Incongruity Detection on ASR Outputs based on EEG Signals*

◎ Sakriani Sakti, Yu Odagaki, Takafumi Sasakura† , Graham Neubig,
Tomoki Toda‡ , Satoshi Nakamura (NAIST)

1 Introduction

Most automatic speech recognition (ASR) systems today are still trained and tuned by minimizing word error rate (WER). In this framework, all words, like keywords, function words, even disfluencies and fillers, are treated uniformly, and all the errors (substitutions, deletions, insertions) of those words are considered equally deleterious. The limitations of WER for measuring accuracy have been the subject of several past studies [1, 2]. But no analysis has investigated how the human brain processes language and perceives the effect of mistaken output by ASR systems. In this study, we utilize event-related brain potential (ERP) and directly analyze impact of ASR errors on brain activity.

2 EEG Event-Related Brain Potential

Electroencephalography (EEG) is an electrophysiological measurement of the brain activity on the human scalp surface where the voltage variations of cortical field potentials are imaged [3]. ERPs are signal-averaged EEG epochs that are time-locked to the presentation of an external event. ERP usually consists of small signals whose amplitudes vary approximately only between 2-10 μV. one common solution to increase the ratio is to average the EEG signal over many presentations of the stimulus.

Specific ERP components have been reported for different aspects of language processing. At least two well-known specific languages related to ERP signatures are: (1) N400 is an enhanced centro-parietal negative-going component for semantically incongruent words in sentences, peaking at about 400 ms after the offending word [4]; and (2) P600 [5] is a late centro-parietal positivity associated with the processing of syntactic anomalies, peaking at about 600 ms after the offending word.

However, most of these ERP studies were based on carefully-designed violation problems, specifically only on word substitutions. In this present study, we examine the ERP components on the impact of real existing speech recognition errors due to substitutions, deletions, and insertions.

3 Experimental Set-Up

3.1 ASR Developments

We used our Japanese ASR system that was built for speech-to-speech translation in the medical domain [6]. For the ASR evaluation, 200 sentences of medical dialog conversation test sets were selected and recorded. There were 27 speakers uttered 100 sentences resulting in 27,000 utterances.

3.2 ERP Materials and Subjects

We used a 32-channel EEG and focused on the central midline electrode site (Cz). A band-pass digital filter between 0.01-30 Hz was applied to remove unwanted frequency components.

Ten native Japanese-speaking subjects (23-24 years old on average) participated in the experiment. All were right-handed and had normal or corrected to normal vision. None had any history of psychiatric or neurological illness, alcohol abuse, or any history of visual deficits. All were fitted with a 32-channel cap containing tin electrodes. They sat in a comfortable chair approximately 1.5 m in front of a 50-inch TV screen from which the stimuli sentences were presented visually.

From the given ASR results, 50 sentences for each stimulus (correct, substitution, deletion, and insertion) were selected. 100 correct filler sentences were also included to equate the numbers of correct and incorrect sentences. The intervals of the word segments were 500 ms, and the pause intervals between

*EEG を用いた ASR 出力に違和感の検出, Sakti Sakriani, 小田垣佑, 笹倉隆史, Neubig Graham, 戸田智基, 中村哲 (NAIST)
†Now at Dai Nippon Printing Co., Ltd.
‡Now at Nagoya University

the sentences were also 500 ms. On average, the experiment lasted about two hours.

4 Experiment Results

Fig. 1 shows overlays of the ERP waveforms at the central midline electrode site (Cz) for the word in the correct condition and the three violation conditions (substitutions, deletions, and insertions), and Fig.2 shows the subjective evaluation results.

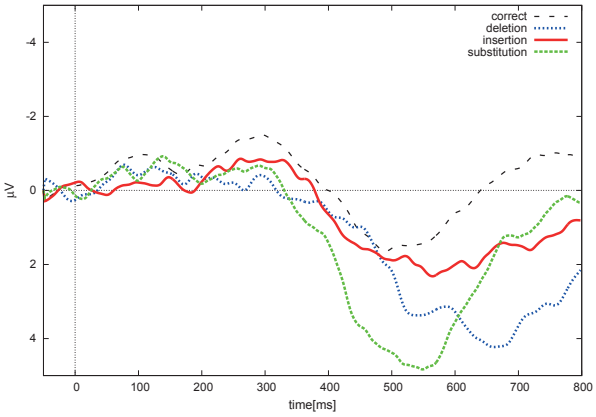

Fig. 1: ERP waveforms for correct and three violation conditions: substitutions, deletions, and insertions.

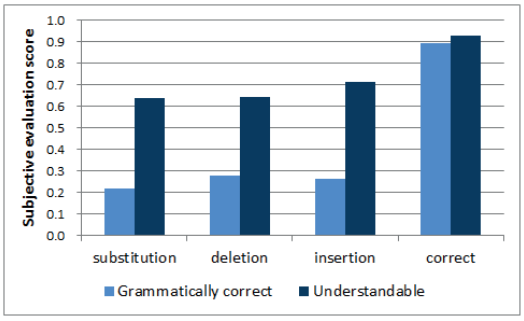

Fig. 2: Subjective evaluations during ERP experiments.

Here, the critical word is presented at time 0, and the ERP results reveal that a positive shift (P600 of the ERP component) appeared at roughly 600 ms after the error words were presented. Moreover, the peak amplitudes of the positive shift after the substitution and deletion violations were much bigger than the insertion violation. This finding indicates that the participants perceived that the substitution and deletion errors had a bigger impact than the insertion errors. From Fig.2, we also confirmed that when the substitution, deletion, and insertion violations were usually presented, the subjects realized that the sentences contained grammatical errors. However, with the insertion errors, the sentences were more understandable than those that contained substitution and deletion errors. The results reveal that the more cognitive effort that is involved to understand the ongoing word contexts the larger is the ERP peak amplitude elicited by that critical word.

5 Conclusion

In this research we utilized event-related brain potential (ERP) studies and directly analyzed the brain activities on the impact of ASR errors. Our results reveal that the peak amplitudes of the positive shift after the substitution and deletion violations are much bigger than those after the insertion violations. This finding indicates that humans perceived each error differently based on its impact of the whole sentence. In the future, we will investigate the utilization of ERP results into ASR systems. One way is to formulate a weighted word error rate to closely reflect human language perception, in which the weights of substitution, deletion, and insertion errors are define based on base-to-peak ERP amplitudes. Further details of our study can be found in [7].

Acknowledgements Part of this work was supported by the Commissioned Research of National Institute of Information and Communications Technology (NICT) Japan, Microsoft CORE 10 Project, and JSPS KAKENHI Grant Number 26870371.

References

[1] A. Morris et al., "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition," in *Proc. of ICSLP*, 2004.

[2] H. Nanjo et al., "A new asr evaluation measure and minimum bayes-risk decoding for open-domain speech understanding," in *Proc. of ICASSP*, 2005.

[3] S. Regel, *The comprehension of figurative language: Electrophysiological evidence on the processing of irony*, MPI for Human Cognitive and Brain Sciences, 2009.

[4] M. Kutas et al., "Event-related brain potentials to semantically inappropriate and surprisingly large words," *Biological Psychology*, vol. 11, 1980.

[5] L. Osterhout, "On the brain response to syntactic anomalies: Manipulations of word position and word class reveal individual differences," *Brain and Language*, vol. 59, 1997.

[6] S. Sakti et al., "Towards multilingual conversations in the medical domain: Development of multilingual medical data and a network-based ASR system," in *Proc. LREC*, 2014.

[7] S. Sakti et al., "An event-related brain potential study on the impact of speech recognition errors," in *Proc. APSIPA ASC*, 2014.