denotes standard error.

answer three research questions: (1) How accurately do EEG-based neural decoding models utilizing neural phase synchronization classify three Japanese spoken sentences in subject-dependent, subject-inclusive, and subject-independent manners?, (2) Which of the three classifiers improved classification accuracy over template matching (the baseline classifier)?, and (3) Do the proposed features including phase patterns in different frequency bands contribute to improve classification accuracy?

Before answering the research questions, I discuss the results of phase synchronization with speech. In the current experiment, PLVs between band-pass filtered EEG data and speech were calculated in multiple frequency bands: delta, theta, alpha, beta, and gamma. The PLV results showed significant phase syn-

Table 11. Multiple comparisons for feature types in subject-independent models

	Estimate (S.E)	p-value
Theta -vs- Delta	0.332 (0.455e-01)**	<0.01
Alpha -vs- Delta	0.106 (0.458e-01)	0.188
Alpha -vs- Theta	-0.226 (0.453e-01)**	<0.01
Beta -vs- Delta	-0.034 (0.462e-01)	0.977
Beta -vs- Theta	-0.366 (0.456e-01)**	<0.01
Beta -vs- Alpha	-0.140 (0.460e-01)*	<0.05
Gamma -vs- Delta	-0.035 (0.462e-01)	0.974
Gamma -vs- Theta	-0.367 (0.456e-01)**	<0.01
Gamma -vs- Alpha	-0.141 (0.460e-01)*	<0.05
Gamma -vs- Beta	-0.001 (0.463e-01)	1.000
Multi -vs- Delta	0.475 (0.455e-01)**	<0.01
Multi -vs- Theta	0.143 (0.449e-01)*	<0.05
Multi -vs- Alpha	0.369 (0.452e-01)**	<0.01
Multi -vs- Beta	0.509 (0.456e-01)**	<0.01
Multi -vs- Gamma	0.510 (0.456e-01)**	<0.01

S.E denotes standard error. Multi: Multiple frequency bands.
** p<0.01, * p<0.05.

chronization in fronto-central regions in the theta frequency band. This coincides with previous phase synchronization research, which has been reported in many research (see [54, 63] for reviews). However, in the current research, right hemisphere lateralization was not observed as the AST theory proposed [66]. This null effect of the right lateralization might be due to the worse spatial resolution of EEG, which signals from several brain regions were overlapped each other on the scalp. If such lateralization is observed, a spatial restriction (i.e., using only signals from the right hemisphere electrodes) might work successfully as feature selection.

The PLVs in alpha, beta and gamma showed similar patterns to theta PLV topographies while they were no significant effects. PLVs in the alpha might be derived from the speech used in the current experiment because the duration of

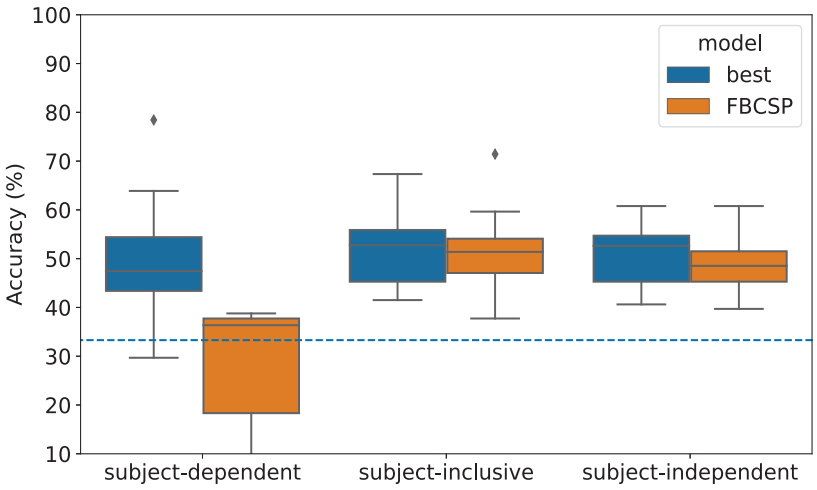

Figure 21. Boxplots of classification accuracies of the best model in subject-dependent and subject-independent model and FBCSP classification. Horizontal line is 33.3% chance level.

mora included in the speech stimuli was extended over alpha frequency bands (see Fig. 11). The PLVs in gamma coincide with Luo and Poeppel (2012) [50] showing that phase synchronization between low-gamma neural oscillations and acoustics. Some neurophysiological models postulate the nested phase synchronization across theta, beta and gamma oscillations (e.g., Ghitza (2011) [28]). Although so far, less evidence of neural phase synchronization in beta oscillations have obtained, this result might partially support such speech perception models.

Contrary to the previous phase synchronization research, there was no tendency of phase synchronization in the delta. The delta synchronization with speech intonation has been observed [13, 22]. Besides, it was reported that the delta oscillation phases are modulated by the existence of syntactic boundaries [22, 55]. The fact that the delta phase synchronization was not observed in the current experiment might be related to nature of the speech stimuli used in the current experiment because the speech stimuli did not include any explicit intonation boundary. Besides, while the speech stimuli includes several subordinate clauses, which might induce phase synchronization at the syntactic boundary (e.g, Anataga kinou muchuude yondeita hon in stimulus 1; *The book that you were ab-*

sorbed in yesterday), the number of the phrase boundaries might not be enough to induce the significant PLV effects. Regardless of that, classifiers showed a significant classification accuracies compared to chance level when using delta phase patterns (see Table 3 and Table 9). Thus, it is considered that phases in the delta oscillations have enough amount of information to discriminate spoken sentences.

As for research question (1), the best performance in subject-dependent models was obtained from SVM trained by theta phase patterns, achieving 50.0% accuracy. The result coincides with the previous MEG classification research [49, 38] showing similar classification performances (around 50% accuracy) using theta oscillations. The best performances of subject-independent models, which was obtained from SVM trained by multiple frequency bands feature, achieved a significant above-chance rate accuracy (50.5%). Thus, the current experiment demonstrated that the following things:

- EEG phase information induced by neural phase synchronization can classify speech significantly with around 50% accuracy.
- This phase-based classification model has generalizability to other different users.

In the comparisons of classification performances across subject-dependent, -inclusive, and -independent classifications, the subject-dependent models showed significantly worse accuracies compared to subject-independent models. As a reason, it is considered that the number of training data in the subjet-dependent classification fewer than -independent. On the other hand, the subject-inclusive models significantly outperformed the subject-independent classifications. Besides, the best classification performances of the experiment were obtained from the subject-inclusive models. Thus, while the phase patterns induced by phase synchronization enable the speech to be classified in subject-independent manners, the classification performances can be improved when the classifier learns the subject-specific EEG patterns.

The current classification performances showed similar performances in the previous MEG-based classification (0.40 – 0.57; [38]). This result was a pleasant surprise because given that the previous neural decoding of the speech showed

that the performances of the EEG-based decoding were worse than the MEG-based one [16], I had expected that the classification performances did not reach such similar performances to MEG-based classification. On the other hand, considering that phase synchronization has been observed in EEG experiments (e.g., [88]), it might be a reasonable result to achieve similar performances to MEG. Of course, it is impossible to compare directly the previous MEG-based classification and the current EEG-based one because of several differences due to differences of language (Japanese vs English) and contents of spoken sentences.

The current research also demonstrated that this decoding method could apply to unknown users. Such subject-independency had predicted because less inter-subject variability had already been demonstrated by Kerlin et al (2010). [41]. As stated in Chapter 4.1, such subject-independency has advantages in terms of BCI application. Thus, this subject-independency emphasizes merit of using phase synchronization as features for neural decoding of speech.

As for the research question (2), the best classification performances were obtained from SVM in both subject-dependent and -independent classification and template matching in subject-inclusive model. While SVM marginally significantly outperformed template matching (baseline) in subject-dependent classification, an effect of classifiers was disappeared in the subject-inclusive and -independent classification. Besides, the significant performance difference across classifiers trained by the best features was not found in any subject-dependent, -inclusive, and independent classification. Thus, it seems to be difficult to insist that SVM is superior to template matching. Rather, taking the results of feature importance topographies in the subject-independent classification into account, template matching is more suitable to capture phase synchronization. One question is: why did template matching succeed in capturing the synchronization? This might be because phase patterns of EEG and speech come to be closer to each other during neural synchronization with speech, thus, the distance-based classification is easier to capture this phenomenon. One clear difference of performances among classifiers is that random forest showed worse performances than other classifiers in the subject-dependent classification. However, the effect was not observed in the subject-independent classification. Given that amount of data in the subject-dependent classification is fewer than the subject-independent one,

random forest in subject-dependent classification fell into overfitting.

Thus, the answer to the research question 2 is:

- There is no significant classifier effect on classification accuracy improvement among the classifiers used in the current experiment.

As for the research question (3), while subject-dependent and -inclusive models did not show the performance improvements by use of multiple frequency bands feature compared to theta features, the subject-independent classification showed a significant accuracy improvement in the multiple frequency bands feature compared to only theta feature. This disagreement is difficult to interpret, however, at least, given that the subject-independent classification showed every frequency bands showed significant classification performances compared to chance rate, phase information in the frequency bands relevant to phase synchronization might contribute to the improvement of classification accuracy under the situation where enough training samples are available.

- Use of phase patterns in the multiple frequency bands relevant to phase synchronization improves classification accuracy in the subject-independent models

However, as the frequency band is higher, the S/N ratio tends to get worse and worse. Thus, whether the phase information in the higher frequency band is discriminative for decoding in the real-life situation need to be investigated.

Finally, to utilize spatial information, features of multiple frequency bands were extracted using FBCSP. However, the features did not show improvement of classification accuracy in both subject-dependent and subject-independent classification. The CSP method is mainly utilized in the EEG-based classification of motor imagery tasks such as imaginary of the left hand, right hand, and foot movements. A CSP method shows effectiveness in such tasks because the motor imagery induces differences of topographical patterns depending on the parts of the body. On the other hand, it is expected that the topographical pattern is similar across speech types because a neural source of phase synchronization does not differ depending on the contents of the speech. This similar topographical patterns across stimulus might hinder the effect of the spatial filter on accuracy.

4.7 Summary of Chapter 4

In Experiment 1, I investigated classification performances of EEG-based neural decoding of speech using phase synchronization during speech perception. Besides, for performance improvement, I tested the effects of different classifiers and features in the multiple frequency bands relevant to phase synchronization. I obtained the following results from the experiment.

- EEG phase information induced by neural phase synchronization can classify three types of speech with around 50% accuracy.
- This phase-based classification model has generalizability to other different users.
- There were no classifier effects on improvement of classification performances, but taking into account of feature importance topographies, template matching might be more suitable classifiers to capture phase synchronization.
- Use of phase patterns in the multiple frequency bands relevant to phase synchronization improves classification accuracy in the subject-independent models

5. EEG phase synchronization with imagined speech

5.1 Purposes of Experiment 2

Experiment 1 had demonstrated the effectiveness of EEG phase patterns for discriminating speech. The next step of the thesis is to apply this phase-based classification to the imagined speech. If phase patterns are modulated depending on a rhythm of the imagined speech, EEG phase patterns also should discriminate the imagined speech. However, as mentioned previously, the modulation of neural oscillations during performing the imagined speech task has not clarified well yet. Thus, the purpose of Experiment 2 is to investigate whether phase synchronization with EEG during the imagined speech and rhythm of the imagined speech.

Recent studies have revealed neural correlations of imagined speech generation: namely, a similar network as the speech production including the left inferior frontal region and the left premotor cortex [67]. Tian and Poeppel (2010) [76] proposed an internal forward model during the imagined speech, where a motor efference copy is sent from the motor planning region to the parietal cortex and a further efference copy is sent from the parietal cortex to the temporal cortex. These parietal/temporal areas activated by the two types of efference copy generate a kinesthetic feeling and auditory perceptual feelings, respectively.

In contrast to such research on neural correlations of imagined speech, the modulation of neural oscillations during imagined speech has not received much attention. Rather than imagined speech, so far, intensive research on neural dynamics has been conducted in the neurophysiological speech perception field. Many studies have shown that theta oscillations (4–8 Hz) in the auditory cortex match their phases to the amplitude envelope of speech during speech processing (see Chapter 3). It has been suggested that this synchronization is based on endogenous theta oscillations in Heschl's gyrus in the right hemisphere dominantly [30].

A similar endogenous fluctuation in theta has also been observed in the part of the ventral premotor cortex related to control of the mouth [30]. Given that the amplitude envelope of speech is mainly derived from vowels voiced by the mouth opening, the speech envelope could conceivably be related to an oscillatory

rhythm in the mouth premotor cortex in a similar way to synchronization in speech processing during perception. Thus, by considering this together with the fact that motor-related regions such as the premotor cortex can be activated by motor imagery [27], it can be hypothesized that neural oscillations during speech imagery synchronize with speech rhythms generated by the imagery. Although it is difficult to observe imagined speech directly, overt speech can be regarded as a counterpart of imagined speech because the phonetic features of imagined speech are similar to those of overt speech [25].

In addition to phase synchronization analysis, I also focused on whether EEG during imagined speech can classify speech stimuli with different speech envelopes as well as Experiment 1. It has demonstrated that neural oscillation phases during synchronization with perceived speech enable speech stimuli to be classified because the neural synchronization induces replicable and stimulus-specific phase patterns of oscillations across trials. Conversely, reliable EEG-based classification of speech stimuli with different speech envelopes suggests that a replicable and stimulus-specific neural phase pattern is induced by imagining the articulatory movements of the speech. Thus, classification accuracy with above-chance level supports evidence for EEG synchronization during the imagined speech.

In sum, Experiment 2 aimed to answer the following research questions:

1. whether do EEG oscillations during imagined speech synchronize with the speech envelope of the overt counterpart?

2. whether can EEG oscillations during imagined speech classify speech stimuli with different amplitude envelopes?

To this end, I regressed the overt speech envelope using EEG and calculated correlation coefficients between the EEG-based regressed envelope and the overt speech envelope. The classification was based on the template matching method which calculates the distance between a test data and a template waveform of each class because Experiment 1 suggests that this method is suitable to capture phase synchronization. Since the duration of the imagined speech was expected to vary across trials, I used a DTW method to correct the durational variability for the classification analysis. To the best of my knowledge, this is the first

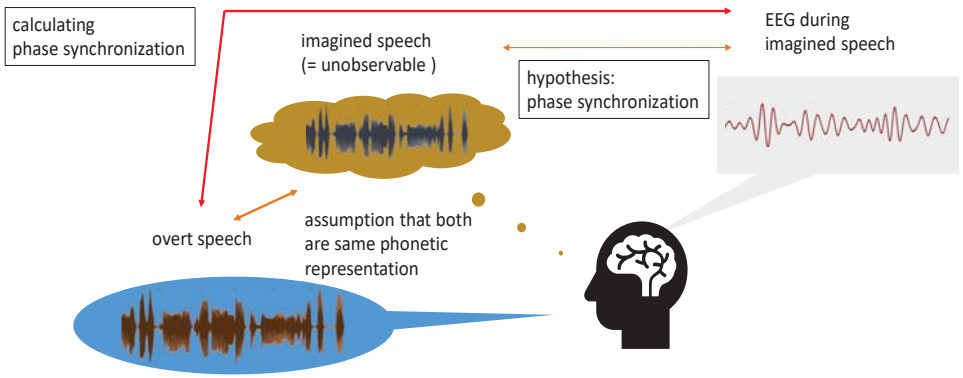

Figure 22. An overview of Experiment 2. I hypothesized that the rhythm of imagined speech and EEG oscillations in the corresponding frequency band during the imagination are synchronized each other. If this hypothesis is true, a similar classification method to Experiment 1 can be applied to an imagined speech classification task.

study to investigate synchronization between the overt speech envelope and EEG oscillations during the imagined speech.

5.2 Methods of Experiment 2

5.2.1 Participants

Eighteen right-handed L1 Japanese speakers participated in the experiment (6 female, 12 male, mean age: 23.8±1.7). They all gave written informed consent to their participation. No participants reported a history of hearing impairment or neurological disorders. The experiment was approved by the ethical review board of the Nara Institute of Science and Technology.

5.2.2 Experimental materials

I recorded three speech stimuli from a female L1 Japanese speaker in a sound-attenuated chamber (44.1 kHz/16 bit). All speech stimuli were nonsense sounds because I wanted to avoid the effect of semantic processing of speech on the synchronization analysis. All stimuli consisted of three [ba] and two [ba:] with a prolonged vowel at different positions to differentiate speech envelopes (stimulus

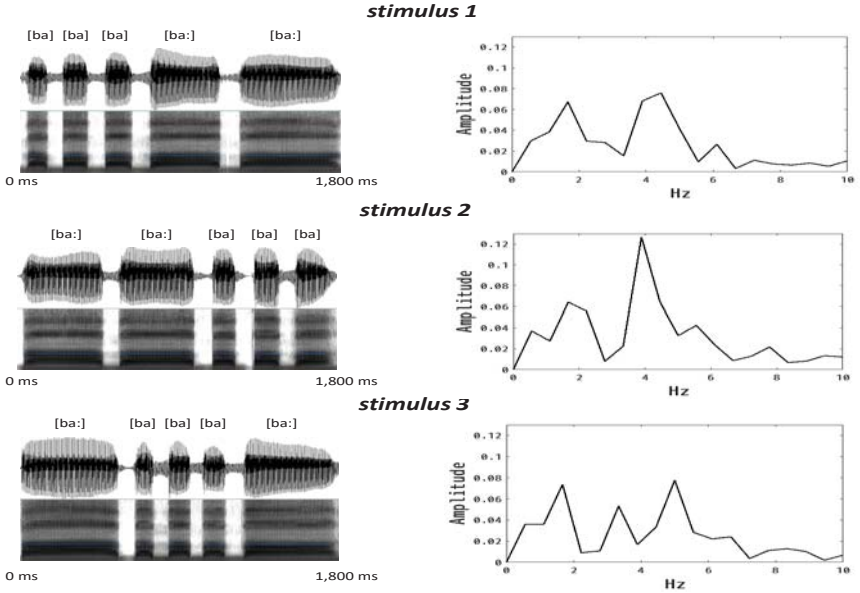

Figure 23. (Left) Waveforms and spectrogram of speech stimuli. (Right) Amplitude spectrum of speech stimuli.

1: [ba] [ba] [ba] [ba:] [ba:], stimulus 2: [ba:] [ba:] [ba] [ba] [ba], stimulus 3: [ba:] [ba] [ba] [ba] [ba:], Fig. 23). The duration of each stimulus was adjusted to 1,800 ms and the volume was normalized. The pitch height was adjusted to 200 Hz by using Praat [12].

5.2.3 EEG recordings

EEGs were recorded with an amplifier (BrainAmp DC, Brain Products GmbH., Germany) from 32 Ag/AgCl electrodes (actiCAP, Brain Products GmbH., Germany). The impedance of the electrodes was kept below 10 kΩ. The EEGs were online-filtered with a 0.016-Hz high-pass and 250-Hz low-pass filter. The sampling rate was 1,000 Hz. An FCz electrode and FPz electrode were used for the reference and ground, respectively. The experiment was controlled using Presentation software (Neurobehavioral Systems, Inc., U.S.A).

Participants sat on a comfortable chair in a dimly lit sound-attenuated chamber. A display monitor, keyboard, and microphone were placed on a desk in front of the chair. Participants were instructed to familiarize themselves with the

speech stimuli and memorize them before the EEG recording. One trial consisted of three tasks: listening, speaking, and imagining speech. During the trial, they were instructed to stay as still as possible. In the imagined speech task, participants were forbidden to move any articulators such as their mouth or lips and were instructed to imagine the articulatory movements of the speech stimulus without actually making those movements.

The experiments consisted of practice and three main blocks, where each block consisted of 21 and 20 trials. Each stimulus was presented to each participant 20 times in a randomized order in the main blocks. The procedure of a trial did not vary between the practice and the main blocks. Each trial started from the appearance of *Ready?* on the display. A procedure of one trial was as follows.

1. Participants initiated a trial by pushing the space key on the keyboard.
2. *LISTEN* was displayed for indicating the task type for 2,000 ms.
3. After a countdown (from 3 to 1) to the start of task execution, a fixation mark (+) appeared, and at the same time, a speech stimulus was played via headphones.
4. In the speaking task, the task indication (*SPEAK*) was followed by a countdown to task execution.
5. After the countdown, participants uttered the speech stimulus that was presented in the listening task at the same speech rate. Participants' utterances were recorded using a microphone.
6. To reduce variation in the duration of uttered speech across trials, I controlled the timing of the start and the end using a progress bar. The progress bar appeared on the display immediately after the countdown, gradually extended horizontally during the imagination task, and stopped at 1,800 ms, which was the same duration as the speech stimuli, relative to the appearance of the progress bar. By having the participants initiate and finish their utterances at the same time as the appearance and stop of the progress bar, respectively, I was able to mitigate duration variabilities across trials.

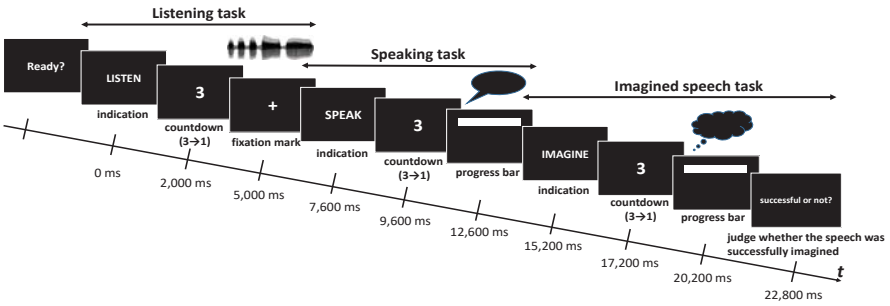

Figure 24. Procedure of an experimental trial. In the listening task, the stimulus was played after the task indication followed by a countdown to task execution. In the speaking task, participants uttered the speech stimuli into a microphone. In the imagined speech task, they imagined the articulatory movement of speech stimuli without making movements. After the imagined speech task, participants reported whether they had successfully imagined the speech by pressing a button.

7. In the imagined speech task, after the task indication (*IMAGINE*), the procedure was the same as the speaking task except that participants imagined articulatory movements of the speech stimulus. Variabilities of the duration of the imagined speech were also adjusted by using the progress bar.

8. After every imagination task, the participants were asked to press a button to indicate whether they had been able to imagine the speech (success: F key, failure: J key).

The procedure of one trial is summarized in Fig. 24. The experiment continued for about 45 minutes.

5.2.4 Preprocessing of EEG data

We analyzed synchronization and its topographical patterns in both perceived speech and imagined speech: synchronization (1) between the amplitude envelope of the speech the participants listened to and EEGs during the time they perceived it and (2) between the amplitude envelope of the speech the participants uttered and the EEGs during the imagined speech.

We used the FieldTrip toolbox [60] for MATLAB (The MathWorks, Inc., U.S.A) for the EEG data analysis. To remove slow drift artifacts, a 4096th order

FIR one-pass zero phase high-pass filter at 0.5 Hz (Hamming window) was applied to the continuous data. EEGs were re-referenced to an average of both mastoids and epoched from –1,000 ms to 3,000 ms relative to the task onset. The task onset was set to the appearance of the fixation mark (in the listening task) and the progress bar (in the imagined speech task). Epochs with large amplitudes exceeding 200μv were rejected. Data from FP1 and FP2 were exempted from this rejection because they included large eye-related artifacts that were later removed during the ICA. Epochs contaminated by large muscle artifacts were identified using an automatic detection method based on z-value of the data distribution (cutoff = 15) and visual inspection (see Chapter 4.2.4). ICA was used to correct the eye-related artifacts and remaining muscle artifacts. Candidates of eye-related ICs were searched based on the average Pearson correlation coefficients between the FP1/FP2 data and ICs. ICs to be removed were identified by visually inspecting their waveforms and spatial distributions.

We separated the EEG datasets on a per condition basis. In the imagined speech dataset, I excluded trials in which participants reported that they had not successfully imagined the speech and trials in which participants uttered the speech incorrectly, such as through a slip of the tongue in the preceding speaking task. The incorrect utterances were annotated manually. One participant was removed from the analysis because of a large total number of rejected trials across conditions (above 30%). As a result, the average total of rejected trials across conditions was 8.7%±5.0. A one-way repeated analysis of variance (ANOVA) test showed no significant differences in the number of rejected trials between speech stimuli (F (2, 32)=1.47, p=0.25).

5.2.5 Analysis pipeline

I analyzed synchronization and its topographical patterns in both perceived speech and imagined speech: synchronization (1) between the envelope of the speech the participants listened to and EEGs during the time they perceived it and (2) between the overt speech envelope uttered by participants and the EEGs during the imagined speech. An analysis pipeline is described in Fig. 25.

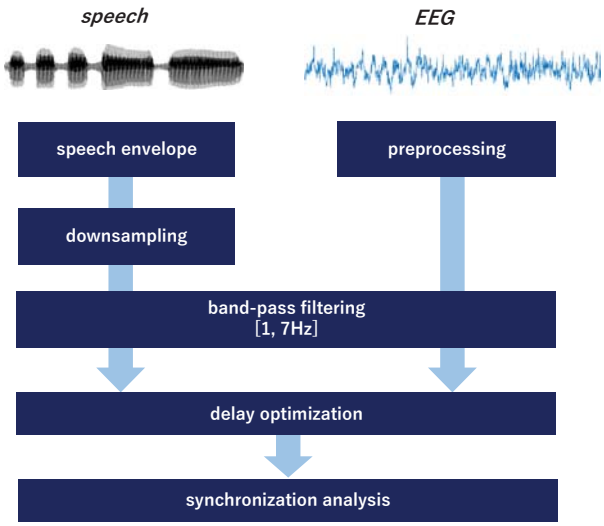

Figure 25. Analysis pipeline for calculating synchronization between speech and EEG.

5.2.6 Speech envelope extraction and band-pass filtering

The amplitude envelope of the speech was extracted using a Hilbert transformation. A two-pass IIR Butterworth band-pass filter of the 8th order (1–7 Hz) was applied to both the speech envelope and the preprocessed EEG, which was further extracted from 0 to 3,000 ms. The frequency ranges of the band-pass filter were decided to extract the low-frequency modulation (see Fig. 23; peaks in the frequency domain were observed around 2 and 5 Hz). To avoid filter artifacts, the flipped data were concatenated to the beginning and the end of the data and were removed after the filtering procedure. The speech envelope was downsampled with the same sampling rate as EEG (i.e., 1,000 Hz).

5.2.7 Optimizing the delay in synchronization

We corrected the delay in synchronization between the EEGs and the speech envelope because a certain delay can be expected (e.g., milliseconds for participants to recognize the appearance of the progress bar and begin imagining the speech). The cross-correlation coefficients between all concatenated trials of the band-pass

filtered EEG and speech envelope were calculated per EEG electrode and then averaged across electrodes. Before calculating the coefficients, both data were demeaned. I searched for the peak showing the highest coefficients within the lag range [10, 200 ms] for perception data and [180, 500 ms] for imagined speech data. A narrower lag range was adopted for perception data because less variability in the delay was expected than for the imagined data, where participants needed to respond to the appearance of the progress bar to initiate the imagined speech. In contrast, the relatively wide range of the lag in the imagined speech was set because larger variability of the reaction times across participants was expected in the imagined speech. The lag range of the imagined speech was chosen with consideration of the fact that human simple reaction time is generally around 200 ms [78]. After each EEG trial was shifted in the time domain by this delay per participant data, EEG data from the new onset time point to 1,800 ms, which was the same duration as the speech stimulus, were used for further analysis of the perceived data. In the case of the imagined speech data, data from the new onset time point to the duration of the corresponding overt speech (i.e., overt speech data in the speaking task immediately before the imagined speech) were extracted under the assumption that the duration of the overt speech and imagined speech were similar.

5.2.8 Synchronization analysis

A multiple linear regression method [7] was used for analyzing the synchronization between the EEG and the speech envelope. Specifically, I used a concatenated, delay-optimized, band-pass filtered EEG matrix $M^{n\times m}$ across trials, where n is the total number of data points and m is the number of electrodes. Each electrode data was demeaned by subtracting the electrode average value. First, the number of dimensions was reduced using principle component analysis (PCA) to avoid collinearity. M was projected into space spanned by eigenvectors of the covariance matrix of M covering 99% of the variance. The projected M and space spanned by eigenvectors were expressed by M_k and P_k, respectively. After the concatenated speech envelope across trials, denoted by, was demeaned, the envelope was modeled by a linear combination of the columns of M_k and noise ϵ:

$$s = M_k b + \epsilon. \tag{22}$$

The optimal coefficients b were estimated using the least-squares method. The regressed speech envelope s is expressed by

$$s = M_k b. \tag{23}$$

The spatial pattern of b on the topography (P_{eeg}) was calculated on the basis of [7] and [61]. First, the pattern P_{pca} was calculated as

$$P_{pca} = \frac{M_k^T \hat{s}}{\hat{s}^T \hat{s}}. \tag{24}$$

From Eq. 24, P_{pca} can be regarded as the coupling between the regressed speech envelope s and the projected EEG data M_k [7]. Finally, P_{eeg} was calculated using the previous PCA space P_k:

$$P_{eeg} = P_k P_{pca}. \tag{25}$$

In terms of visualizing the topography, the absolute value of each participant's P_{eeg} was normalized to have a certain range [0, 1] (1 represents the maximum coefficients of the synchronization). I calculated the coefficients of the Spearman rank correlation (Spearman's rho) between the EEG-based regressed envelope ($\hat{s}$) and speech envelope (s) as an index of the synchronization. Spearman's rho was converted using the Fisher-Z transform to approximate a normal distribution.

5.2.9 Classification analysis

We performed a classification analysis to investigate whether EEG during speech perception and speech imagery included a signature of the envelope of the perceived speech and the overt speech, respectively. Classifiers were trained using the delay-optimized, band-pass filtered EEG data in speech perception and imagined speech task. The classification method was similar to Zhang et al's ECoG-based classification of overt speech [86]. Since I expected the duration of the imagined speech to vary across trials regardless of the duration variability mitigation afforded by the progress bar, each EEG data in the imagined speech task was

realigned using the DTW method, which is an algorithm to find a path that minimizes the distance between two signals. The performance of the classifier was evaluated by leave-one-out cross-validation.

The classification was based on the Euclidean distance between a test data and a template waveform of each class. The templates were constructed by using the training data. To this end, first, training data were separated based on the class labels. In the imagined speech data, each training data was realigned to the envelope of the speech stimulus corresponding to the class label using the DTW to correct duration variability. In the perception data, realignment was not performed because less durational variability was expected. EEG data and the envelope of the speech stimulus data were standardized by using z-score for all data to take values in a similar scale. Each template waveform of the class label was constructed by averaging the training data belonging to the class in the time domain. The class label of the template waveform showing the least square Euclidean distance to a test data was considered the prediction result. In the case of the imagined data, each test data was also realigned to each template using the DTW before the classification and the distance between the realigned test data and the template were calculated because the duration of the test data also differed across trials.

For classification, I used the five electrodes showing the highest absolute values of the coefficients of the EEG pattern among electrodes positioned in the frontal and central region (i.e., Fz, F3, F4, F7, F8, FC5, FC6, FCz, FC1, FC2, Cz, C3, C4, CP5, CP6, CP1, and CP2), as the synchronization was mainly observed in these regions (see Chapter 5.3 Results of Experiment 2). The final prediction of the speech stimuli was decided by a voting system across the results of the five electrodes.

There is a possibility that the classification was performed only based on event-related responses to the perceived speech or the task onset, not based on the neural synchronization with the slow modulations of the perceived speech and imagined speech. To exclude this possibility, I performed an additional classification using data in which the first part of the waveform was excluded. In this classification, when calculating the distances between the test data and each template, the first 150 ms of data in the waveform were ignored.

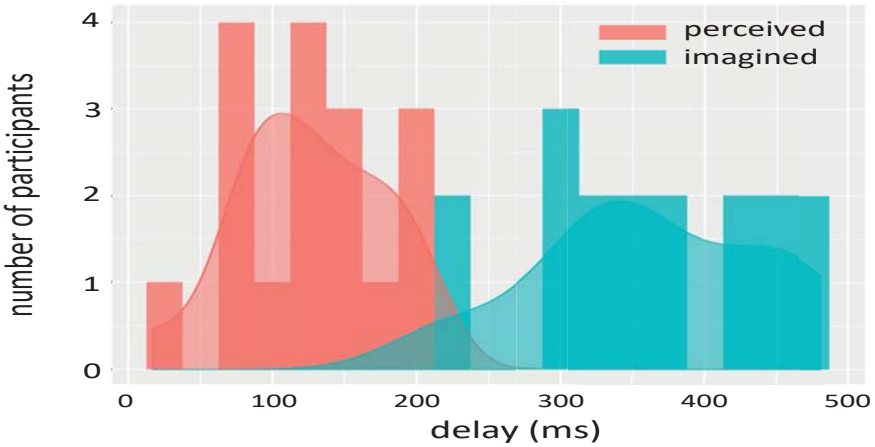

Figure 26. Histograms of estimated delays in perceived and imagined speech. Filled curves represent the densities of the distributions.

5.3 Results of Experiment 2

First, I plotted the histograms of all participants' estimated delays of perceived and imagined speech EEG data (Fig. 26). The average delays across participants were 126 ms (SD=49) and 364 ms (SD=77) for the perceived and imagined speech, respectively. As for the synchronization analysis, the Spearman's rho averaged across participants was 0.15 (SD=0.04) and 0.10 (SD=0.03) for perceived and imagined speech, respectively. One sample t-test revealed that both Spearman's rhos significantly differed from zero (perceived: $t(16)=14.3$, $p<0.01$, imagined: $t(16)=15.0$, $p<0.01$). Box plots of the Spearman's rho in the perceived and imagined speech are provided in Fig. 27A. I also plotted an example of synchronization between the EEG-based regressed envelope and a corresponding speech envelope (Fig. 27B). Both envelopes were standardized using z-score for visualization. The grand averages of EEG patterns across participants are shown in Fig. 28A. The EEG pattern of the perceived data showed the synchronization at the electrodes in a central region. In contrast, the pattern of imagined data was distributed in a more frontal region. This indicates that the neural generator differs across perceived and imagined speech. In the classification analysis, mean accuracies across participants were 54.7% (SD=10.8) for perceived speech and 38.5% (SD=5.3) for imagined speech (Fig. 28B). One sample t-test revealed

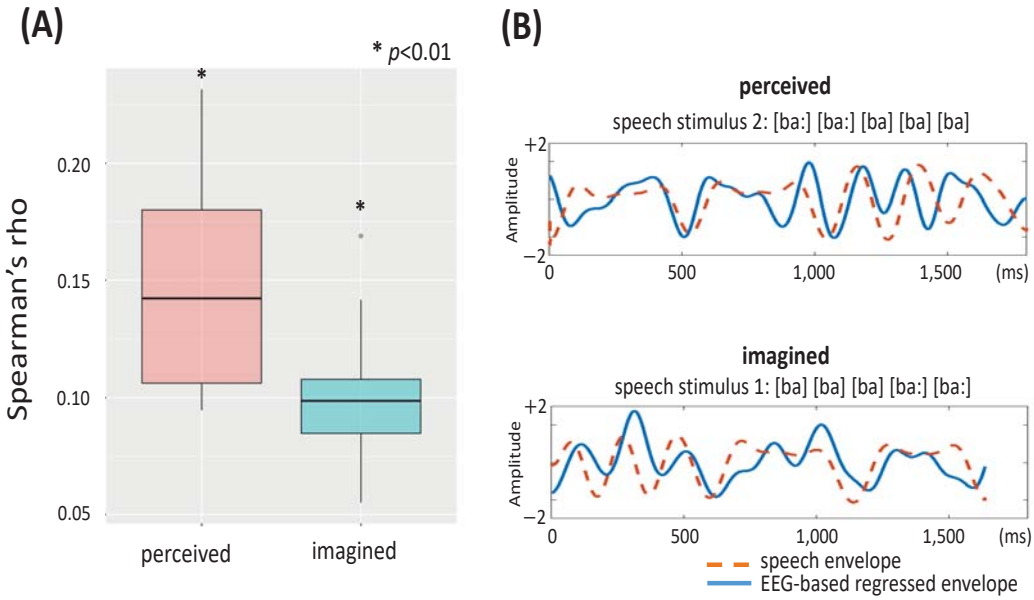

Figure 27. (A) Box plots of Spearman's rho between EEG-based regressed speech envelope and speech envelopes per condition. (B) An example of the EEG-based regressed envelope and the corresponding speech envelope from subject 03 in perceived and imagined speech.

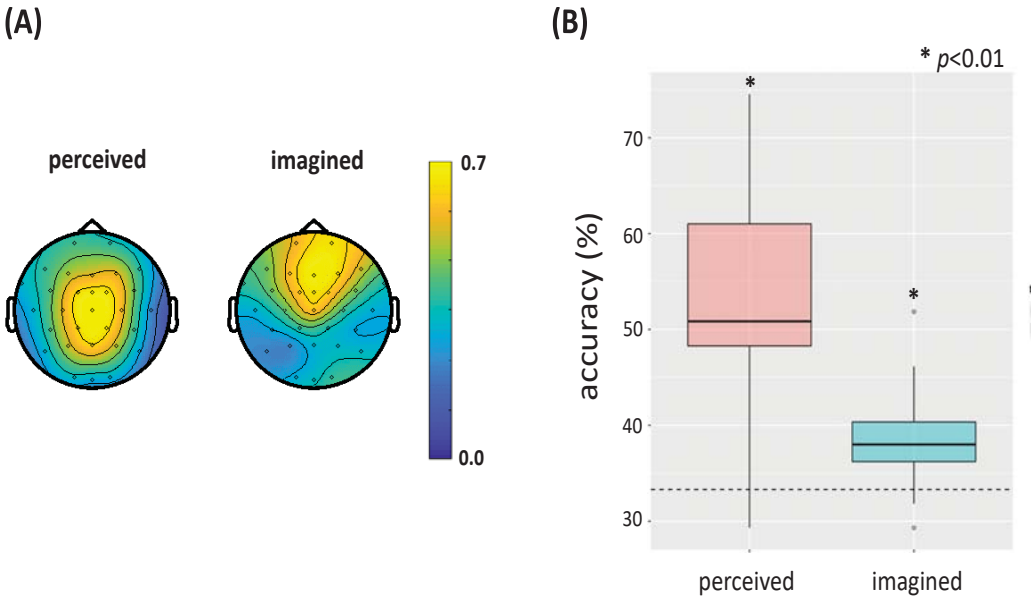

Figure 28. (A) Grand averaged synchronization patterns across participants in perceived and imagined speech. (B) Box plots of accuracies in EEG-based classification of speech stimuli with different amplitude envelopes in perceived and imagined speech. The dotted horizontal line represents the level of chance (33.3%).

that the accuracies were significantly above the 33.3% chance rate (perceived: $t(16)=7.9$, $p<0.01$, imagined: $t(16)=3.9$, $p<0.01$). The accuracies of the classification based on perceived data significantly outperformed the ones based on imagined speech (paired sample t-test: $t(16)=5.2$, $p<0.01$). As for the classification using the data from which the first 150 ms were excluded, data in both the perceived and imagined speech showed similar results to the previous classifications: 53.8% (SD=9.3) for perceived speech and 37.9% (SD=6.0) for imagined speech. Both accuracies significantly outperformed the chance rate (one-sample t-test, perceived: $t(16)=8.8$, $p<0.01$, imagined: $t(16)=3.0$, $p<0.01$).

5.4 Discussion of Experiment 2

The purpose of this research was to investigate the synchronization of imagined speech and neural oscillations during the generated speech imagination. To this end, I substituted participants' overt speech for imagined speech because it is im-

possible to observe imagined speech physically. Multiple regression-based analysis [7] revealed a significant correlation between the EEG-based regressed speech envelope and the overt speech envelope as well as the perceived data. Besides, the classification performances of speech stimuli with different envelopes achieved the accuracy of 38.5% using the EEG during the imagined speech. These results indicate that EEG oscillations during the imagined speech contain the signature of the speech envelope of the overt counterpart of the imagined speech. Thus, I obtained a positive answer to my two research questions.

The EEG patterns of synchronization in the perceived and imagined data were differently distributed over the scalp: to the central region and the frontal region, respectively. Although it is difficult to identify the generator by using EEG due to the low signal-to-noise ratio and volume conduction, the difference in the topographies, at least, indicates that the neural generators of the synchronization differ from each other. In the case of speech perception, the neural source of the synchronization is the auditory cortex [31, 42], while the generator of the synchronization during the imagined speech seems to be in the more frontal parts, for example, in the frontal lobe including the motor-related area.

As mentioned in Chapter 5.1, one candidate of the generator of the imagined speech synchronization is the ventral premotor cortex, as Giraud et al. (2007) [30] revealed the endogenous fluctuation at the theta frequency band in the ventral premotor cortex of the mouth. Considering that the waveform of the speech envelope is formed based on the cycling of the mouth opening, conceivably the endogenous theta oscillations in the region are modulated depending on the timing of the cycling of the mouth opening during the imagined speech.

Assuming that the generator is the ventral premotor cortex, one question is raised: what is the functional role of the phase synchronization in the premotor cortex during the imagined speech? I know that during the periodical auditory stimulus process in the brain, neural oscillations show phase-locked responses to the periodicity of the stimulus [81]. The functional role of the phase-locked responses to the external periodicity is to enable the brain to predict the timing of future input stimuli and process them at the state of high excitability (see [5] for a review) because the neural oscillation phases are related to neuronal excitability [46] and the state of the ongoing oscillatory activity in the pre-stimulus period

affects the processing of sensory stimuli [4, 70]. This phase-locked response is also widely observed during speech perception: synchronization between the speech amplitude envelope, which features pseudo-periodical fluctuations around 4–8 Hz, and the neural oscillation phase in the theta frequency band [1, 49, 38, 50, 64]. Because syllabic information is dominant in the amplitude envelope of speech, it has been suggested that the syllabic information is sampled and segmented via this neural phase synchronization (see [54] for a review). As evidence to support this notation, Hyafil et al. (2015)'s computational model [39] demonstrated that the syllable boundaries can be reliably predicted from neural oscillations during speech perception.

Considering that the phase synchronization during speech perception may be related to the syllable duration, phase synchronization during imagined speech might be also related. The Directions Into Velocities of Articulators (DIVA) model [32, 33] defines the role of the ventral premotor cortex during speech production as a speech sound map that stores a repository of the motor programs of frequently observed production units such as syllables. Together with the DIVA model and the current result, I speculate that during syllable production, a motor program of the syllable is read from the ventral premotor region and the syllable duration (i.e., how long the syllable is pronounced) is encoded in the timing of the neural oscillation phases in the region. Alternatively, Zhang et al (2012) [86] revealed that the ECoG-based classification of overt sentences, which features a classification method similar to the current research, was successful at an electrode corresponding to the Broca area. This result suggests that the Broca region also might show synchronization with the imagined or overt speech envelope. To further investigate the neural source and the functional role of the neural synchronization during the imagined speech, I aim to localize the neuronal source using other brain imaging techniques such as fMRI in the future.

In this study, I used only two types of the syllable ([ba] and [ba:]) to control the segmental information across speech stimuli. However, even if I used different syllables such as [ku] and [ku:] or [de] and [de:] with the same sequence as the current research, I expect a similar result would be observed. This is because I analyzed the similarity between the band-pass filtered EEG and the envelope pattern of overt speech in both synchronization and classification analysis (synchro-

nization analysis: correlation, classification analysis: Euclidean distances) and did not use the information related to the segmental information itself. Thus, if I had used different syllables from [ba] and [ba:], the synchronization would probably be observed as long as the stimuli include a rhythm of the speech envelope. Supporting this, Deng et al. (2010) performed EEG-based classification of trials with different timings of syllable imagination (3 different timings 2 syllables: [ba] and [ku]) [21], showing significant accuracies of three-class of the syllable imagination timing using EEG data of both [ba] and [ku].

Another question is whether the observed synchronization in the current research is specific to language processing or not. It is possible that the synchronization and reliable classification performances were obtained when the participants imagined the rhythms of non-linguistic content, such as beating a drum or blowing a whistle. I have stated that the syllable duration information might be encoded in the low-frequency oscillations in the ventral premotor cortex related to mouth control. If so, it is predicted that the synchronization would not be observed when non-linguistic content, which is not related to controlling mouth movements, is imagined. Alternatively, it might be possible that the synchronization to non-linguistic rhythms might be observed in other motor-related brain regions. Comparing the results of tasks where participants imagine the rhythm of producing syllables and non-linguistic rhythms such as the rhythm of a beat will lead to further understanding of the functional role of synchronization during the imagined speech.

We obtained classification accuracies (38.5%) significantly above the level of chance in classifying speech stimuli with different envelopes using EEG during the imagined speech. This result partially supports the neural synchronization with the overt speech envelope because it indicates that EEG oscillations during imagined speech induce a stimulus-specific, replicable oscillation pattern. While the neural synchronization with speech envelope during speech perception has been applied to M/EEG-based sentence classification [49, 38, 80] , Experiment 2 applied it to the classification of imagined speech. So far, there have been many studies on non-invasive neural decoding of imagined speech performed on the classification of vowels [20], syllables [23], and words [71] (see Chapter 3). As mentioned in Chapter 1, so far, the neurophysiological mechanisms enabling the

neural decoding of imagined speech have not been sufficiently investigated. Experiment 2 successfully provides a novel, neurophysiologically motivated feature for neural decoding of imagined speech.

One of the difficulties in the neural decoding of imagined speech stems from the variable duration of imagined speech, which cannot be observable directly. I demonstrated that duration control using a progress bar and the DTW-based correction inhibited this variability across trials and enabled a reliable classification performance, at least, above the chance level. However, because the classification performances based on the EEG during the imagined speech were significantly lower than those during the perceived data, further considerations to improve the classification accuracy, at least, to achieve a similar performance level to the perceived data, are required.

As an alternative possible explanation of the above-chance level classification performances, one might argue that the trained classifier relied on the speech onset, as opposed to the dynamics of the low-frequency modulation of the speech stimuli or imagined speech, because there is a possibility that the difference in the segmental information of the onset of the speech stimuli (e.g., [ba] and [ba:]) evoked different responses to the event onset. However, when I performed classification using the EEG data with the first 150 ms omitted, the classification performances remained significantly above the level of chance. Thus, in the current classification, I tend to conclude that the whole waveform pattern of EEG, not the onset information, contributed to the classification of speech stimuli with different envelope patterns.

5.5 Summary of Chapter 5

To reveal the effectiveness of neural phase synchronization for neural decoding of the imagined speech, Experiment 2 investigated whether EEG phase patterns during the imagined speech synchronize with a rhythm of the imagined speech, which was replaced by the overt counterpart. I obtained the following results from Experiment 2.

- EEG oscillations during imagined speech synchronize with the speech envelope of the overt counterpart.

- Based on this phenomenon, EEG oscillations during imagined speech can classify speech stimuli with different amplitude envelopes.
- The DTW method is successful to correct the durational variability of the imagined speech.

6. EEG-based neural decoding of imagined speech

6.1 Purposes of Experiment 3

In the previous Experiment, I demonstrated that EEG during imagined speech synchronizes with a rhythm of imagined speech (which was replaced by overt counterpart). As well as neural phase synchronization during speech perception, the imagined speech with different amplitude rhythm can be discriminated using EEG phase patterns because the phase synchronization induces stimuli-specific phase patterns. However, the effectiveness of neural phase synchronization for neural decoding of speech is not fully clear because it is unknown if the synchronization is observed when using meaningful sentences. Thus, Experiment 3 aims to replicate Experiment 2 using meaningful sentences.

Another purpose is to investigate whether the synchronization during the imagined speech is observed in different frequency bands relevant to linguistic processing (i.e., delta, theta, alpha, beta, and gamma) as well as Experiment 1. In Experiment 2, the analysis of phase synchronization was focused on the lower frequency band (delta and theta). However, considering the existence of neural synchronization in other frequency bands during speech perception, it is expected that the imagined EEG phases synchronize with the speech imagery in other frequency bands.

In sum, research questions of Experiment 3 are:

1. Whether does EEG phase synchronization during the imagined speech is observed in multiple frequency bands (i.e., delta, theta, alpha, beta, and gamma) using meaningful sentences?

2. Whether do EEG phase patterns during the imagined speech discriminate three types of meaningful sentences?

To this end, I used the same sentences to Experiment 1 because it had already demonstrated that these sentences could be classified using phase synchronization with perceived speech. The analysis method of PLV was also quite similar to Experiment 1. I used SVM trained by features which were extracted by the CSP method (Experiment 1) and a template matching method used in Experiment 2 for classification. To mitigate a variance of the duration of the imagined speech,

the EEG data were realigned to the model speech using DTW, which was the same method to Experiment 2. Furthermore, as well as Experiment 1, the classifiers were trained by subject-dependent, subject-inclusive, and subject-independent manner.

6.2 Methods of Experiment 3

6.2.1 Participants

Six right-handed L1 Japanese speakers participated in data recordings (1 female, 5 males). The average age was 24.7±2.7 across participants. All participants agreed to participate and gave informed consent in writing. They all reported no history of neurological illness and no hearing abnormalities. This experiment was approved by the ethical review board of the Nara Institute of Science and Technology.

6.2.2 Experimental materials

The same speech materials to Experiment 1 was used in the current experiment (see Chapter 4.2.2).

6.2.3 EEG recordings

EEGs were recorded with the same apparatus and settings to Experiment 2. The recording procedure was also similar to Experiment 2. Participants sat on a comfortable chair in a dimly lit sound-attenuated chamber. A display monitor, keyboard, and microphone were placed on a desk in front of the chair. Participants were instructed to familiarize themselves with the speech stimuli and memorize them before the EEG recording. One trial consisted of three tasks: listening, speaking, and imagining speech. During the trial, they were instructed to stay as still as possible. In the imagined speech task, participants were forbidden to move any articulators such as their mouth or lips and were instructed to imagine the articulatory movements of the speech stimulus without actually making those movements.

The experiments consisted of practice and three main blocks. Both blocks consisted of 16 trials. Each stimulus was presented to each participant 16 times

in a randomized order in the main blocks. The procedure of a trial did not vary between the practice and the main blocks. A procedure of one trial is as follows.

1. Each trial started from the appearance of *Ready?* on the display. Participants initiated a trial by pushing the space key on the keyboard.
2. A letter *LISTEN* was displayed for indicating the task type for 2,000 ms.
3. A fixation mark (+) appeared for 2,500 ms, the color of the fixation mark was changed from white to blue. At the same time, a speech stimulus was played via headphones.
4. After 5,500 ms, the color was changed to white again to indicate the end of the task.
5. In the speaking task, the task indication (*SPEAK*; 2,000 ms) was followed by the appearance of a fixation mark for 2,500 ms.
6. After the color of the fixation mark changed from white to blue, participants uttered the speech stimulus that was presented in the listening task at the same speech rate.
7. In the imagined speech task, after the task indication (*IMAGINE*; 2,000 ms), the procedure was the same as the speaking task except that participants imagined articulatory movements of the speech stimulus.

The procedure of one trial is summarized in Fig. 29. The experiment continued for about 40 minutes.

6.2.4 Preprocessing of EEG

Data from one participant was removed due to the recording error. For EEG data analysis, the FieldTrip toolbox for MATLAB (The MathWorks, Inc., U.S.A) was used [60]. First, one-pass zero-phase FIR high-pass filter at 0.5 Hz (filter order: 4,096th, a window type: hamming) was applied to continuous EEG data. After EEG data were re-referenced to average values of TP9 and TP10 electrodes, EEG data belonging to the imagined speech task were epoched from −1,000 to 3,500 ms relative to the onset of the task.

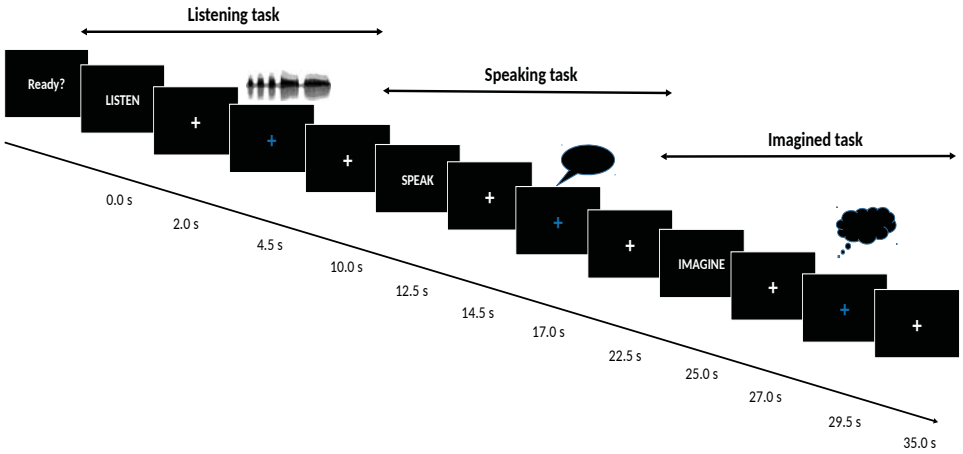

Figure 29. Procedure of an experimental trial in Experiment 3. In the listening task, the stimulus was played at the same time to the color change of the fixation mark. In the speaking task, participants uttered the speech stimuli into a microphone. In the imagined speech task, they imagined the articulatory movement of speech stimuli without making movements.

Next, I rejected trials contaminated with large amplitude artifacts and muscle artifacts. For large amplitude artifacts, trials exceeding ± 200 μv were removed from further analysis. This rejection procedure was not applied to data from FP1 and FP2 electrodes because data from these electrodes were very often contaminated by eye movement-related artifact which was removed by ICA later. Trials including muscle artifacts were detected by a z-score-based method and by visual inspection (see Chapter 4.2.4). In average, 5.0 ± 3.2 of trials across all participants were removed.

EEG data were decomposed of ICs by ICA. The ICs reflecting blinks, eye movements, electrocardiograms, electromyograms, and noise derived from electrodes were selected by inspecting the waveforms and topographies of the ICs visually. The selected ICs were removed from the EEG data.

Finally, I excluded trials in which participants uttered the speech incorrectly, such as through a slip of the tongue in the preceding speaking task. The incorrect utterances were annotated manually. The number of average percents judged as the incorrect speech was 0.6 ± 0.89 across participants.

6.3 Quantification of neural phase synchronization

To quantify a degree of phase synchronization, I calculated PLV in each frequency band using the same method to Experiment 1. A different point from Experiment 1 is to apply a delay optimization to EEG data as well as Experiment 2 because lags between the trial onset and the initiation of the imagined speech are expected. The same delay optimization procedure to Experiment 2 was adopted (see Chapter 5.2.7).

6.4 Imagined speech classification

I used both the template matching classification which was the same method to Experiment 2 and SVM trained by features extracted using CSP. Before model training, single-trial EEG data were realigned to the corresponding model speech (see Chapter 5.2.9) to mitigate variations in duration of the imagined speech. Each model was trained by phase patterns in a single frequency band. In the template matching classification, classification was performed per electrode in a fronto-central region (i.e., Fz, FC2, FCz, FC1, C3, Cz and C4 electrodes) where phase synchronization tended to be observed in Experiment 2. The final prediction result was decided based on a voting system across these electrodes. In SVM trained by CSP-based features, the feature was extracted using a CSP method from a single frequency band. The trained models were evaluated by LOSO cv and LOO cv to confirm whether the models have generalizability to unknown users (see Chapter 4.4.2).

6.5 Results of Experiment 3

6.5.1 EEG phase synchronization with meaningful imagined speech

The average delay in the delay optimization procedure was 293±74 ms, 323±33 ms, 358±88 ms, 419±48 ms and 386±127 ms for delta, theta, alpha, beta, and gamma, respectively. I plotted the PLV topographies in each frequency band in Fig. 30. Visual inspection suggests that the theta frequency band shows relatively strong PLV responses in a fronto-central region. Given that this pattern is consistent with the previous results, it can be considered as phase synchroniza-

tion with speech rhythm. There is no clear tendency of phase synchronization in other frequency bands.

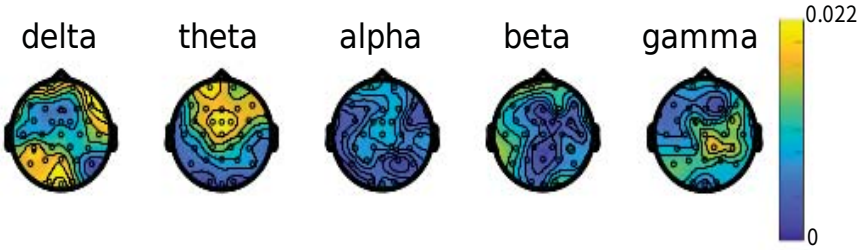

Figure 30. PLV topographies in each frequency band in Experiment 3.

The values of PLV across front-central regions (FCz, Cz, FC1, FC2, F3, and F4) were plotted in Fig. 31 and the average values across the electrodes were summarized in Table 12. In the average values, the PLV in the theta frequency band was larger than the other frequency bands. As a result of multiple comparisons using Tukey-Kramer, the differences between PLVs in theta and other frequency bands were statistically significant (p<0.01, respectively).

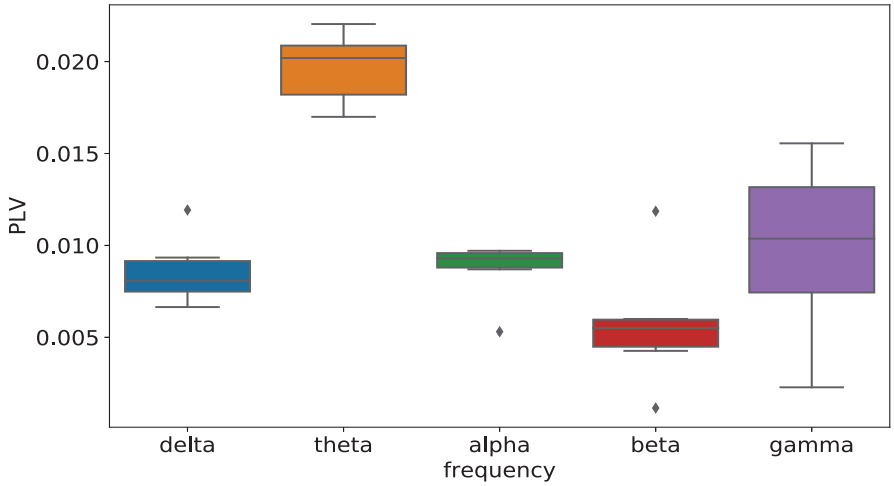

Figure 31. Boxplots of PLVs in the fronto-central regions per frequency band.

To decide whether this synchronization is larger than the null-hypothesis distribution, I performed a permutation test (see Chapter 4.3). To construct the

Table 12. Average PLVs across the fronto-central electrodes per frequency band.

	Delta	Theta	Alpha	Beta	Gamma
Mean	0.86e-2	1.97e-2	0.87e-2	0.57e-2	0.99e-2
SD	0.19e-2	0.20e-2	0.17e-2	0.35e-2	0.49e-2

null-hypothesis distribution, I calculated PLV values using data that a temporal relationship between EEG and speech was broken by splitting EEG data at random time point and swapping the data before and after the split point. I decided to test the significance only in the theta frequency band because there was no tendency of phase synchronization in other frequency bands. The maximum PLV value among electrodes was used to construct the distribution to correct multiple comparisons in statistical tests [51]. The number of iterations was 1,000. The number of values exceeding the observed PLVs for each electrode in the null-hypothesis distribution was divided by the number of iterations for calculating p values. The one-sided test was used for determining the significance because the purpose of the statistical test was to determine whether the observed PLV exceeds the random distribution. As a result of the permutation test, FCz and FP1 electrodes reached the marginally significant level (p=0.086, p=0.091, respectively).

6.5.2 Classification performances

I calculated the classification performances in the theta frequency band because the other frequency band did not show phase synchronization with speech. Classification accuracies were summarized in Fig. 32 and Table 13. The template matching and CSP-based SVM in the subject-dependent model outperformed the chance level (33.3%). The one-sample t-test revealed that the accuracies of the template matching were marginally significant compared to the 33.3% chance rate ($t(4)$=2.75, p=0.051). On the other hand, the CSP-based SVM did not reach the significance ($t(4)$=0.99, p=0.38).

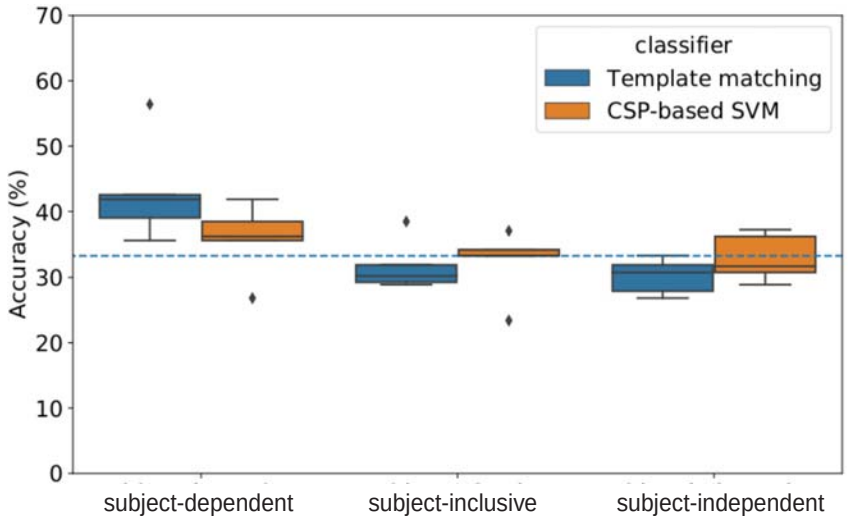

Figure 32. Boxplots of classification accuracies in each model.

Table 13. Mean accuracies per classification model (SD).

	subject-dependent	subject-inclusive	subject-independent
Template matching	43.1 (7.1)	31.8 (3.5)	30.2 (2.4)
CSP-based SVM	35.8 (5.0)	32.3 (4.6)	32.9 (3.2)

6.6 Discussion of Experiment 3

Experiment 3 aims to replicate the results of Experiment 2 using meaningful sentences. Research questions were (1) Whether do EEG phase synchronization during the imagined speech is observed in multiple frequency bands (i.e., delta, theta, alpha, beta, and gamma) using meaningful sentences? and (2) Whether do EEG phase patterns during the imagined speech discriminate the meaningful sentences?

The answer to the research question (1) is positive in the theta frequency band. Although the result of the statistical test was marginally significance against the null-hypothesis distribution, considering that the topographical pattern is consistent across previous results, it can be considered as an EEG phase synchronization phenomenon. This result further reinforces the hypothesis that neural

phases during the imagined speech synchronize with a rhythm of the imagined speech.

One difference between the current result and Experiment 1 is that tendencies of phase synchronization in other frequency bands (alpha, beta, and gamma) were not observed. I expected phase synchronization in the low-gamma frequency band because Giraud et al. (2007) [30] showed that EEGs in a premotor region related to tongue movement fluctuates at a gamma frequency band endogenously. Considering the facts that tongue movements are related to phoneme production (e.g., velar stop: [k] or [g], alveolar fricative: [s] or [z]) and that the phase synchronization in low-gamma band was observed in [50] during acoustic processing, it is expected that this gamma fluctuation in the premotor cortex is related to generation of phoneme level rhythm during the imagined speech. However, one difficult point in PLV calculation is that an effect of phase lags gets more critical in the higher frequency bands than the low-frequency band. Thus, the phase patterns in the higher frequency bands are more prone to be affected by duration variability of speech imagery across trials. Besides, the S/N ratio of EEG in the higher frequency bands is worse than the lower one because of smaller amplitudes EEG amplitudes in the higher bands.

The delta frequency band also did not show the synchronization, which coincides with the results of Experiment 1. As mentioned in Chapter 3.1, considering that delta oscillations are related to intonation [13] and syntactic boundary [22, 55], I expected that internal syntactic or prosodic chunking during the imagined speech might be related to delta oscillations. However, as discussed in Chapter 4.6, it might be reasonable to investigate the delta effect using sentences including explicit prosodic or syntactic boundaries.

The answer to the research question (2) is also positive, but it depends on the settings. The template matching classification using the DTW, which was used in Experiment 2, discriminated the imagined sentences successfully in the subject-dependent classification, although the statistical test showed a marginal significance compared to the chance level. The classification performances (43.1%) seem to be better than nonsense imagined speech classification in Experiment 2 (38.5%). Some research reported that intelligibility enhanced phase synchronization during speech perception [64]. Thus, the meaningfulness of the imagined

speech might enhance the synchronization and leads to more robust phase patterns to obtain better classification accuracies. Alternatively, a longer duration of speech stimuli than Experiment 2 affects the classification accuracies (around 3,000 ms vs 1,800 ms). To investigate the effect of meaningfulness of the imagined speech, further control of the experimental condition is necessary. However, if meaningfulness of the imagined speech affects the synchronization during the imagined speech, it might be lead to further understanding of this phenomenon.

Contrary to the template matching, the performance of the any CSP-based SVM classification did not outperform the chance rate significantly. As discussed in Experiment 1, this might be because the neural source of the EEG synchronization is considered to be common across sentence types. Although the neural source of this synchronization is not revealed, at least, null-effect of the CSP methods suggests the common neural sources across sentence types.

As for the effects of the subject-dependency on the classification accuracies, in contrast to the subject-dependent model, any models did not achieve the chance level in the subject-inclusive and subject-independent models. This result is against Experiment 1 demonstrating that the increase in number of training data in the subjet-independent and subject-adaptive classification improves classification accuracy compared to subject-dependent classification. This opposite result to Experiment 1 is might be due to the following two reasons. First, temporal duration of imagined speech largely varied across participants because speech rate is influenced by individual differences. Second, increase in number of training data by subject-inclusive and independent was not large compared to Experiment 1 because the number of participants of Experiment 3 was fewer than Experiment 1 (Experiment 1: 17 participants, Experiment 3: 6 participants, one participant was removed). However, the subject-independency is a merit of the BCI application. Thus, in the future, it is required to correct such individual differences of the imagined speech rhythm across participants.

6.7 Summary of Chapter 6

In Experiment 3, I obtained the following results:

- EEG theta phase synchronization during the imagined speech is observed

when using meaningful sentences.

- Three meaningful imagined sentences can be discriminated successfully in subject-dependent classification using the DTW-based template matching trained by phase patterns in the theta frequency band. To realize subject-independent classification, a method to correct individual differences of the imagined speech across participants is required.

7. Summary and future directions

7.1 Summary and achievement

The purpose of the thesis is to reveal that the EEG phase information induced by neural synchronization successfully classify speech information. This is motivated from the fact that the underlying mechanisms remain unclear in neural decoding of speech. In Experiment 1, I investigated classification performances using EEG phases during speech perception. As the results of Experiment 1, I demonstrated the effectiveness of EEG phase patterns for neural decoding of speech:

- EEG phase information can classify three types of speech with around 50% accuracy. This phase-based classification model has generalizability to other different users in the perceived speech classification.
- Template matching is a better method for phase synchronization-based classification in terms of accuracy and analysis of feature importance.
- Use of phase patterns in the multiple frequency bands relevant to phase synchronization improves classification accuracy in the subject-independent models.

Experiment 2, I investigated whether similar phase synchronization is induced during the speech imagery: synchronization with EEGs during the imagined speech and speech rhythms of the imagined speech, which was replaced with the overt counterpart. The results showed that:

- EEG oscillations during imagined speech synchronize with the speech envelope of the overt counterpart.
- Based on this phenomenon, EEG oscillations during imagined speech can classify speech stimuli with different amplitude envelopes.

Experiment 3 investigated whether phase synchronization during the imagined speech can apply to meaningful sentences using the same sentence to Experiment 1. As results,

- EEG theta phase synchronization during the imagined speech is observed when using meaningful sentences.
- Three meaningful imagined sentences can be discriminated successfully in subject-dependent classification using the DTW-based template matching trained by the theta frequency band.

Over these experiments, I demonstrated that neural phase synchronization is effective for neural decoding of speech in both speech perception and imagined speech task. Thus, the main purpose of the thesis was successfully achieved.

7.2 Future directions: limitations and possible directions

However, much work for realizing neural decoding of speech as BCI application are left. In this section, I discuss the future direction of the thesis: limitations and possible solutions.

7.2.1 Classification accuracy

Limitation. The best classification accuracies were 50.5% (SVM trained by all in subject-independent classification) in speech perception and 43.1% in imagined speech (DTW-based template matching trained by theta phase patterns in the subject-dependent model) in the three-class classification. This performance seems not to be enough for practical application.

Possible solutions. A first possible approach is to test the performances of the state-of-the-art algorithms using EEG phase patterns, which manages both the performance-oriented classification and neurophysiological validness. Note that the current classification has focused on the interpretability of the model and there are more powerful algorithms have been developed in the machine learning field. As introduced in Chapter 2.5.2, the state-of-the-art research of neural decoding of imagined speech achieved the 97% in the best case in a binary words classification task using a Riemannian manifold [59].

Another possible solution is to improve the S/N ratio of EEG signals to obtain better accuracy. The one the methods is an averaging method. In the domain

of ERP-based BCI such as the P300-speller, averaging several numbers of trials achieves around 99% accuracy (see Chapter 2). Although it is unclear whether such high accuracy is obtained in the neural decoding of speech, it is no doubt that averaging multiple trials are the simplest method to improve the S/N ratio. On the other hand, the disadvantages of the averaging process are that it requires more time to classify speech compared to single-trial classification because it requires more trials. Thus, the best balance between the S/N ratio and the necessary time for it need to be investigated if there is need to use an averaging method.

7.2.2 Number of classes

Limitation. Another major limitation of the current method is that a few numbers of classes. As summarized in Table 1, the almost previous research focused on binary classification for neural decoding of imagined speech. Thus, at present, given that the classification accuracies in the limited number of classes, it seems to be difficult to increase the number largely in the neural decoding.

Possible direction. One possible solution to use the decoding as a BCI system is to combine the existing EEG-based spelling system and the neural decoding of speech. For example, while the conversation in daily life is performed using the spelling system, greetings (e.g., good morning, have a nice day and how are you?) was outputted via neural decoding of speech. The greetings are fixed phrases and quick responses should be preferable. Compared to the one-by-one selection of the characters in the spelling system, neural decoding of speech have the potential to output speech more quickly (This is just in terms of that no trial averaging is necessary in case of single-trial classification such as the current experiments).

One might argue that the above-mentioned things can be solved by putting icons corresponding to the greetings on the speller. The statement is completely true. However, icon selection (or spelling) is difficult to convey para-linguistic information such as emotional intonation. Thus, I would like to propose paralinguistic estimation in the neural decoding of speech in parallel to the content estimation. The neural decoding of speech might have the potential to discriminate paralinguistic information such an emotional intonation pattern because it

is also related to linguistic rhythm. If the neural decoding of speech output the fixed phrases with paralinguistic information and other conversation is performed by the existing spelling system, neural decoding of the limited number of speech might be useful.

7.2.3 Effectiveness in other speech stimuli

Limitations. The performances of other different three sentences remain unclear because I investigated the classification of only three Japanese sentences. The classification performance is expected to be worse if the speech rhythm which is quite similar between the target speech is chosen.

Besides, it is not clear whether speech stimuli with a shorter duration than the current speech stimuli can be classified successfully. Especially, the single syllable or phoneme is too difficult to be classified using phase synchronization because they do not have a speech rhythm.

Possible directions. The solution is to investigate the performances of other different speech set from the current experiments. Given that the previous neural phase-based classification showed a similar classification performance (around 50%) when using different three sentences [49, 50], the expectation values of single-trial classification accuracy might be around the accuracy in neural decoding of perceived speech. To solve this problem, further research is necessary in the future.

As stated above, if speech rhythms are similar to each other, classification is difficult using phase synchronization. However, I do not particularly want to use only phase synchronization for neural decoding of speech. Thus, another future direction is to find a relationship between speech processing and neural oscillations dynamics leading to better and more robust classification performances.

Besides, it is necessary to investigate the performances using stimuli with a shorter duration than the current stimuli. Note that there is no need to persist in sentence classification and the sentences were used as stimuli only for inducing phase synchronization. Thus, it is also good to investigate classification performances of a more shorter linguistic unit such as words. However, as mentioned above, this classification method is not an optimal tool to classify single syllable

and phoneme. To discriminate them, other features need to be investigated [79].

7.2.4 Performances by ALS patients

Limitation. While participants of the current experiments were healthy people, it is assumed that patients with severe motor disabilities utilize this system. Regardless of that, classification performances remain unclear in the decoding by patients with severe motor disabilities such as ALS. Because the activity of the motor-related region of such patients might be impaired, the classification performances might be worse the current results.

Possible direction. It is necessary to investigate whether this classification method is available when the patients use the system. To this end, one thing needs to be solved: the timing of initiation of the imagined speech. In Experiments, the timing of initiation of imagined speech was controlled by the experimenter. However, when assuming that patients use the system, it is difficult to indicate the initiation of the speech imagery. One possible solution is to utilize SSVEP: A part of the display is flickering and the user starts to direct attention to the flickering light at the timing when they want to convey something, which evokes an SSVEP response. The system sends a signal to initiate the imagined speech when it detects the SSVEP and the user start the imagined speech. This is just one example. In the future, a system that ALS patients can use alone need to be developed and need to investigate the performances by the patients.

Acknowledgements

I owe my deepest gratitude to my academic advisor Professor. Satoshi Nakamura (NAIST, Graduate School of Science and Technology). He always supported and guided me a lot during the doctoral course. Without his insightful advice, guidance, and comments, it was impossible to complete this project.

I am also heartily thankful to Professor. Yuji Matsumoto (NAIST, Graduate School of Science and Technology) whose comments in the interim report of the thesis and advice for thesis writing had been a lot of help.

I am also heartily thankful to Assistant Professor. Hiroki Tanaka (NAIST, Graduate School of Science and Technology), Associate Professor. Sakriani Sakti (NAIST, Graduate School of Science and Technology) and Dr. Lars Meyer (Max Planck Institute for Human Cognitive and Brain Sciences). Assistant Professor. Hiroki Tanaka, who is in charge of the cognitive communication group of our laboratory, guided me to plan experiments, write academic papers, and progress my doctoral project. Associate professor. Sakriani Sakti gave me a lot of advice about signal processing and statistical machine learning algorithms from a view of a specialist of spoken language processing. Dr. Lars Meyer supported and instructed me during my research internship at Max Planck Institute of Human Cognitive and Brain Sciences, Leipzig, Germany. I learned a lot of things about handling and analyzing neurophysiological data from him.

I would like to thank Ms. Manami Matsuda and Ms. Miho Hayashi who are secretaries of the augmented human communication laboratory. They not only supported the project, but their kind and encouraging words helped me.

Special thanks go to the participants of the EEG experiment and the members of the augmented human communication laboratory.

Finally, I would like to express my sense of gratitude to my family.

References

[1] E. Ahissar, S. Nagarajan, M. Ahissar, A. Protopapas, H. Mahncke, and M. M. Merzenich. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences*, 98(23):13367–13372, 2001.

[2] B. Z. Allison, C. Brunner, C. Altstätter, I. C. Wagner, S. Grissmann, and C. Neuper. A hybrid ERD/SSVEP BCI for continuous simultaneous two dimensional cursor control. *Journal of Neuroscience Methods*, 209(2):299–307, 2012.

[3] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan. Filter bank common spatial pattern (fbcsp) in brain-computer interface. *Proceedings of International Joint Conference on Neural Networks*, pages 2390–2397, 2008.

[4] A. Arieli, A. Sterkin, and A. Grinvald A. D. Aertsen. Dynamics of ongoing activity: explanation of the large variability in evoked cortical responses. *Science*, 273(5283):1868–1871, 1996.

[5] L. H. Arnal and A. L. Giraud. Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, 16(7):390–398, 2012.

[6] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[7] Z. Bayraktaroglu, K. von Carlowitz-Ghori, F. Losch, G. Nolte, G. Curio, and V. V. Nikulin. Optimal imaging of cortico-muscular coherence through a novel regression technique based on multi-channel EEG and un-rectified EMG. *NeuroImage*, 57(3):1059–1067, 2011.

[8] H. Berger. über das elektrenkephalogramm des menschen. *Arch. Psychiatr Nervenkr*, 87:527–570, 1929.

[9] L. Bi, J. Lian, K. Jie, R. Lai, and Y. Liu. A speed and direction-based cursor control system with P300 and SSVEP. *Biomedical Signal Processing and Control*, 14:126–133, 2014.

[10] N. Birbaumer, N. Ghanayim, T. Hinterberger, I. Iversen, B. Kotchoubey, A. Kübler, J. Perelmouter, E. Taub, and H. Flor. A spelling device for the paralysed. *Nature*, 398:297–298, 1999.

[11] B. Blankertz, G. Dornhege, M. Krauledat, M. Schröder, J. Williamson, R. Murray-Smith, and K. R. Müller. The Berlin Brain-Computer Interface presents the novel mental typewriter Hex-o-Spell. In *Proceedings of the 3rd International Brain-Computer Interface Workshop and Training Course*, pages 1–2, Graz, Austria, 2006.

[12] P. Boersma. Praat, a system for doing phonetics by computer. *Glot International*, 5:341–345, 2002.

[13] M. Bourguignon, X. De Tiege, M. O. de Beeck, N. Ligot, P. Paquier, P. Van Bogaert, S. Goldman, R. Hari, and V. Jousmäki. The pace of prosodic phrasing couples the listener's cortex to the reader's voice. *Human Brain Mapping*, 34(2):314–326, 2013.

[14] K. Brigham and B. V. K. Kumar. Imagined speech classification with EEG signals for silent communication: a preliminary investigation into synthetic telepathy. In *Proceedings of International Conference on Bioinformatics and Biomedical Engineering*, pages 1–4, 2010.

[15] P. Brunner, A. L. Ritaccio, F. J. Emrich, H. Bischof, and G. Schalk. Rapid communication with a "P300" matrix speller using electrocorticographic signals (ECoG). *Frontiers in Neuroscience*, 5(5), 2011.

[16] A. M. Chan, E. Halgren, K. Marinkovic, and S. S. Cash. Decoding word and category-specific spatiotemporal representations from MEG and EEG. *NeuroImage*, 54(4):3028–3039, 2011.

[17] X. Chi, J. B. Hagedorn, D. Schoonover, and M. D'Zmura. EEG-based discrimination of imagined speech phonemes. *International Journal of Bioelectromagnetism*, 13(4):201–206, 2011.

[18] Z. Y. Chin, K. K. Ang, C. Wang, C. Guan, and H. Zhang. Multi-class filter bank common spatial pattern for four-class motor imagery BCI. In *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 571–574, 2009.

[19] J. M. Correia, B. Jansma, L. Hausfeld, S. Kikkert, and M. Bonte. EEG decoding of spoken words in bilingual listeners: from words to language invariant semantic-conceptual representations. *Frontiers in Psychology*, 6(71):1–10, 2015.

[20] C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike. Single-trial classification of vowel speech imagery using common spatial patterns. *Neural Networks*, 22(9):1334–1339, 2009.

[21] S. Deng, R. Srinivasan, T. Lappas, and M. D'Zmura. EEG classification of imagined syllable rhythm using Hilbert spectrum methods. *Journal of Neural Engineering*, 7:046006, 2010.

[22] N. Ding, L. Melloni, H. Zhang, X. Tian, and D. Poeppel. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, 19(1):158–164, 2016.

[23] M. D'Zmura, S. Deng, T. Lappas S. Thorpe, and R. Srinivasan. Toward EEG sensing of imagined speech. In *Jacko J.A. (eds) Human-Computer Interaction. New Trends. HCI 2009. Lecture Notes in Computer Science*, volume 5610, pages 40–48, Springer, Berlin, Heidelberg, 2009.

[24] L. A. Farwell and E. Donchin. Talking off the top of your head: toward a mental prosthesis utilizing event-related brain potentials. *Electroencephalography and Clinical Neurophysiology*, 70(6):510–523, 1988.

[25] R. Filik and E. Barber. Inner speech during silent reading reflects the reader's regional accent. *PloS One*, 6(10):e25782, 2011.

[26] J. Fox and S. Weisberg. *An R companion to applied regression*. Sage, Thousand Oaks CA, 2nd edition, 2011.

[27] E. Gerardin, A. Sirigu, S. Lehéricy, J. B. Poline, B. Gaymard, C. Marsault, and Y. Agid D. L. Bihan. Partially overlapping neural networks for real and imagined hand movements. *Cerebral Cortex*, 10:1093–1104, 2000.

[28] O. Ghitza. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, 2(130):1–13, 2011.

[29] O. Ghitza and S. Greenberg. On the possible role of brain rhythms in speech perception: intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, 66:113–126, 2009.

[30] A. L. Giraud, A. Kleinschmidt, D. Poeppel, T. E. Lund, R. S. J. Frackowiak, and H. Laufs. Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron*, 56:1127–1134, 2007.

[31] A. L. Giraud, C. Lorenzi, J. Ashburner, J. Wable, I. Johnsrude, R. Frackowiak, and A. Kleinschmidt. Representation of the temporal envelope of sounds in the human brain. *Journal of Neurophysiology*, 84(3):1588–1598, 2000.

[32] F. H. Guenther. Cortical interactions underlying the production of speech sounds. *Journal of Communication Disorders*, 39(5):350–365, 2006.

[33] F. H. Guenther, S. S. Ghosh, and J. A. Tourville. Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, 96(3):280–301, 2006.

[34] C. Guger, S. Daban, E. Sellers, C. Holzner, G. Krausz, R. Carabalon, F. Gramatic, and G. Edlinger. How many people are able to control a P300-based brain-computer interface (BCI)? *Neuroscience Letters*, 462:94–98, 2009.

[35] V. Guy, M. H. Soriani, M. Bruno, T. Papadopoulo, C. Desnuelle, and M. Clerc. Brain computer interface with the P300 speller: Usability for disabled people with amyotrophic lateral sclerosis. *Annals of Physical and Rehabilitation Medicine*, 61:5–11, 2018.

[36] E. Hortal, D. Planelles, A. Costa, E. Lá nez, A. Úbeda, J. M. Azorín, and E. Fernández. SVM-based Brain-Machine Interface for controlling a robot arm through four mental tasks. *Neurocomputing*, 151:116–121, 2015.

[37] T. Hothorn, F. Bretz, and P. Westfall. Simultaneous inference in general parametric models. *Biometrical Journal*, 50(3):346–363, 2008.

[38] M. F. Howard and D. Poeppel. Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *Journal of Neurophysiology*, 104(5):2500–2511, 2010.

[39] A. Hyafil, L. Fontolan, C. Kabdebon, B. Gutkin, and A. L. Giraud. Speech encoding by coupled cortical theta and gamma oscillations. *Elife*, 4:1–23, 2015.

[40] S. Inoue, Y. Akiyama, Y. Izumi, and S. Nishijima. The development of BCI using alpha waves for controlling the robot arm. *IEICE transactions on communications*, 91(7):2125–2132, 2008.

[41] J. R. Kerlin, A. J. Shahin, and L.M. Miller. Attentional gain control of ongoing cortical speech representations in a "cocktail party". *The Journal of Neuroscience*, 30(2):620–628, 2010.

[42] J. Kubanek, P. Brunner, A. Gunduz, D. Poeppel, and G. Schalk. The tracking of speech envelope in the human cortex. *PloS One*, 8(1):e53398, 2013.

[43] A. Kübler, B. Kotchoubey, T. Hinterberger, N. Ghanayim, J. Perelmouter, M. Schauer, C. Fritsch, E. Taub, and N. Birbaumer. The thought translation device: a neurophysiological approach to communication in total motor paralysis. *Experimental Brain Research*, 124:223–232, 1999.

[44] A. Kübler, N. Neumann, J. Kaiser, B. Kotchoubey, T. Hinterberger, and N. P. Birbaumer. Brain-computer communication: self-regulation of slow cortical potentials for verbal communication. *Archives of Physical Medicine and Rehabilitation*, 82:1533–1539, 2001.

[45] J. P. Lachaux, E. Rodriguez, J. Martinerie, and F. J. Varela. Measuring phase synchrony in brain signals. *Human Brain Mapping*, 8(4):194–208, 1999.

[46] P. Lakatos, A. S. Shah, K. H. Knuth, I. Ulbert, G. Karmos, and C. E. Schroeder. An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of Neurophysiology*, 94(3):1904–1911, 2005.

[47] J. Long, Y. Li, H. Wang, T. Yu J. Pan, and F. Li. A hybrid brain computer interface to control the direction and speed of a simulated or real wheelchair. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 20(5):720–729, 2012.

[48] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger. A review of classification algorithms for EEG-based brain-computer interfaces: a 10 year update. *Journal of Neural Engineering*, 15(3):031005, 2018.

[49] H. Luo and D. Poeppel. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, 54(6):1001–1010, 2007.

[50] H. Luo and D. Poeppel. Cortical oscillations in auditory perception and speech: evidence for two temporal windows in human auditory cortex. *Frontiers in Psychology*, 3(170):1–10, 2012.

[51] E. Maris and R. Oostenveld. Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, 164(1):177–190, 2007.

[52] D. J. McFarland, W. A. Sarnacki, and J. R Wolpaw. Electroencephalographic (EEG) control of three-dimensional movement. *Journal of Neural Engineering*, 7, 2010.

[53] D. J. McFarland and J. R. Wolpaw. Brain-computer interfaces for communication and control. *Communications of the ACM*, 54(5):60, 2011.

[54] L. Meyer. The neural oscillations of speech processing and language comprehension: state of the art and emerging mechanisms. *Europen Jorunal of Neuroscience*, 48:2609–2621, 2017.

[55] L. Meyer, M. J. Henry, P. Gaston, N. Schmuck, and A. D. Friederici. Linguistic bias modulates interpretation of speech via neural delta-band oscillations. *Cerebral Cortex*, pages 1–10, 2016.

[56] J. D. R. Millán, R. Rupp, G. R. Müller-Putz, R. Murray-Smith, C. Giugliemma, M. Tangermann, C. Vidaurre, F. Cincotti, A. Kübler, R. Leeb, C. Neuper, K.-R. Müller, and D. Mattia. Combining brain-computer interfaces and assistive technologies: state-of-the-art and challenges. *Frontiers in Neuroscience*, 4(161), 2010.

[57] K. R. Müller and B. Blankertz. Toward noninvasive brain-computer interfaces. *IEEE Signal Processing Magazine*, 23:128–126, 2006.

[58] C. Neuper and G. Pfurtscheller. Event-related dynamics of cortical rhythms: frequency-specific features and functional correlates. *International Journal of Psychophysiology*, 43:41–58, 2001.

[59] C. H. Nguyen, G. K. Karavas, and P. Artemiadis. Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features. *Journal of Neural Engineering*, 15, 2018.

[60] R. Oostenveld, P. Fries, E. Maris, and J. M. Schoffelen. FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, 1:1–9, 2011.

[61] L. Parra, C. Alvino, A. Tang, B. Pearlmutter, N. Yeung A. Osman, and P. Sajda. Linear spatial integration for single-trial detection in encephalography. *NeuroImage*, 17(1):223–230, 2002.

[62] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[63] J. E. Peelle and M. H. Davis. Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Pyschology*, 3:1–17, 2012.

[64] J. E. Peelle, J. Gross, and M. H. Davis. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, 23(6):1378–1387, 2012.

[65] G. Pfurtscheller, C. Brunner, A. Schlögl, and F. H. Lopes da Silva. Mu rhythm (de)synchronization and EEG single-trial classification of different motor imagery tasks. *NeuroImage*, 31:153–159, 2006.

[66] D. Poeppel. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication*, 41:245–255, 2003.

[67] C. J. Price. A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, 62(2):816–847, 2012.

[68] R Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[69] H. Ramoser, J. Muller-Gerking, and G. Pfurtscheller. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE transactions on rehabilitation engineering*, 8(4):441–446, 2000.

[70] V. Romei, J. Gross, and G. Thut. On the role of prestimulus alpha rhythms over occipito-parietal areas in visual input regulation: correlation or causation? *Journal of Neuroscience*, 30(25):8692–8697, 2010.

[71] M. Salama, L. ElSherif, H. Lashin, and T. Gamal. Recognition of unspoken words using electrode electroencephalograhic signals. In *Sixth International Conference on Advanced Cognitive Technologies and Applications, Cognitive 2014 (IARIA)*, pages 51–55, Venice, Italy, 2014.

[72] W. Speier, C. Arnold, N. Chandravadia, D. Roberts, S. Pendekanti, and N. Pouratian. Improving P300 spelling rate using language models and predictive spelling. *Brain-Computer Interfaces*, 5:13–22, 2018.

[73] P. Suppes, B. Han, and Z. L. Lu. Brain-wave recognition of sentences. *Proceedings of the National Academy of Sciences*, 95(26):15861–15866, 1998.

[74] P. Suppes, Z. L. Lu, and B. Han. Brain wave recognition of words. *Proceedings of the National Academy of Sciences*, 94(26):14965–14969, 1997.

[75] S. Sutton, M. Braren, J. Zubin, and E. R. John. Evoked-potential correlates of stimulus uncertainty. *Science*, 150(3700):1187–1188, 1965.

[76] X. Tian and D. Poeppel. Mental imagery of speech and movement implicates the dynamics of internal forward models. *Frontiers in Psychology*, 1(166), 2010.

[77] R. Turn. The use of speech for man-computer communication. *Tech. Rep. RAND Report-1386-ARPA*, RAND-Corp, 1974.

[78] R. Ulrich, G. Rinkenauer, and J. Miller. Effects of stimulus duration and intensity on simple reaction time and response force. *Journal of Experimental Psychology: Human Perception and Performance*, 24(3):915–928, 1998.

[79] R. Wang, M. Perreau-Guimaraes, C. Carvalhaes, and P. Suppes. Using phase to recognize English phonemes and their distinctive features in the brain. *Proceedings of the National Academy of Sciences*, 109(50):20685–20690, 2012.

[80] H. Watanabe, H. Tanaka, S. Sakti, and S. Nakamura. Subject-independent classification of Japanese spoken sentences by multiple frequency bands phase pattern of EEG response during speech perception. In *Proceedings of Interspeech*, pages 2431–2435, Stockholm, Sweden, 2017.

[81] U. Will and E. Berg. Brain wave synchronization and entrainment to periodic acoustic stimuli. *Neuroscience Letter*, 424:55–60, 2007.

[82] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113(6):767–791, 2002.

[83] J. R. Wolpaw, D. J. McFarland, G. W. Neat, and C. Forneris. An EEG-based brain-computer interface for cursor control. *Electroencephalography and Clinical Neurophysiology*, 78:252–259, 1991.

[84] J. R. Wolpaw and E. W. Wolpaw, editors. *Brain-computer interfaces: principles and practice*. Oxford University Press, New York, 2012.

[85] Y. Yokota, S. Miyamoto, and Y. Naruse. Estimation of human workload from the auditory steady-state response recorded via a wearable electroencephalography system during walking. *Frontiers in Human Neuroscience*, 11(314), 2017.

[86] D. Zhang, E. Gong, W. Wu, J. Lin, W. Zhou, and B. Hong. Spoken sentences decoding based on intracranial high gamma response using dynamic time warping. In *Proceedings of Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3292–3295, San Diego, CA, USA, 2012.

[87] S. Zhao and F. Rudzicz. Classifying phonological categories in imagined and articulated speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 992–996, 2015.

[88] B. Zoefel and R. VanRullen. EEG oscillations entrain their phase to high-level features of speech sound. *NeuroImage*, 124:16–23, 2016.

A list of publication (As the first author)

- Journal papers

  1. Watanabe, H., Tanaka, H., Sakriani S., & Nakamura, S. (In Press). Synchronization between overt speech envelope and EEG oscillations during imagined speech, *Neuroscience Research*.
  2. Watanabe. H., Tanaka. H., Sakriani. S., & Nakamura. S. (2019). Neural oscillation-based classification of Japanese spoken sentences during speech perception, *IEICE Transactions on Information and Systems, Vol.E102-D (2)*, 383–391.

- International conference

  1. Watanabe, H., Tanaka, H., Sakti, S., & Nakamura, S. (2017). Subject-independent classification of Japanese spoken sentences by multiple frequency bands phase pattern of EEG response during speech perception, *Interspeech 2017*, pp. 2431-2435, Stockholm, Sweden, August, 2017.

- Domestic conference

  1. Watanabe. H., Tanaka, H., Sakti, S. & Nakamura, S. (2018). Sentence classification based on phase patterns in EEG neural oscillation during imagined speech, *The 41st Annual Meeting of the Japan Neuroscience Society*, Kobe, Japan, July. (peer reviewed)
  2. Watanabe. H., Tanaka, H., Sakti, S. & Nakamura, S. (2017). Subject-independent consistency of EEG phase patterns during auditory perception of Japanese spoken sentences, *The 40st Annual Meeting of the Japan Neuroscience Society*, Chiba, Japan, July. (peer reviewed)
  3. 渡部 宏樹, 田中 宏季, サクティ サクリアニ, 中村 哲. (2017). 音声知覚時における脳神経活動の位相パターンと SVM に基づく日本語音声刺激の識別, 信学技報*, 116(435)*, pp.9-14, 福岡, 日本, 2017 年 1 月.

A list of publication (As the second author)

- Journal papers

  1. Tanaka. H., Watanabe. H., Maki. H., Sakriani. S., & Nakamura. S. (2019). EEG-based single-trial detection of semantic and syntactic anomalies in listening to speech. *Frontiers in Computational Neuroscience*, 13: 15.

- International conferences

  1. Tanaka, H., Watanabe, H., Maki, H., Sakti, S., & Nakamura, S. (2018). Single-trial detection of semantic anomalies from EEG during listening to spoken sentences, *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 977-980, Honolulu, U.S.A., July, 2018.

2. Pintér, G., & Watanabe, H. (2016). Do GMM phoneme classifiers perceive synthetic sibilants as humans do?, *Interspeech 2016*, pp. 1363-1367, San Francisco, U.S.A., September, 2016.

- Domestic conference

1. 田中 宏季, 渡部 宏樹, 真木 勇人, サクティ サクリアニ, 中村 哲. (2017). 音声文聴取時における意味違反が生じた際の脳波自動判別, 信学技報, *117(375)*, pp.5-8, 宮城, 日本, 2017 年 12 月.

Competetive fundings

- International Information Science Foundation, *2016 Kenkyuusya Kaigai Haken Jyosei* ((公財) 情報科学国際交流財団 平成 28 年度研究者海外派遣助成)
- Foundation for Nara Institute of Science and Technology, *Kyoiku Katsudo Shien*, term: 2016/4/1 – 2017/3/31 ((公財) 奈良先端科学技術大学院大学支援財団 教育研究活動支援)
- Japan Society for the Promotion of Science, *Grants-in-Aid for JSPS Fellows*, term: 2018/04/01 - 2020/03/30 ((独)　日本学術振興会　特別研究員奨励費)