INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Subject-independent Classification of Japanese Spoken Sentences by Multiple Frequency Bands Phase Pattern of EEG Response during Speech Perception

Hiroki Watanabe, Hiroki Tanaka, Sakriani Sakti, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology, Japan
{watanabe.hiroki.vx6, hiroki-tan, ssakti, s-nakamura}@is.naist.jp

Abstract

Recent speech perception models propose that neural oscillations in theta band show phase locking to speech envelope to extract syllabic information and rapid temporal information is processed by the corresponding higher frequency band (e.g., low gamma). It is suggested that phase-locked responses to acoustic features show consistent patterns across subjects. Previous magnetoencephalographic (MEG) experiment showed that subject-dependent template matching classification by theta phase patterns could discriminate three English spoken sentences. In this paper, we adopt electroencephalography (EEG) to the spoken sentence discrimination on Japanese language, and we investigate the performances in various different settings by using: (1) template matching and support vector machine (SVM) classifiers; (2) subject dependent and independent models; (3) multiple frequency bands including theta, alpha, beta, low gamma, and the combination of all frequency bands. The performances in almost settings were higher than the chance level. While performances of SVM and template matching did not differ, the performance with combination of multiple frequency bands outperformed the one that trained only on single frequency bands. Best accuracies in subject dependent and independent models achieved 55.2% by SVM on the combination of all frequency bands and 44.0% by template matching on the combination of all frequency bands, respectively.

Index Terms: phase-locking, EEG, subject-independent classification, neural oscillation

1. Introduction

In human speech communication, an acoustic speech waveform conveys a message from a speaker to a listener. This communication medium involves a rhythmic phenomenon where the dominant modulation frequency reflects the sequential rate of the syllabic information. The ability of the listener to extract the linguistic information contained in the acoustic speech depends on how the brain oscillations entrain the speech input rhythm and parse the incoming information. One mechanism that supports such auditory processing is the phase-locking between the speech amplitude and theta band oscillation (4-8 Hz; ∼125-250 ms) in the auditory cortex (for a review see [1]). Because neural oscillation is related to the excitability of neuronal populations [2], phase-locking between acoustic information and neural oscillation enables acoustic processing during the high excitability of neuronal populations [1].

Acoustic information in speech also includes such rapid temporal information as segmental features. Recently, for processing such rapid information, parallel and concurrent information processing were postulated in theta neural oscillation and the distinct higher frequency bands of neural oscillation [3], [4], [5], [6]. For example, the Asymmetric Sampling in Time (AST) hypothesis [3] argued that theta oscillation in the right hemisphere extracts acoustic information in a relatively long time scale (150-250 ms; e.g., a syllable) and gamma oscillation in the left hemisphere extracts rapid spectral changes (20-40 ms; e.g., formant transition). A previous MEG experiment, which supports the AST hypothesis, demonstrated that brain responses to non-speech with the corresponding temporal structure to theta (∼200 ms; ∼5 Hz) and low gamma (∼25 ms; ∼40 Hz) showed reliable phase-locking, but non-speech stimuli with a temporal structure that corresponds to alpha (∼80 ms; ∼12.5 Hz) did not [7].

The phase-locked responses to input speech rhythms can be used for neural response-based spoken sentence discrimination [8], [9]. If the phase in the neural oscillation entrains the input speech rhythm, the phase patterns in the neural oscillation during processing the same spoken sentence are highly replicable. Recent MEG experiments demonstrated such replicable phase patterns across trials and found that subject-dependent template matching with theta phase-locked responses during speech processing discriminated three English spoken sentences [8], [9]. Such a neural response-based spoken speech recognition system might be an advantage of brain-computer interface (BCI) systems. Subject independency is another characteristic that phase-locked responses validate for BCI applications because such responses are consistent across other listeners [10], [11]. Given that subject-independent models can be trained based on phase-locked responses, a data collection benefit is obtained since a large amount of neuronal response data does not have to be measured from one participant for model training. However, classification performances with other languages, other classifiers, or other frequency bands remain unclear in spoken sentence classification using phase-locked responses. In addition, while EEG, which can be measured non-invasively with relatively low cost and a compact apparatus, is suitable for BCI systems, its poor spatial resolution might degrade classification performances compared to MEG [12].

In this study, we adopt EEG to spoken sentence discrimination on Japanese and investigate the performances of EEG-based spoken sentence classification in various settings using the following three schemes: (1) template matching and SVM classifiers; (2) subject-dependent and -independent models; (3) multiple frequency bands including theta, alpha, beta, low gamma, and a combination of all the frequency bands.

2. Method

2.1. EEG data collection

2.1.1. Participants

Ten L1 Japanese speakers participated in our experiment (males: 6, females: 4, mean age=24.3, SD=1.8), all of whom reported right-handness, no history of neurological problems, and normal hearing. Our experiment was approved by the ethical

http://dx.doi.org/10.21437/Interspeech.2017-854

review board of the Nara Institute of Science and Technology. Written informed consent was obtained from all of them.

2.1.2. Speech stimuli

Three Japanese spoken sentences were recorded by a female L1 Japanese speaker in a soundproof room at 44.1 kHz and 16-bit resolution:
(1) Anataga kinou muchuude yondeita honwa omoshirokatta
(*The book that you were absorbed in yesterday is interesting.*),
(2) Tsui sakki onnanokoga watashini itta kotowa hontouno-hanashi
(*What the girl said to me just now is true.*),
(3) Mukou no kabeni kazatteirunowa kareno oniisanga kaita e
(*The picture on the other wall was drawn by his older brother.*).

The speaker read these three sentences aloud at a normal speed with a declarative intonation and without any pauses. The average duration of the sentences was 3,146 ms. The duration of the moras, which is the rhythm unit of Japanese, were calculated manually using the Praat software [13]. Fig. 1 shows their duration in all of the spoken sentence stimuli. The peak was obtained around 6-8 Hz, which corresponds to the theta frequency range.

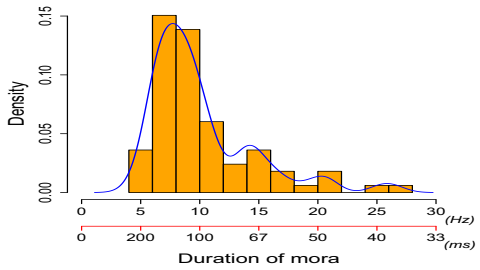

Figure 1: *Histogram of average duration of moras in spoken sentences used in experiment*

2.1.3. Apparatus

The EEG data were recorded using 32 Ag/AgCl electrodes on an EEG cap based on a 10% system (Easycap) at a 1,000-Hz sampling rate with a high-pass filter at 0.016 Hz and a low-pass filter at 250 Hz. The additional reference and ground electrodes were respectively placed on the AFz position and the right earlobe. Impedance was kept below 10 kΩ during the experiment. Because an electrode from one participant showed an increase in impedance above 10 kΩ after the experimental session, that electrode was rejected from further analysis. Speech stimuli were binaurally presented to the participants using software (Presentation, Neurobehavioral Systems) by earphones (ER-1, Etymotic Research).

2.1.4. Experimental procedure

Participants sat on the chair located in front of a PC display and placed their left index finger on the keyboard's F key, their right index finger on the J key, and remained motionless without blinking (as much as possible) while the speech stimuli were playing. The speech stimuli were presented based on previous research [9]. First, all possible nine pairs of spoken sentences (including pairs of the same sentence) were created. The order of the pairs was randomized, and each spoken sentence was presented as follows: (1) The sentence "Are you ready?" was presented on the display. Participants started a trial by pushing the space key; (2) at 1,500 ms from the trial onset, the first sentence of the pair was presented, and the second sentence was presented at 7,500 ms; (3) at 12,000 ms from the onset, a short beep sound was played. In order to direct participants' attention to speech stimuli, participants were instructed to judge whether the two spoken sentences were the same or different. If the two sentences were the same, they pressed the F key, and if not, they pressed the J key; (4) the trial duration was fixed to 14,500 ms.

Participants repeated the trials until all the pairs were presented. After all the pairs were presented, the order of the pairs was randomized again. This randomization was conducted four times. 24 trials per spoken sentence were presented to the participants. EEG recording was conducted in a dimly lit soundproof room. The experimental sessions lasted about ten minutes.

2.2. Extraction of phase patterns from EEG data

2.2.1. Preprocessing of EEG data

We used the EEGLAB toolbox [14] for preprocessing the recorded raw EEG data. Artifacts were removed from the data using a zero-phase FIR high-pass filter (passband edge frequency: 1Hz, -6-dB cutoff frequency: 0.5 Hz) and a zero-phase FIR low-pass filter (passband edge frequency: 45 Hz, -6-dB cutoff frequency: 50.6 Hz). The EEG data were extracted from -500 to 4,000 ms relative to the speech onset and decomposed using independent component analysis (ICA). Trials containing artifacts were estimated based on data improbability using each component [15]. If the joint log probability of a trial was over five standard deviations above the mean of the probability distribution, the trial was rejected [15]. As a result, 6.5 % of the trials were rejected from further analysis.

ICA was performed again, and independent components related to such artifacts as muscle artifacts, eye blinks, and eye movements were subtracted from the EEG data with the ADJUST plugin [16] for EEGLAB and visual inspection. Finally, the preprocessed EEG data were extracted from the $0 \sim 2{,}900$ ms time-locked to the stimulus onset.

2.2.2. Cross-trial phase coherence as phase-locking index

To identify the electrode channels that show the phase-locked responses to input the speech rhythm, we calculated the cross-trial phase coherence (Cphase) [7], [8], [9] per electrode channel. The Cphase index quantifies the coherence of the phase patterns in a frequency bin among trials. The range is from 0 to 1, where 1 represents the maximum coherence. Phase values were calculated in each window using Short-Time Fourier Transform (STFT; FFT points: 500, shift points: 100, Hanning window). Cphase is defined by the following formula:

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

Here, n, j, i, and k respectively represent each trial (24 trials per sentence), each window in STFT, a frequency bin (2-Hz interval), and the type of spoken sentence. $Cphase_{kij}$ was averaged in the temporal domain and among three sentences.

2.3. Classification method on Japanese spoken sentences using EEG phase patterns

We calculated the averaged Cphase among each frequency bin in the theta band (frequency bins: 4, 6, and 8 Hz), the alpha band (10, 12, and 14 Hz), the beta band (16, 18, and 20 Hz), and the low gamma band (38, 40, and 42 Hz). Then we selected the top three channels, except for Fp1, Fp2, O1, and O2, with the highest averaged Cphase per frequency band. Phase patterns were extracted using STFT (FFT points: 500, shift points: 100, Hanning window) from the top three channels in each frequency band. The phase patterns in each single frequency band and the combination of all the frequency bands were used as classification features.

For three-class Japanese spoken sentence classification, we constructed both subject-dependent and -independent models based on SVM and template matching. In subject-dependent classification, the models were trained using each participant's data (five frequency bands $\times$ two classifiers = ten models per participant) and evaluated by leave-one-out cross validation. The accuracy in each model (frequency bands and classifiers) was obtained by summing up the numbers of correct classifications obtained in all the participant models and dividing them by the total trials used for the analysis. In subject-independent classification, the models were evaluated by leave-one-subject-out cross validation. As with the subject-dependent models, we obtained the accuracies by summing up the total numbers of correct classifications in each fold and dividing them by the total trials used for the analysis.

For template matching classification, templates were created by averaging the data in the training set, and the template with the minimum squares to a test sample was regarded as the classification result. For SVM, a linear kernel was used, and in order to construct the optimized SVM for this task, the cost parameter was tuned using a grid search.

3. Results

3.1. Cphase in each frequency band

To confirm the phase-locked responses in each frequency band, we compared the averaged Cphase values among the three highest channels. One participant was removed from further analysis because her Cphase values exceeded 2.5 SD from each average in the alpha, beta, and low gamma frequency bands. As shown in Fig. 2, the theta phase patterns showed the highest Cphase patterns among all of the frequency bands.

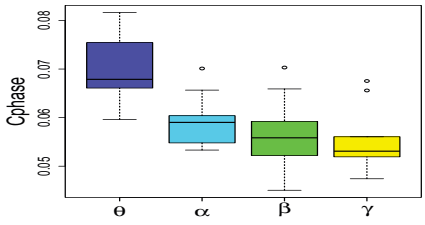

Figure 2: *Averaged Cphase values among highest three channels in each frequency band*

One-way repeated measures ANOVA revealed a main effect of frequency bands ($F(3, 24)=10.33$, $p<0.01$). Pairwise comparisons using paired t-tests with Holm's p-value adjustment revealed that Cphase in theta was significantly higher than alpha, beta, and low gamma ($p<0.05$). It is indicated that this higher Cphase in theta is derived by phase-locked responses to Japanese speech envelope [7], [8], [9].

We plotted the topographic maps of Cphase distribution (Fig. 3). Coinciding with the AST hypothesis [3], a clear right-lateralized Cphase distribution was observed in the theta frequency band. As for the low gamma, which is another window for processing the rapid temporal changes in the AST hypothesis, the left-lateralization was not clear, coinciding with a previous result [7]. The Cphase analysis identified the phase-locked responses in the theta band.

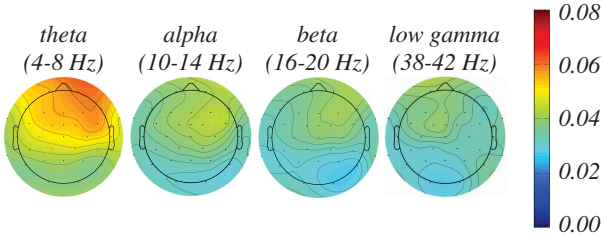

Figure 3: *Topographic maps of averaged Cphase among participants in each frequency band*

3.2. Classification performances

3.2.1. Subject-dependent models

Figure 4 shows the accuracies of SVM and template matching in each frequency band. The best performance achieved 55.2 % accuracy from the SVM on the combination of all frequency bands. This performance showed 8.6 % improvement compared to template matching on the theta phase patterns (46.6 %). A Pearson's Chi-squared test revealed that this improvement was significant ($\chi^2(1)=8.942$, $p<0.01$); however, the SVM and template matching performances based on a combination of all the frequency bands did not differ significantly (template matching: 54.2 %, $\chi^2(1)=0.120$, $p=0.73$).

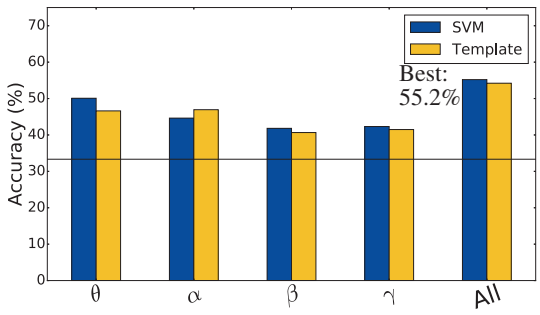

Figure 4: *SVM and template matching accuracies for subject dependent models in each frequency band and combination of all frequency bands. A horizontal line represents a chance level (33.3%).*

In the classification on a single frequency band, the performances show a similar tendency to the Cphase values in each frequency band (c.f., Fig. 2). We calculated the correlation between the Cphase values and model accuracies from each frequency band and each participant (Fig. 5). The correlation test revealed a positive correlation between the Cphase values and the classification performances: SVM: Kendall's

τ = 0.526, p<0.01, template matching: Kendall's τ = 0.473, p<0.01. This suggests that consistent phase patterns in neural oscillations across trials are related to the classification performances.

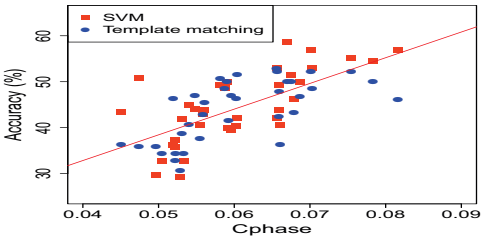

Figure 5: *Relationship between Cphase and accuracies. Red line represents regression.*

3.2.2. Subject-independent models

Figure 6 shows the SVM and template matching accuracies per frequency band. The best performance achieved 44.0 % from template matching on the combination of all the frequency bands. Compared to template matching with theta phase patterns, 5.5 % improvement was observed (template matching on theta: 38.5 %). The difference was marginally significant (Pearson's Chi-squared test: $\chi^2(1)$=3.714, p=0.054). The difference between the performances of the two classifiers on this feature was not significant (Pearson's Chi-squared test: $\chi^2(1)$=0.760, p=0.38). In the single frequency classification, the performances based on theta or alpha frequency bands were above the chance level in both classifiers (one-tailed exact binomial test: p<0.05). On the other hand, whether performances based on beta or gamma were above the chance level depended on the classifier. For the models trained on the beta phase patterns, the SVM accuracy was significantly above chance level (37.2 %, p<0.05), but template matching accuracy only showed a marginally significant difference (36.5 %, p=0.053). In the case of the gamma phase patterns, SVM failed to reach chance level (33.1 %), but the template matching accuracy significantly exceeded chance level (37.0 %, p<0.05).

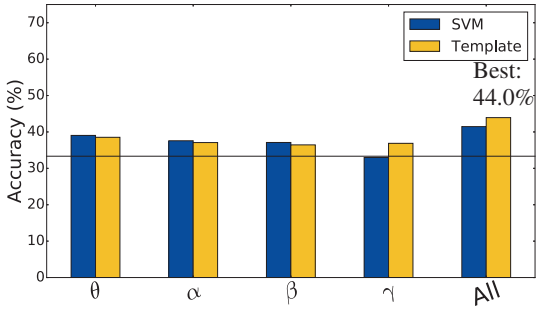

Figure 6: *SVM and template matching accuracies for subject-independent models in each frequency band and combination of all frequency bands. A horizontal line represents a chance level (33.3%).*

4. Discussion

In this study, we investigated the performances of EEG-based Japanese spoken sentence classification by two classifiers (SVM and template matching), multiple frequency bands, and a combination of all the frequency bands in subject-dependent and -independent models based on phase patterns.

In the analysis of phase-locked responses, we observed that theta neural oscillation showed phase-locking to envelopes in Japanese spoken sentences. Because phase-locked responses are driven by acoustics in speech [9], we expected such responses to speech rhythm in Japanese spoken sentences. As for the types of classifiers, we expected that SVM showed a better performances than template matching, but there were no differences between classifiers.

We found that the best performances were obtained from models based on a combination of all the frequency bands. The feature improved the accuracies compared to template matching based on theta phase patterns (subject-dependent: 8.6%, subject-independent: 5.5%). Reliable performances in subject-independent classification are one piece of evidence that phase-locked responses showed consistent patterns across other listeners [10], [11]. In subject-independent classification based on a single frequency band, the phase patterns in theta and alpha showed above chance level performances in both classifiers. Recent perception models identified the role of cross-coupled oscillations in the lower ($\leq$10 Hz) and distinct higher frequency bands [3], [4], [5], [6], and an MEG experiment showed that neural oscillation in alpha did not show phase-locking to non-speech stimuli with a temporal structure that corresponds to alpha [7]. Thus, the theta phase patterns that consistently track envelopes across listeners realize reliable classification performances, but it is unlikely that the performances of the models trained on the alpha phase patterns were derived by phase-locked responses to input speech rhythm. In the beta and low gamma frequency bands, whether classification performances were above chance depended on the classifiers. Considering the above perception models, the above chance level classification performances in the two frequency bands in either classifier might be one piece of evidence for the existence of cross-coupled neural oscillation in the theta and higher frequency bands.

5. Conclusions and future work

We demonstrated the improvement of the classification accuracy in models trained on a combination of multiple frequency bands. This improvement is based on recent neurophysiological perception models. Replicable phase patterns across trials were related to classification performances, and consistent phase-locked responses across other listeners enable subject-independent classification. In the future, for application to BCI systems, we must extend the number of sentence classes and investigate robustness to acoustical variabilities in spoken sentences (e.g., sentences spoken by multiple speakers).

In this study, we found no performance differences between classifiers. In the future, we will also investigate SVM with other kernels (e.g., RBF kernel) or such other classification algorithms as deep neural networks (DNN). Generally, a DNN needs more datasets than other algorithms for model training, but the subject independency of phase-locked responses might overcome this obstacle.

6. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Numbers 16K16172 and 17K00237, and by a joint research project with Suntory Global Innovation Center.

7. References

[1] J. E. Peelle and M. H. Davis, "Neural oscillations carry speech rhythm through to comprehension," *Frontiers in Pyschology*, vol. 3:320, 2012.

[2] 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*, vol. 94, no. 3, pp. 1904–1911, 2005.

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

[4] O. Ghitza, "Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm," *Frontiers in Psychology*, vol. 2:130, 2011.

[5] A. L. Giraud and D. Poeppel, "Cortical oscillations and speech processing: emerging computational principles and operations," *Nature Neuroscience*, vol. 15, no. 4, pp. 511–517, 2012.

[6] A. Hyafil, L. Fontolan, C. Kabdebon, B. Gutkin, and A. L. Giraud, "Speech encoding by coupled cortical theta and gamma oscillations," *eLife*, vol. 4, pp. 1–23, 2015.

[7] 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. 3:170, 2012.

[8] ——, "Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex," *Neuron*, vol. 54, no. 6, pp. 1001–1010, 2007.

[9] 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.

[10] 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.

[11] B. Zoefel and R. VanRullen, "EEG oscillations entrain their phase to high-level features of speech sound," *Neuroimage*, vol. 124, pp. 16–23, 2016.

[12] 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.

[13] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [Computer program]," Version 6.0.14, retrieved from http://www.praat.org/, 2016.

[14] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[15] A. Delorme, T. Sejnowski, and S. Makeig, "Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis," *Neuroimage*, vol. 34, no. 4, pp. 1443–1449, 2007.

[16] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features," *Psychophysiology*, vol. 48, no. 2, pp. 229–240, 2011.