PAPER

Neural Oscillation-Based Classification of Japanese Spoken Sentences During Speech Perception

Hiroki WATANABE[†a)], ***Nonmember***, **Hiroki TANAKA**[†], **Sakriani SAKTI**[†,††], ***and*** **Satoshi NAKAMURA**[†], ***Members***

SUMMARY Brain-computer interfaces (BCIs) have been used by users to convey their intentions directly with brain signals. For example, a spelling system that uses EEGs allows letters on a display to be selected. In comparison, previous studies have investigated decoding speech information such as syllables, words from single-trial brain signals during speech comprehension, or articulatory imagination. Such decoding realizes speech recognition with a relatively short time-lag and without relying on a display. Previous magnetoencephalogram (MEG) research showed that a template matching method could be used to classify three English sentences by using phase patterns in theta oscillations. This method is based on the synchronization between speech rhythms and neural oscillations during speech processing, that is, theta oscillations synchronized with syllabic rhythms and low-gamma oscillations with phonemic rhythms. The present study aimed to approximate this classification method to a BCI application. To this end, (1) we investigated the performance of the EEG-based classification of three Japanese sentences and (2) evaluated the generalizability of our models to other different users. For the purpose of improving accuracy, (3) we investigated the performances of four classifiers: template matching (baseline), logistic regression, support vector machine, and random forest. In addition, (4) we propose using novel features including phase patterns in a higher frequency range. Our proposed features were constructed in order to capture synchronization in a low-gamma band, that is, (i) phases in EEG oscillations in the range of 2–50 Hz from all electrodes used for measuring EEG data (all) and (ii) phases selected on the basis of feature importance (selected). The classification results showed that, except for random forest, most classifiers perform similarly. Our proposed features improved the classification accuracy with statistical significance compared with a baseline feature, which is a phase pattern in neural oscillations in the range of 4–8 Hz from the right hemisphere. The best mean accuracy across folds was 55.9% using template matching trained by all features. We concluded that the use of phase information in a higher frequency band improves the performance of EEG-based sentence classification and that this model is applicable to other different users.

key words: *brain-computer interface, electroencephalogram (EEG), neural decoding, neural oscillations, phase-locking*

1. Introduction

A brain-computer interface (BCI) provides a way for physically handicapped people to compensate for lost bodily functions [1]. For example, locked-in syndrome patients cannot coordinate voluntary movements except for limited eye movements and blinks, even though they are awake with normal consciousness, and this means that it is difficult for them to express their intentions. Hence, a BCI is required for them to communicate with others without making body movements.

Event-related potential (ERP)-based typing systems that use scalp electroencephalograms (EEGs) are one of the most famous applications for communicating without movements, e.g., the P300 speller [2]. On this system, the user can select a letter on a monitor when the system detects a P300 that corresponds to a particular letter. However, regardless of the high usefulness of this ERP-based BCI, the system has some limitations [3]. First, a time-lag occurs from the timing when a user attempts to express his/her intention until all letters are selected. This is derived from an averaging process for detecting ERP. Second, communication is limited to places where a monitor display can be set.

Attempts to recognize speech information from single-trial brain activity during linguistic processing, e.g., a speech comprehension [4]–[8] or articulatory imagination [9]–[12], overcome the above-mentioned limitations. A dramatic reduction in time-lag shows the potential for real-time brain-based communication [3]. One possible neurophysiological mechanism for enabling such brain-based speech recognition is the synchronization between neural oscillations during speech processing and speech rhythms, that is, the phase of theta oscillations (~4–8 Hz) in listeners' auditory cortexes matches speech envelopes (this is also mainly ~4–8 Hz) consisting of syllabic rhythm, i.e., phase-locking (see [13] for a review). This mechanism reproduces consistent neural phase patterns for a specific sentence and different patterns for different ones. On the basis of this mechanism, magnetoencephalogram (MEG) phase patterns in theta oscillations during speech processing were used to classify three spoken English sentences [7], [8].

The current research aims to fill a gap between the previous, neurophysiologically-motivated classification methods and BCI applications. First, applying MEG-based sentence classification as in previous research for a BCI system is hard. This is because the apparatuses for measuring MEGs are too large for BCI systems, and the running costs of the apparatuses are generally high. EEGs can be measured by using a relatively compact device with a lower running cost than MEGs, but it is argued that although the characteristics of EEGs seem more suitable for BCI systems, the accuracy of EEG-based decoding is lower than MEG-based decoding as reported in previous neural decoding research [14]. However, the performance of EEG-based

Manuscript received August 21, 2018.
Manuscript publicized November 14, 2018.
[†]The authors are with Nara Institute of Science and Technology, Ikoma-shi, 630–0192 Japan.
[††]The author is with RIKEN, Center for Advanced Intelligence Project AIP, Ikoma-shi, 630–0192 Japan.
a) E-mail: watanabe.hiroki.vx6@is.naist.jp
DOI: 10.1587/transinf.2018EDP7293

sentence classification with phase synchronization between neural oscillations and speech rhythms has not been explored. Therefore, in this study, we investigated the performance of EEG-based sentence classification.

Second, the existing approach was used to construct only subject-dependent models. Concerning BCI applications, this, however, indicates that we need 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 we train subject-independent models, we prepare a model for 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 we want to construct a BCI application without the need to retrain a model with a new subject. However, according to previously reported studies, the classification performance when using subject-independent neural decoding [4], [14] might be worse than the performance when using subject-dependent decoding [15]. However, Kerlin et al. [16] demonstrated that phase-locked responses are replicable phase patterns across different listeners. Thus, in the current research, we performed subject-independent classification on the basis of phase-locked responses. In particular, we investigated how accurately our model classified spoken Japanese sentences when evaluating models by leave-one-subject-out cross-validation (LOSO), which tests performances using the data of unknown subjects.

Third, in previous research, only template matching was used for classifying spoken sentences, which is based on the Euclidean distance between a template and test data. In the domain of BCI, various types of algorithms have been utilized for classification [17]. This indicates that the use of different types of classifiers leads to an improvement in classification accuracy in EEG-based sentence classification. Thus, in the current study, we investigated the performances of four types of classifiers: (a) template matching (baseline), (b) logistic regression, (c) support vector machine (SVM), and (d) random forest. These models work well for small-scale and high-dimensional data such as EEGs.

Finally, the previous method of classification used phase information in theta oscillations as a feature to capture the synchronization between theta oscillation phases and syllabic rhythm. Neural oscillations in a higher frequency band synchronize with phonemic level rhythm [18], [19]. This parallel synchronization in a distinct frequency band shows functional lateralization; syllabic information extraction preferentially relies on the right hemisphere (RH), and phonemic information extraction relies preferentially on the left hemisphere (LH) [18]. Such synchronization in the higher frequency band leads to the hypothesis that phase patterns in broader frequency ranges in different hemispheres provide additional information for classification. Thus, in addition to the baseline feature, we propose three features for the purpose of capturing synchronization in a higher frequency band: (i) a theta-RH feature (baseline), i.e., phase patterns extracted from the theta frequency range (4–8 Hz) in the right hemisphere, (ii) all features, i.e., phase patterns extracted from a broader frequency range (2–50 Hz) from all electrodes used for measuring EEG data, and (iii) selected features, i.e., phase patterns selected from all features on the basis of feature importance. The selected features were prepared for the sake of avoiding overfitting because the high dimensionality of EEGs may hinder classification performance.

In summary, in this study, we perform EEG-based classification for three Japanese sentences by using a subject-independent model and four types of classifiers (template matching, logistic regression, SVM, and random forest). We also propose using new features including phase patterns in a higher frequency range. Specifically, we focus our research on addressing the following questions[†]:

(1) How accurately do our subject-independent EEG-based models classify the three Japanese spoken sentences?
(2) Which of the three classifiers improved classification accuracy over template matching (the baseline classifier)?
(3) Do our proposed features including phase patterns in a higher frequency band improve classification accuracy?

2. EEG Data Recordings

2.1 Participants

Seventeen right-handed L1 Japanese speakers participated in data recordings. One female participant was excluded from the analysis because she changed her dominant hand from left to right during childhood. The average age and lateralization quotient for the handedness [20] of the remaining participants (6 females, 10 males) was 24.2 ± 2.0 and 90.15 ± 12.3, respectively. All participants agreed to participate and gave informed consent in writing. They all reported no any 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.

2.2 Spoken Sentence Stimuli

We constructed three Japanese sentences for our classification task (Table 1). 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).

[†]This study is an extended work from our conference proceeding [15].

Table 1 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.*)

2.3 Apparatus for EEG Data Recording

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 (EASYCAP GmbH., Germany) according to the 10% system. Electrode impedance was kept below 10 kΩ before the recording. Raw EEG data were 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).

2.4 Procedure for Recording EEG Data

The sentences were presented to each participant aurally. A trial included one behavioral task based on previous research [8] with the aim of keeping participants' attention on the stimuli. 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. 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. Before the 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.

Before presenting the sentences, 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. The sentences in a pair were played in sequence as follows. (1) A sentence, "Are you ready?", appeared on the display. Participants started a trial by pushing the space key. (2) A fixation mark (+) appeared at the center of the display. (3) The first sentence in a pair was played at 1,500 ms from the trial onset. (4) The second sentence was played at 7,500 ms. (5) A short tone sound was played at 12,000 ms. (6) Participants

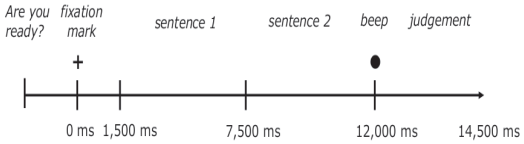

Fig. 1 Procedure of one trial.

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. We summarized the procedure of one trial in Fig. 1.

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 exact same procedure. Each sentence was presented 24 times to each participant. The EEG data recording lasted approximately for 1 to 1.5 hours including preparation.

3. EEG Data Analysis

3.1 Preprocessing of EEG

The FieldTrip toolbox for MATLAB (The MathWorks, Inc., U.S.A) was used for EEG data analysis [21]. EEG data were epoched from −500 to 2,900 ms relative to the onset of speech. To remove slow drift, a sixth-order two-pass IIR Butterworth high-pass filter (1 Hz) was applied to the EEG data. Line noise at 60 Hz was removed by using a discrete Fourier transform filter. During the filtering procedure, buffer zones were added before and after EEG epochs in order to avoid edge artifacts. EEG data were detrended and baseline-corrected [−500 ms, 0 ms]. Electrodes contaminated with large noise over most all of the trials were detected by visual inspection. One electrode was removed from the data of two participants.

Next, we rejected trials contaminated with large amplitude artifacts and muscle artifacts. For large amplitude artifacts, trials exceeding ±350 μv were removed from further analysis. 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 15. After this automatic procedure, we judged whether a trial that was automatically judged as having an artifact included muscle artifacts visually. In total, 4.8% of trials across all participants were removed. One-way factorial analysis of variance revealed no significant differences in the number of rejected trials among sentence types (average number of rejected trials across participants, *sentence 1* = 1.00, *sentence 2* = 1.25, *sentence 3* = 1.19,

$F(2, 45) = 0.116$, $p = 0.89$).

EEG data were decomposed of independent components (ICs) by independent components analysis. 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, EEG data were re-referenced by using a common average reference. Rejected electrodes were interpolated by using neighboring electrodes. EEG data were low-pass filtered by a sixth-order two-pass IIR Butterworth filter (50 Hz) to improve the signal-to-noise ratio.

3.2 Cross-Trial Phase Coherence (Cphase)

For the purpose of determining whether phase-locked responses occurred, we quantified the degree of the response in each frequency band (theta: 4–8 Hz, alpha: 10–14 Hz, beta: 16–20 Hz, low-gamma: 38–42 Hz) by using cross-trial phase coherence (Cphase) [7], [8], [19]. This is an index of the phase coherence among EEG responses to a spoken sentence (range = [0, 1], 0: no phase coherence, 1: perfect coherence). Cphase was calculated per electrode and frequency band.

First, we extracted phase information from EEG trials by using short-time Fourier transform (STFT; FFT points: 500, shift points: 100, Hanning window tapering, 29 windows in total; duration of EEG trials: 2,900 ms). Cphase was calculated with concatenated data across all participants by using Eq. (1) [7], [8], [19]:

$$Cphase_{i,j} = \frac{1}{K}\sum_{k=1}^{K}\left[\left[\frac{\sum_{n=1}^{N}\cos(\theta_{knij})}{N}\right]^2+\left[\frac{\sum_{n=1}^{N}\sin(\theta_{knij})}{N}\right]^2\right]. \quad (1)$$

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. To calculate Cphase for each frequency band, $Cphase_{i,j}$ was averaged across shifting windows, followed by averaging the mean Cphase among frequency bins belonging to the frequency band, e.g., $Cphase_{theta}$ was averaged values among $Cphase_{4Hz}$, $Cphase_{6Hz}$, and $Cphase_{8Hz}$. Because a speech onset produces an ERP, which induces strong coherence across trials, we used a time range after 500 ms from the onset of speech in order to average the Cphase values in the time domain.

4. Spoken Sentence Classification

4.1 Feature Extraction

We created a vector (*hereafter, a phase pattern*); each element of the vector is a phase value calculated in each shifting window (total of 29 windows) in STFT (see Sect. 3.2). A phase pattern was prepared per frequency bin (2–50 Hz; 2-Hz interval). We created the following three types of feature vectors from the phase patterns.

(1) Theta-RH
This feature was a concatenated vector of phase patterns at 4, 6, and 8 Hz from the right hemisphere electrodes (C4, CP2, CP6, F4, F8, FC2, FC6, FT10, Fp2, O2, P4, P8, T8, and TP10). The theta-RH features had 1,218 dimensions (14 electrodes × 29 shifting windows × 3 frequency bins: 4, 6, and 8 Hz). In previous research, theta phase patterns were extracted from electrodes selected by using Cphase values [7], [8], while we did not use Cphase values themselves to select electrodes. Thus, this feature was not exactly the same as that in previous research. However, considering that phase-locking in theta oscillations was observed dominantly in the right hemisphere [18], we treated this feature as a baseline.

(2) All
All features were a concatenated vector of phase patterns in the range of 2–50 Hz from all 32 electrodes. All features had 23,200 dimensions (32 electrodes × 29 shifting windows × 25 frequency bins: 2–50 Hz, 2-Hz interval). This feature was constructed in order to utilize the synchronization between speech and neural oscillations at a phonemic level.

(3) Selected
To avoid overfitting due to the high dimensionality of all features, features were selected by using feature importance calculated by each trained classifier. We identified the best 20 important combinations of a frequency bin × an electrode. The importance of the combination was calculated as follows. (a) Classifiers were trained by using all features at each cross-validation step (see Sect. 4.2). (b) Feature importances were averaged among each training dataset. (c) The averaged feature importances were z-normalized and averaged across the shifting windows per combination of an electrode and a frequency bin. The selected features were a concatenated vector of phase patterns from the highest 20 electrode × frequency combinations. The number of dimensions was 580 (20 combinations × 29 shifting windows).

4.2 Classifiers and Evaluation Method

We trained template matching, logistic regression, SVM, and random forest. The classification task was to predict a sentence by using phase patterns in a single-trial EEG. We used a Python library, Scikit-learn [22], and custom Python scripts for training and evaluating classifiers. We evaluated the classification performance by using LOSO with the aim of confirming the generalizability of our models to other users. At each validation step, the feature importance was calculated per classifier in order to use it for feature selection (Sect. 4.1).

Except for template matching, the best parameters of each classifier were estimated by using a grid search. We used the data of one subject as a test set and the data of the

next subject as a development set for parameter tuning. The remaining data were used for training. In template matching, we used the data of one subject as a test set and the remaining data as a training set. The descriptions of classifiers are as follows.

(1) Template matching
We 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. 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
Classifiers were trained by using L2 regularization. We used a one-vs-the-rest multiclass strategy. The best cost parameter was searched for in the range [−4, 4] in log space. We used the variance of coefficients assigned to each feature among one-vs-the-rest classifiers as feature importance.

(3) SVM
A linear kernel was used, and the one-vs-the-rest multiclass strategy was adopted. The tuning of a cost parameter, type of regularization, and procedure for calculating feature importance were the same as logistic regression.

(4) Random forest
The best number of trees and maximum tree depth were searched for in the ranges [10, 50, 100, 150] and [5, 10, 15], respectively.

5. Results

5.1 Phase-Locking to Speech Rhythms

To determine whether phase-locked responses to speech stimuli occurred, we plotted Cphase topographies per frequency band (Fig. 2). A visual inspection suggested a relatively strong Cphase on the fronto-central region for theta. Whereas similar distributional patterns to theta were observed for alpha and low-gamma, Cphase for beta was lower than the other frequency bands.

To determine whether the Cphase at each electrode in each frequency band was statistically significantly different from a null-hypothesis, i.e., the Cphase at an electrode in a frequency band is located at the chance level, we created a permutation distribution of null-hypotheses from observed data by shuffling the sentence labels [23]. For the sake of controlling multiple comparisons, we took the maximum and minimum Cphase from randomly shuffled data among all electrodes in all frequency bands to distribute the null-hypotheses [23]. The alpha level was set to 0.05. The

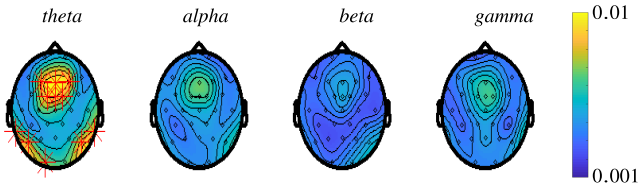

Fig. 2 Topographies of Cphase for each frequency band. Red star represents electrode with statistically significant difference from null-distribution.

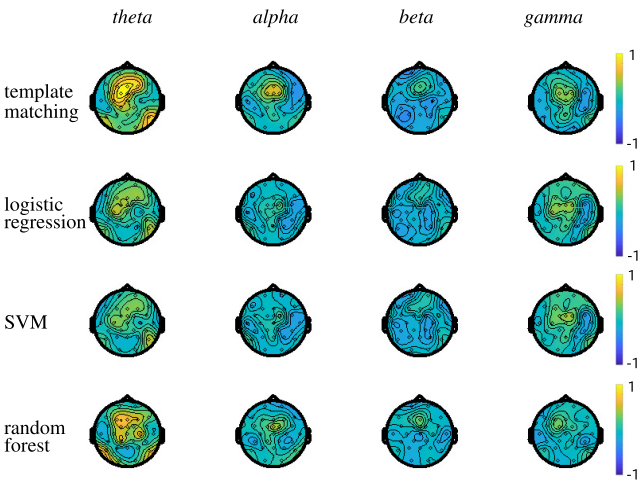

Fig. 3 Topographies of feature importance for each frequency band and each classifier. Feature importance was z-normalized.

number of iterations was 2,000. We counted the number of values of the null-hypothesis distribution exceeding the observed Cphase for each electrode and each frequency band. The number was divided by the number of iterations for calculating p values.

A two-tailed test revealed that the Cphase values in the fronto-central region for theta were statistically significantly larger that the null-hypothesis (electrodes represented by red stars in Fig. 2). The other Cphase values were not statistically significant.

5.2 Topographies of Feature Importance

We determined the feature importance topographies per frequency band and classifier (Fig. 3) because an obvious mismatch between the topographies of feature importance and phase-locked responses suggested that the classifiers failed to utilize phase-locked responses for classification. An overall visual inspection suggested that the feature importance topographies of all classifiers had similar distributional patterns to Cphase topographies: the fronto-central region for theta and low-gamma. In addition, in the fronto-central region for alpha, random forest and template matching were relatively high, but they had weaker feature importance than theta.

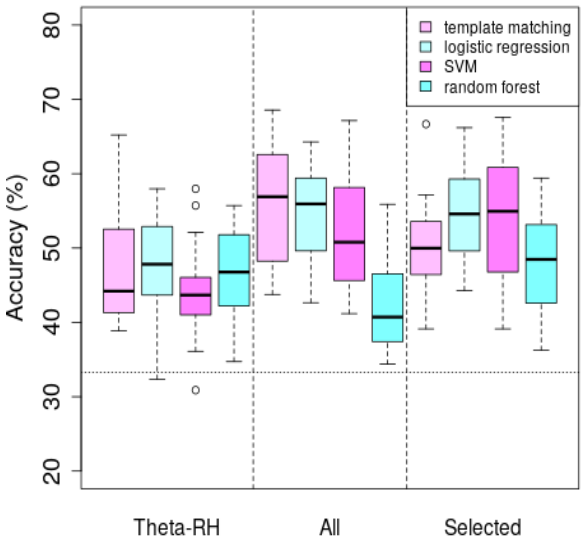

Fig. 4 Classification accuracies for each classifier and feature. Each box represents accuracies from all folds. Horizontal line is 33.3% chance level.

Table 2 Mean accuracies across folds per classifier and feature type. Sample standard deviations are given in parentheses. Best accuracy is shown in bold.

	Theta-RH	All	Selected
Template matching	46.9 (7.1)	**55.9** (7.9)	50.3 (6.2)
Logistic regression	46.5 (7.5)	54.3 (6.6)	54.6 (7.0)
SVM	44.2 (6.6)	52.3 (7.6)	53.9 (8.7)
Random forest	46.4 (6.1)	42.7 (6.5)	47.7 (6.6)

5.3 Classification Performances

Figure 4 summarizes the classification accuracies from folds per combination of classifiers and features. The best mean accuracy across folds was 55.9% for template matching based on all features (Table 2).

To evaluate the effect of classifiers and features on accuracy, we constructed a generalized linear mixed model (GLMM) by using the number of correct classifications as response variables. We used R [24] and an *lme4* package [25] 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 fold). A logit link function was used for the model. Types of classifiers and features were incorporated into a GLMM as fixed effects. Our 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, we found a statistically significant effect of classifiers [$\chi^2(3) = 31.33$, $p < 0.01$] and features [$\chi^2(2) = 36.19$, $p < 0.01$].

For the purpose of determining what classifiers and features affected accuracies, we performed multiple comparisons for each fixed effect by using a *multcomp* package [27]. *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 significantly (Table 3). The other classifiers did not differ from each other. Multiple comparisons among feature types revealed that both all and selected features improved accuracy compared with the baseline feature (Table 4).

Finally, we determined whether a combination of each classifier and feature improved accuracy compared with a baseline method (template matching based on theta-RH). To that end, classification accuracies from all models were compared to that from the baseline by using the Dunnett method for *p*-value adjustment. The results of the comparisons are summarized in Table 5. The best performances, from template matching (all features), showed a statistically significant improvement from the baseline. In addition, logistic regression (all features and selected features) and SVM (selected features) also showed a statistically significant improvement.

Table 3 Multiple comparisons for classifier types.

	Estimate (S.E)
Logistic regression -vs- Template matching	0.032 (4.955e-2)
SVM -vs- Template matching	−0.034 (4.954e-2)
Random forest -vs- Template matching	−0.221 (4.964e-2)**
SVM -vs- Logistic regression	−0.066 (4.955e-2)
Random forest -vs- Logistic regression	−0.253 (4.965e-2)**
Random forest -vs- SVM	−0.187 (4.963e-2)**

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

Table 4 Multiple comparisons for feature types.

	Estimate (S.E)
All -vs- Theta-RH	0.220 (4.297e-2)**
Selected -vs- Theta-RH	0.228 (4.297e-2)**
Selected -vs- All	0.007 (4.291e-2)

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

Table 5 Multiple comparisons between baseline model and all classifiers trained by each feature using Dunnett method for p-value adjustment.

	Estimate (S.E)
Template matching	
All -vs- baseline	0.376 (8.609e-2)**
Selected -vs- baseline	0.140 (8.577e-2)
Logistic regression	
Theta-RH -vs- baseline	−0.011 (8.587e-2)
All -vs- baseline	0.309 (8.594e-2)**
Selected -vs- baseline	0.313 (8.594e-2)**
SVM	
Theta-RH -vs- baseline	−0.104 (8.606e-2)
All -vs- baseline	0.228 (8.582e-2)*
Selected -vs- baseline	0.287 (8.590e-2)**
Random forest	
Theta-RH -vs- baseline	−0.018 (8.589e-2)
All -vs- baseline	−0.167 (8.624e-2)
Selected -vs- baseline	0.037 (8.582e-2)

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

6. Discussion

In the current research, we investigated the performance of sentence classification based on single-trial EEG phase patterns during speech processing. Our research aimed to answer three research questions: (1) How accurately do our subject-independent EEG-based models classify the three Japanese spoken sentences?, (2) Which of the three classifiers improved classification accuracy over template matching (the baseline classifier)?, and (3) Do our proposed features including phase patterns in a higher frequency band improve classification accuracy?

As for research question (1), our classification accuracy was 55.9% over template matching trained with all features. Considering that the spatial resolution of an EEG is worse than that of an MEG, this accuracy might be promising. Model evaluation by LOSO indicates that this model can be applied to other different users. We expected that such subject-independent classification was possible [15] because phase-locked responses to speech rhythms are consistent across listeners [16]. The second research question focused on the effect of classifier types on classification performances. Whereas the best accuracy was obtained from template matching [7], [8], in our GLMM analysis, the classifiers used in the current research showed similar performances to template matching except for random forest, which showed a statistically worse accuracy. The current spoken sentence classification is based on a phase synchronization mechanism between neural oscillations and speech rhythms. This means that an EEG phase value at each time point takes a value close to a phase value in speech envelopes. Thus, it is thought that all linear classifiers used in the current experiment captured such a linear relationship between EEGs and speech successfully.

Finally, as for the third research question, our two proposed features (all and selected) contributed to improving the classification accuracy over the baseline feature. As discussed in the introduction, this performance improvement is thought to be obtained from phase-locked responses in a higher frequency band at a phonemic level.

7. Conclusion and Future Directions

We concluded that the use of phase patterns in a higher frequency range improved accuracy in EEG-based sentence classification by capturing phase-locked responses at a phonemic level.

There are some future directions for approximating this classification method closer to a BCI application. A first direction is related to improving accuracy further. The accuracy in our classification is still not enough for practical use in BCI applications. Thus, other neural oscillation features related to language processing might have a role in further improvement. Recent neurophysiological research demonstrated a relationship between neural oscillation phases and the boundaries of syntactic phrasing [28], [29], sentences [29], and intonation phrases [30]. Utilizing this information in neural oscillatory dynamics might improve accuracy. Another direction is to use multiple speech stimuli recorded from various speakers to test robustness against acoustic variability. In the current research, each speech stimulus was recorded from one single female speaker; thus, this robustness needs to be investigated in the future. In connection with this direction, a final direction is the use of a larger number of sentences; if the number of sentences to be classified were increased, this EEG-based classification would be closer to a BCI application.

Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Numbers 18J14871, 16K16172, 17K00237, and 17H06101.

References

[1] J.R. Wolpaw and E.W. Wolpaw, eds., Brain-computer interfaces: principles and practice, Oxford University Press, New York, 2012.
[2] 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, vol.70, no.6, pp.510–523, 1988.
[3] J.S. Brumberg, A. Nieto-Castanon, P.R. Kennedy, and F.H. Guenther, "Brain-computer interfaces for speech communication," Speech Communication, vol.52, no.4, pp.367–379, 2010.
[4] P. Suppes, Z.-L. Lu, and B. Han, "Brain wave recognition of words," In Proceedings of the National Academy of Sciences, vol.94, no.26, pp.14965–14969, 1997.
[5] 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, vol.6, no.71, pp.1–10, 2015.
[6] P. Suppes, B. Han, and Z.-L. Lu, "Brain-wave recognition of sentences," In Proceedings of the National Academy of Sciences, vol.95, no.26, pp.15861–15866, 1998.
[7] H. Luo and D. Poeppel, "Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex," Neuron, vol.54, no.6, pp.1001–1010, 2007.
[8] 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, vol.104, no.5, pp.2500–2511, 2010.
[9] C.S. DaSalla, H. Kambara, M. Sato, and Y. Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," Neural Networks, vol.22, no.9, pp.1334–1339, 2009.
[10] X. Chi, J.B. Hagedorn, D. Schoonover, and M. D'Zmura, "EEG-based discrimination of imagined speech phonemes," International Journal of Bioelectromagnetism, vol.13, no.4, pp.201–206, 2011.
[11] M. D'Zmura, S. Deng, T. Lappas, S. Thorpe, and R. Srinivasan, "Toward EEG sensing of imagined speech," In Human-Computer Interaction, New Trends, Springer, Berlin, Heidelberg, vol.5610, pp.40–48, 2009.
[12] K. Brigham and B.V.K.V. 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, pp.1–4, 2010.
[13] J.E. Peelle and M.H. Davis, "Neural oscillations carry speech rhythm through to comprehension," Frontiers in Pyschology, vol.3, no.320, pp.1–17, 2012.

[14] A.M. Chan, E. Halgren, K. Marinkovic, and S.S. Cash, "Decoding word and category-specific spatiotemporal representations from MEG and EEG," Neuroimage, vol.54, no.4, pp.3028–3039, 2011.
[15] 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, Stockholm, Sweden, pp.2431–2435, 2017.
[16] 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, vol.30, no.2, pp.620–628, 2010.
[17] F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi, "A review of classification algorithms for EEG-based brain-computer interfaces," Journal of neural engineering, vol.4, no.2, pp.1–24, 2007.
[18] D. Poeppel, "The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'," Speech Communication, vol.41, no.1, pp.245–255, 2003.
[19] H. Luo and D. Poeppel, "Cortical oscillations in auditory perception and speech: evidence for two temporal windows in human auditory cortex," Frontiers in Psychology, vol.3, no.170, pp.1–10, 2012.
[20] R.C. Oldfield, "The assessment and analysis of handedness: The Edinburgh inventory," Neuropsychologia, vol.9, no.1, pp.97–113, 1971.
[21] 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, vol.2011, pp.1–9, 2011.
[22] 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, vol.12, pp.2825–2830, 2011.
[23] E. Maris, J.-M. Schoffelen, and P. Fries, "Nonparametric statistical testing of coherence differences," Journal of Neuroscience Methods, vol.163, no.1, pp.161–175, 2007.
[24] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2018.
[25] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," Journal of Statistical Software, vol.67, no.1, pp.1–48, 2015.
[26] J. Fox and S. Weisberg, An R companion to applied regression, 2nd ed., Sage, Thousand Oaks CA, 2011.
[27] T. Hothorn, F. Bretz, and P. Westfall, "Simultaneous inference in general parametric models," Biometrical Journal, vol.50, no.3, pp.346–363, 2008.
[28] 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, vol.27, no.9, pp.4293–4302, 2016.
[29] N. Ding, L. Melloni, H. Zhang, X. Tian, and D. Poeppel, "Cortical tracking of hierarchical linguistic structures in connected speech," Nature neuroscience, vol.19, no.1, pp.158–164, 2016.
[30] M. Bourguignon, X.D. Tiège, M.O. de Beeck, N. Ligot, P. Paquier, P.V. 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, vol.34, no.2, pp.314–326, 2013.

Hiroki Watanabe received the B.A. degree from Ritsumeikan University, Kyoto, Japan in 2012 and the M.A. degree from Kobe University, Hyogo, Japan in 2014. He is currently a Ph.D. student at Nara Institute of Science and Technology. His research interest includes neurolinguistics and brain-computer interface.

Hiroki Tanaka received the master's and Ph.D. degrees from the Nara Institute of Science and Technology, Japan, in 2012 and 2015, respectively. He is an Assistant Professor with the Graduate School of Information Science, Nara Institute of Science and Technology. His research interest is assisting people with disabilities through human-computer interaction.

Sakriani Sakti received her B.E. degree in Informatics (cum laude) from Bandung Institute of Technology, Indonesia, in 1999. In 2000, she received DAAD-Siemens Program Asia 21st Century Award to study in Communication Technology, University of Ulm, Germany, and received her MSc degree in 2002. During her thesis work, she worked with Speech Understanding Department, DaimlerChrysler Research Center, Ulm, Germany. Between 2003–2009, she worked as a researcher at ATR SLC Labs, Japan, and during 2006–2011, she worked as an expert researcher at NICT SLC Groups, Japan. While working with ATR-NICT, Japan, she continued her study (2005–2008) with Dialog Systems Group University of Ulm, Germany, and received her PhD degree in 2008. She actively involved in collaboration activities such as Asian Pacific Telecommunity Project (2003–2007), A-STAR and U-STAR (2006–2011). In 2009–2011, she served as a visiting professor of Computer Science Department, University of Indonesia (UI), Indonesia. From 2011, she has been an assistant professor at the Augmented Human Communication Laboratory, NAIST, Japan. She served also as a visiting scientific researcher of INRIA Paris-Rocquencourt, France, in 2015–2016, under "JSPS Strategic Young Researcher Overseas Visits Program for Accelerating Brain Circulation". She is a member of JNS, SFN, ASJ, ISCA, IEICE and IEEE. She is now also a board member of SLTU (Spoken Language Technologies for Under-resourced languages) and a member of ISCA/ELRA SIG-UL (a joint Special Interest Group for Under-resourced Languages). Her research interests lie in deep learning & graphical model framework, statistical pattern recognition, zero-resourced speech technology, multilingual speech recognition and synthesis, spoken language translation, social-affective dialog system, and cognitive communication.

Satoshi Nakamura is Professor of Graduate School of Information Science, Nara Institute of Science and Technology, Japan, Honorarprofessor of Karlsruhe Institute of Technology, Germany, and ATR Fellow. He received his B.S. from Kyoto Institute of Technology in 1981 and Ph.D. from Kyoto University in 1992. He was Associate Professor of Graduate School of Information Science at Nara Institute of Science and Technology in 1994–2000. He was Director of ATR Spoken Language Communication Research Laboratories in 2000–2008 and Vice president of ATR in 2007–2008. He was Director General of Keihanna Research Laboratories and the Executive Director of Knowledge Creating Communication Research Center, National Institute of Information and Communications Technology, Japan in 2009–2010. He is currently Director of Augmented Human Communication laboratory and a full professor of Graduate School of Information Science at Nara Institute of Science and Technology. He is interested in modeling and systems of speech-to-speech translation and speech recognition. He is one of the leaders of speech-to-speech translation research and has been serving for various speech-to-speech translation research projects in the world including C-STAR, IWSLT and A-STAR. He received Yamashita Research Award, Kiyasu Award from the Information Processing Society of Japan, Telecom System Award, AAMT Nagao Award, Docomo Mobile Science Award in 2007, ASJ Award for Distinguished Achievements in Acoustics. He received the Commendation for Science and Technology by the Minister of Education, Science and Technology, and the Commendation for Science and Technology by the Minister of Internal Affair and Communications. He also received LREC Antonio Zampolli Award 2012. He has been Elected Board Member of International Speech Communication Association, ISCA, since June 2011, IEEE Signal Processing Magazine Editorial Board Member since April 2012, IEEE SPS Speech and Language Technical Committee Member since 2013, and IEEE Fellow since 2016.