Single-trial Detection of Semantic Anomalies from EEG during Listening to Spoken Sentences

Hiroki Tanaka, Hiroki Watanabe, Hayato Maki, Sakriani Sakti, and Satoshi Nakamura

Abstract— **We propose a method for the automatic detection of mismatched feelings that occur in communication. As our first step, we examined the semantically anomalous feelings from EEGs when participants listened to spoken sentences. Previous studies have shown that the event-related potentials (ERP) of an electroencephalogram (EEG) are evoked in the auditory and visual modalities where a semantic anomaly occurs. We expand this knowledge and detect it from a single-trial ERP using machine learning techniques. We recorded the brain activity of eight participants as they listened to sentences that contained semantic anomalies and found that a combination of feature selection using linear discriminant analysis and linear kernel support vector machines achieved the highest accuracy that exceeded 60%. By applying this technique, we plan to detect other types of anomalies in practical situations.**

I. INTRODUCTION

In speech communication, we often recognize semantic and syntactic errors, among other types, specifically in language learners and machine output (e.g., machine translation results). In traditional methods, evaluators directly ask questions to observe such perceived errors. However, since this approach includes problems in which the evaluations of the participants contain ambiguity and are time consuming and cost ineffective, we propose a new method that automatically detects such error (mismatched) feelings from biomedical signals.

An electroencephalogram (EEG) is a non-invasive tool that records the electrical activity of the human brain with the electrodes placed along the scalp. Among types of EEG measurements, event-related potential (ERP) is a measured brain response that is a direct result of a specific sensory, cognitive, or motor event. Since ERP generally has a low signal/noise ratio in individual trials, many consecutive trials are usually averaged to diminish the random noise.

Even though single-trial detection of ERP is challenging, it is useful for applying the realtime assessments of the cognitive states of users. Most previous works have shown that P300 components, which have relatively high signal/noise ratios, can be detected with around 50-70% accuracy using several machine learning algorithms [1], [2]. Work-detecting keyboard auto-correction errors from EEGs indicated that the accuracy of single-trial detection was around 70% [3].

A few studies have investigated the single-trial detection of semantic anomalies. For example, Geuze et al., tried single-trial detection of semantic priming, classifying visually presented related and unrelated words [4]. The semantic anomaly was measured as an ERP of N400, which is a well-known ERP component evoked in auditory and visual modalities where semantic anomalies occur [5], [6]. N400 is a phenomenon in which the potential shift in the negative direction increases around the brain's parietal region at 400 ms from the onset of the semantic anomalies. Because N400 is strongly influenced by background noise, artifacts, and variations among trials, multiple times must be averaged.

In this study, we focus on the single-trial detection of semantic anomalies while listening to spoken sentences. This method can be applied to many practical situations such as evaluating errors in automatic speech recognition (ASR), spoken dialogue (SD) systems, and machine translation (MT) as well as assessing people with autism spectrum disorders [7] or those who show anomalies of semantic context sensitivity [8].

We recorded EEG data while participants listened to sentences with semantic anomalies and analyzed the N400 effects. In addition, we detected semantic anomalies from single-trial EEGs with a technique that classified them from multi electrodes/integration of time and spectral information with machine learning as well as feature selection based on linear discriminant analysis.

II. METHODS

A. Materials

Japanese semantic anomalies were manually created by referring to Takezawa et al. [9]. We created a matched number of semantically correct and incorrect sentences. The following is an example of two types of sentences:

(semantic)

a.	Taro-ga	ryoko-ni	dekake-ta
	Taro-NOM	a journey-DAT	set out-PAST
	Taro set out on a journey.		
b.	*Taro-ga	jisho-ni	dekake-ta
	Taro-NOM	a dictionary-DAT	set out-PAST
	Taro set out on a dictionary.		

NOM: nominative case marker;
DAT: dative case marker;
PAST: past tense morpheme.

*Part of this work was supported by a joint research project with Fujitsu Laboratories, as well as JSPS KAKENHI Grant Numbers JP17H06101, JP18K11437, and SCAT.

H. Tanaka, H. Watanabe, H. Maki, S. Sakti, and S. Nakamura are with Graduate School of Information Science, Nara Institute of Science and Technology, Takayama-cho 8916-5, Ikoma-shi, Nara, Japan. hiroki-tan@is.naist.jp

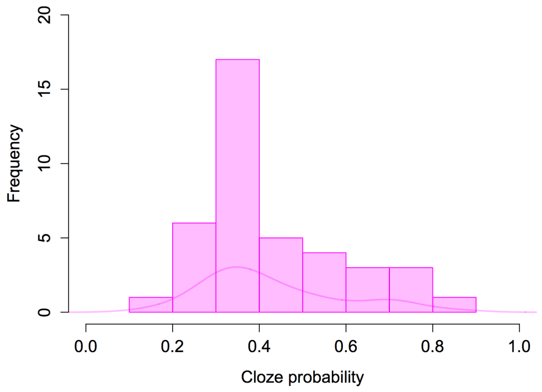

Fig. 1. Cloze probability of hidden final words from semantically incorrect sentences

Here an asterisk indicates semantically incorrect sentences. Matched sentences corresponded in the first and third phrases. The error is a selectional restriction between a verb and its arguments. Due to speech stimulus, we controlled the phonemes in the third phrase to begin with plosive sounds: e.g., /t/, /k/, /d/.

A group comprised of the first author (A), the second author (B), and a graduate student who did not join our experiment (C) confirmed and corrected each sentence and reached a consensus about whether a semantic anomaly occurred. We selected from a total of 360 sentences as follows: 40 semantically correct, 40 semantically incorrect, 40 syntactically correct, 40 syntactically incorrect[1], and 40 fillers sentences (200 sentences in total). Here, fillers were correct sentences used as dummy. We transcribed them into text and recorded the speech naturally spoken by a female professional narrator. The length of the audio file ranged from 1.8 to 3.0 s. Persons A and C marked the synchronized onset, which is the speech's start position of the third phrases.

Moreover, we investigated the predictability of subsequent words (cloze probability) that may affect amplitudes of N400. 100 crowdsourcing workers from CrowdWorks were given a list of 40 semantically incorrect sentences from which the final word had been removed. They read the sentences and filled the blanks at the position of the hidden sentence-final words with the first word that popped into their heads. After that, we manually calculated the cloze probability of the most frequently selected words. The distribution of the cloze probability is shown in Fig. 1 (mean: 0.42, SD: 0.16, range: 0.15-0.85).

B. Participants

This experiment was approved by the research ethical committee of the Nara Institute of Science and Technology. Ten graduate students (nine males and one female) from the Nara Institute of Science and Technology participated, all of whom were native Japanese speakers, right-handed, and without any history of psychiatric problems.

C. Procedure

The participants entered a soundproof room, sat on a chair, and were instructed to look at the attention point on the monitor and refrain from blinking and body movements as much as possible. The following was the experimental procedure: (1) watch the "+" mark for 1 s on the screen, (2) listen to one randomly presented speech sound within 4 s, and (3) press a key to determine whether each sound is correct Japanese within 2 s. We conducted subjective evaluations to support the attention of the participants and prepared practice trials before the EEG recordings. All these steps were completed within 25 minutes. For speech listening, insert earphones (ER1) were used.

D. EEG Recording and Preprocessing

As an EEG cap, we used ActiCAP by Brain Products with 32 ch and active electrodes and a BrainAmp DC from the same company as an amplifier. For pre-processing the recorded EEGs, we used EEGLAB [10] and ERPLAB [11] as follows: (1) the recorded EEGs were downsampled to 250 Hz; (2) re-referencing was performed on the average of the TP9 and TP10 electrodes; (3) independent component analysis was performed, and specific components were removed using ADJUST [12]; (4) a two-pass IIR Butterworth filter was applied through a high-pass filter of 0.1 Hz and a low-pass filter of 30 Hz (filter order: 2, cutoff freq. (half-amp.): -6 dB); (5) for each trial condition (excluding fillers), epoching was carried out at 200-800 ms of the synchronous onset. Regarding the baseline corrections, the time before the onset was specified; and (6) we finally applied the moving window peak-to-peak elimination method (voltage threshold: 100 μV, moving window full width: 200 ms, window step: 100 ms). Since we had to remove two participants because of a high noise ratio (more than half epochs were rejected), 24.8% of the trials were rejected from subsequent analysis. We found no effects of the number of rejected trials between semantic correct and incorrect.

E. N400 Analysis

For each of the eight participants, we plotted the signal average of the semantically correct/incorrect trials. We confirmed whether N400 was evoked and computed the grand average of all the participants. Based on a previous study [6], we calculated a statistical test of the mean amplitude values at the Pz, Cz, and Fz electrodes in 350-500 ms. Assuming the normality and the equal variance, a paired one-tailed t-test was used. The significance level is 5%.

F. Feature Sets and Classifiers

Based on previous feature sets [6], we calculated the average value of the 350-500 ms of the Pz, Cz, Fz electrodes (time domain: Pz, Cz, Fz) and extracted the average value of the amplitudes of 200-300 ms, 350-500 ms, and 500-750 ms from all of the electrodes (93 of time domain). In addition,

[1] In this study, the double nominative case was syntactically anomalies. We will investigate syntactic errors in the future.

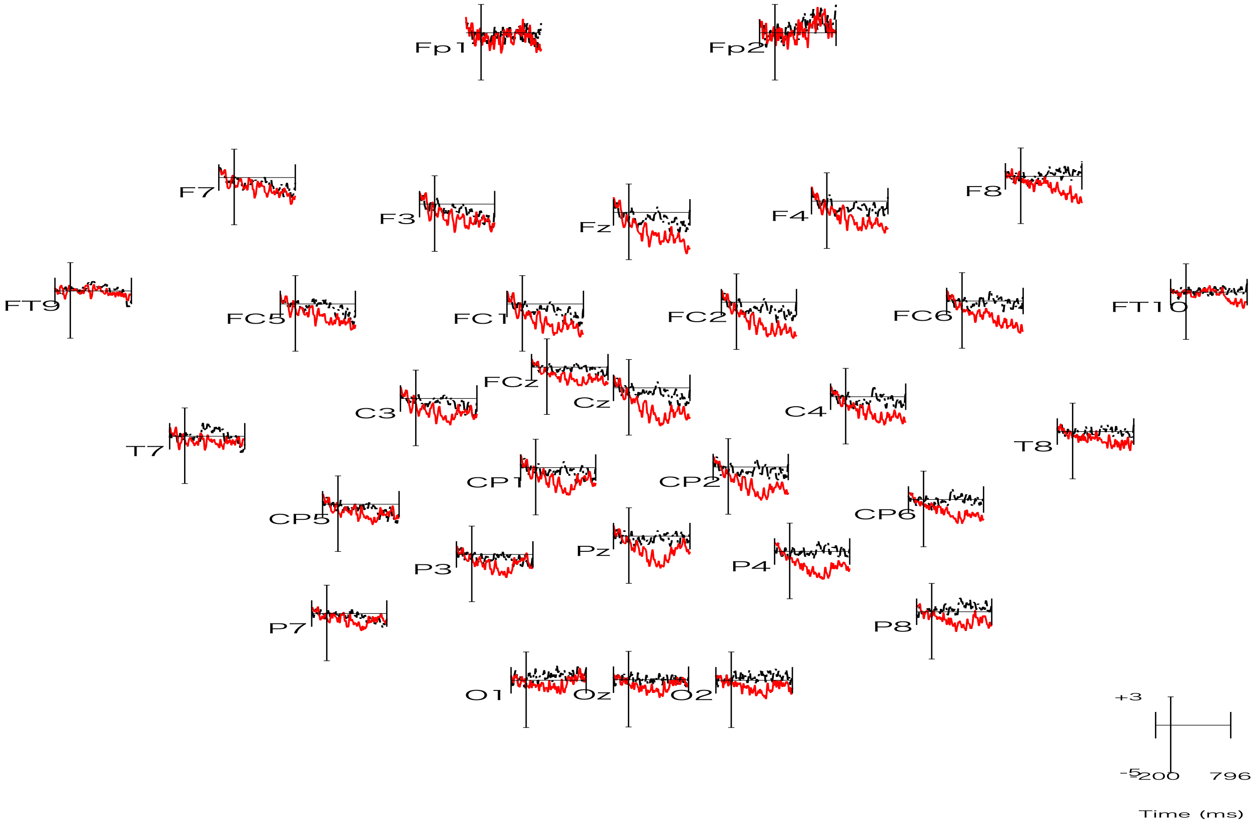

Fig. 2. Grand average of all channels from all participants

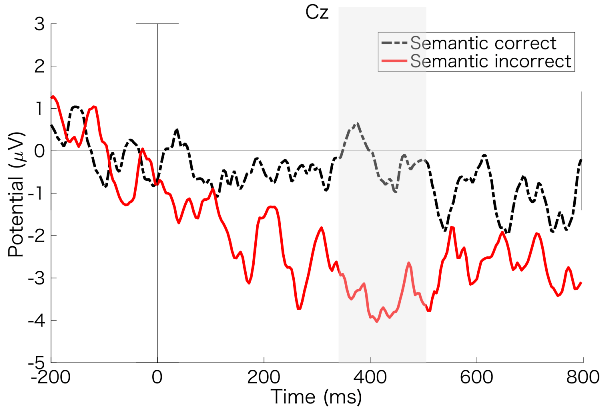

Fig. 3. Grand average of Cz channel

we performed a fast Fourier transform on the waveform after the onset and calculated the average value of the power spectrum of θ (4-7 Hz), α (8-12 Hz), and β (13-28 Hz) (93 of spectral domain).

As classifiers, we used a linear support vector machine (SVM) and a random forest (RF), which is capable of non-linear separation. We used the following parameters (SVM: cost=1.0, RF: # of variables tried at each split: 13). We performed 10-fold cross-validation to evaluate our model (we split data into training and test sets).

In addition, we calculated the weight value (LDR: linear discriminant ratio) from Fisher's linear discriminant analysis in the training set for the time and spectral features and selected the top 20% of weighted features. By a binomial test, we compared the chance rate (50.6%) and the model that achieved the highest accuracy.

III. Results

A. N400 Effects

Figure 2 plots the ground average at all the electrodes. The potential shift to the negative around 400 ms can be observed under the semantically incorrect condition over the parietal region. Fig. 3 shows a plot of the Cz electrode. A difference in the negative potential shift was confirmed around 400 ms (highlighted in gray). All three channels were significantly different: Fz (t(7)=3.07, $p = 0.008$), Pz (t(7)=3.29, $p = 0.006$), and Cz (t(7) = 3.19, $p = 0.007$).

B. Single-trial Detection

Table I indicates the accuracy of the feature sets and the classifiers. For the Pz, Cz, and Fz electrodes, the values of the

TABLE I

UNWEIGHTED ACCURACIES [%] OF FEATURE SETS AND CLASSIFIERS. BEST MODEL IS INDICATED IN BOLD.

Feature	SVM	RF
Time domain (Pz, Cz, Fz)	54.43	46.23
Time domain	56.48	54.81
Spectral domain	53.97	55.23
Time and spectral domain	56.48	57.14
Time and spectral domain (LDR: $>$ 80%)	**60.67**	59.62

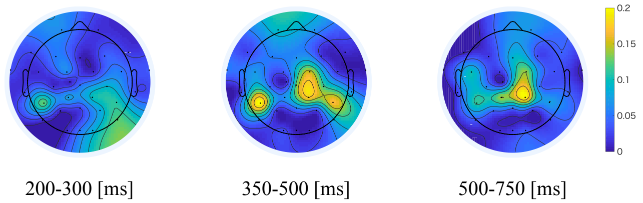

Fig. 4. Weighted time domain channel of LDR

accuracy were 54.43% (SVM) and 46.23% (RF). The time and spectral domains slightly contributed to the accuracy. The combination of the time and spectral domain improved the accuracy to 57.14% (RF). Although there were no large differences between the classifiers, feature selection based on LDR was effective, achieving the highest accuracy of 60.67% (SVM), 59.62% (RF). Regarding this accuracy, we confirmed a statistical difference compared to the chance rate ($p < 0.05$). In a case study of the first two participants, we found that a sentence with the highest cloze probability (0.85) can be correctly predicted.

The LDR weights of the electrode in each time domain are shown in Fig. 4, which represents selected features in time domain. The 350-500 ms time areas and the parietal region were highly weighted.

IV. DISCUSSION

A. N400 Analysis

In the Japanese language and semantically incorrect condition, we confirmed the N400 effects. One of this experiment's limitations is that semantically incorrect sentences were limited to anomalies of selectional restrictions at the end of sentences. Future work must consider the positions and the types of anomalies. Also, to evaluate machine output, SD and ASR errors, etc., we need to consider the effects of the subjective evaluations obtained from behavior tasks.

B. Single-trial Detection

Our classification model achieved 60% detection accuracy and outperformed the chance rate. Such accuracy resembles previous related works [2], [3] although these works tried to detect the P300 or error potentials. However, a detection accuracy of 50-60 % still would not be suitable for detection in practical situations. This result suggests that the feature selection (top 20%) in the time and spectral domain was effective, since maximum accuracy was obtained from the linear SVM with feature selection. We also observed the LDR weights in the time domain and the channel, and they might be related to the brain's parietal region that agrees with previous N400 studies [6].

V. CONCLUSION AND FUTURE WORK

We detected semantic anomalies from a single-trial EEG using a machine learning technique. We measured the EEGs of eight participants while they listened to semantically anomalous sentences and confirmed the N400 effects in the parietal region. In addition, when using feature selection and linear SVM, we achieved detection accuracy over 60%.

Future work will improve our model using [13], which we previously proposed, as well as finding best parameters. We will detect semantically anomalous feelings with a method that doesn't use an offline approach, such as independent component analysis, for a realtime communication evaluation. Automatic onset detection and technique of artificial shifted trials are also needed for completely automated anomalies detection.

REFERENCES

[1] N. Sharma, "Single-trial p300 classification using pca with lda, qda and neural networks," *arXiv preprint arXiv:1712.01977*, 2017.

[2] H. Higashi, T. M. Rutkowski, T. Tanaka, and Y. Tanaka, "Subspace-constrained multilinear discriminant analysis for erp-based brain computer interface classification," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2015 Asia-Pacific*. IEEE, 2015, pp. 934–940.

[3] F. Putze and W. Stuerzlinger, "Automatic classification of auto-correction errors in predictive text entry based on EEG and context information," *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, pp. 137–145, 2017.

[4] J. Geuze, M. A. van Gerven, J. Farquhar, and P. Desain, "Detecting semantic priming at the single-trial level," *PloS one*, vol. 8, no. 4, p. e60377, 2013.

[5] P. Hagoort and C. M. Brown, "ERP effects of listening to speech compared to reading: The P600/SPS to syntactic violations in spoken sentences and rapid serial visual presentation," *Neuropsychologia*, vol. 38, no. 11, pp. 1531–1549, 2000.

[6] ——, "ERP effects of listening to speech: Semantic ERP effects," *Neuropsychologia*, vol. 38, no. 11, pp. 1518–1530, 2000.

[7] H. Tanaka, H. Negoro, H. Iwasaka, and S. Nakamura, "Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders," *PLoS ONE*, vol. 12, no. 8, pp. 1–15, 2017.

[8] J. Pijnacker, B. Geurts, M. Van Lambalgen, J. Buitelaar, and P. Hagoort, "Exceptions and anomalies: An ERP study on context sensitivity in autism," *Neuropsychologia*, vol. 48, pp. 2940–2951, 2010.

[9] S. Takazawa, N. Takahashi, K. Nakagome, O. Kanno, H. Hagiwara, H. Nakajima, K. Itoh, and I. Koshida, "Early components of event-related potentials related to semantic and syntactic processes in the Japanese language," *Brain Topography*, vol. 14, no. 3, pp. 169–177, 2002.

[10] 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.

[11] J. Lopez-Calderon and S. J. Luck, "Erplab: an open-source toolbox for the analysis of event-related potentials," *Frontiers in human neuroscience*, vol. 8, pp. 1–14, 2014.

[12] 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.

[13] M. Hayato, T. Hiroki, S. Sakriani, and N. Satoshi, "Graph regularized tensor factorization for single-trial eeg analysis," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.