An Event-Related Brain Potential Study on the Impact of Speech Recognition Errors

Sakriani Sakti, Yu Odagaki, Takafumi Sasakura, Graham Neubig, Tomoki Toda, and Satoshi Nakamura
Graduate School of Information Science, Nara Institute of Science and Technology, Japan
E-mail: {ssakti,yu-o,sasakura.takafumi.sl4,neubig,tomoki,s-nakamura}@is.naist.jp
Tel: +81-743-72-5264

***Abstract*—Most automatic speech recognition (ASR) systems, which aim for perfect transcription of utterances, are trained and tuned by minimizing the word error rate (WER). In this framework, even though the impact of all errors is not the same, all errors (substitutions, deletions, insertions) from any words are treated in a uniform manner. The size of the impact and exactly what the differences are remain unknown. Several studies have proposed possible alternatives to the WER metric. But no analysis has investigated how the human brain processes language and perceives the effect of mistaken output by ASR systems. In this research we utilize event-related brain potential (ERP) studies and directly analyze 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 the insertion violations. This finding indicates that humans perceived each error differently based on its impact of the whole sentence. To investigate the effect of this study, we formulated a new weighted word error rate metric based on the ERP results: ERP-WWER. We re-evaluated the ASR performance using the new ERP-WWER metric and compared and discussed the results with the standard WER.**

I. Introduction

Researchers have been working in speech recognition technology for many decades. ASR approaches have progressed from a simple machine that responds to a small set of sounds to a more sophisticated system in mobile phones that responds to real spoken language. The task of state-of-the-art statistical ASR systems has shifted from recognizing a well-formed speech into purely spontaneous speech. Consequently, extracting the speaker's underlying message is more crucial than just transcribing all of the spoken words. However, most ASR systems today are still trained and tuned by minimizing WER and evaluating the measurement of recognition accuracy that calculates the minimum string edit distance (Levenshtein distance) between the correct transcription and the recognition hypothesis. In this framework, all words, like keywords, functional words, even disfluencies and fillers, are treated uniformly, and all the errors (substitutions, deletions, insertions) of those words are considered equally deleterious. In fact, the impact of errors is not the same; the size of the impact and exactly what the differences are remain unknown.

The limitations of WER for measuring accuracy have been the subject of several past studies. Research has shown that the alignment between WER and spoken language understanding is not linear [1]. Possible alternatives to the WER metric have also been investigated. Morris et al. [2] took an information theoretic approach and proposed metrics that compute the proportion of the information communicated or lost by automatically transcribing spoken messages. Other research proposed a new evaluation measure and minimum Bayes-risk decoding for open-domain speech understanding [3]. A "Human Perceived Accuracy" metric was developed by the direct regression of human subjective ratings (mean opinion score ratings) of recognition performance [4]. However, none of these works reflected directly on how the human brain processes language and perceives the effect of mistakes during communication.

On the other hand, the electrophysiological measurements of event-related brain potentials hold great promise as tools for studying the cognitive processes that underlie language comprehension. ERP measures the electrical brain activities that are the direct result of a specific cognitive event. ERP can image brain activity online (i.e., immediately at the time point of stimulus processing) with high-temporal resolution in a millisecond range that reflects rapidly occurring cognitive processes and the integration of different types of linguistic information, such as phonologic, semantic, syntactic, and pragmatic data as they unfold over time [5].

In this study, we utilize ERP studies in which the brain activities on the impact of mistaken words (substitutions, deletions and insertions errors) are directly investigated. To investigate the effect of this study, we formulate a new weighted word error rate metric based on ERP results: ERP-WWER. Here, we compare the performance using our new ERP-WWER metric with a standard WER and investigate within a speech recognition framework.

II. Related Work

Over the last two decades, a growing number of studies report specific ERP components for different aspects of language processing. At least two well-known specific languages related to ERP signatures have been identified and analyzed, including N400 and P600/SPS. N400 is an enhanced centro-parietal negative-going component for semantically incongruent words in sentences, peaking at about 400 ms after the offending word [6], and P600 [7] is a late centro-parietal positivity associated with the processing of syntactic anomalies or grammatical violations, peaking at about 600 ms after the offending word. These peak amplitudes of N400 or P600 are inversely affected by the semantic or syntactic compatibility of a given word

978-616-361-823-8

and its context. The more cognitive effort that is involved in integrating a word into an ongoing context the larger is the peak amplitude elicited by that word.

Several ERP studies across a variety of languages have obtained similar sentence-semantic N400 and sentence-syntactic P600 effects with written, spoken, and signed sentence materials. The pioneering work of Kutas and her colleagues [8] has shown that the N400 component is larger for sentence final words that are anomalous in English. Another study [9] reported syntax and semantic-related ERP effects in Dutch. Other than Indo-European languages, other work [10], [11] found that N400 and P600 were elicited in Japanese and Chinese, respectively.

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.

III. Stimuli Procedure and Quantification of ERP Components

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 [5]. It records the electrical signals generated by the brain through electrodes placed on different points on the scalp and compares the voltage between two or more different sites. We used a 32-channel EEG and focused on the central midline electrode site (Cz).

ERPs are signal-averaged EEG epochs that are time-locked to the presentation of an external event. They evoked potentials that contain negative ("N") and positive ("P") voltage deflections. However, the potential recorded at the scalp can be influenced by the sources of the electrical activity that do not arise from the brain (i.e., eye movements and tension muscles in the head). Here, a band-pass digital filter between 0.01-30 Hz was applied to remove unwanted frequency components. Furthermore, ERP usually consists of small signals whose amplitudes vary approximately only between 2-10 μV [12]. Due to the background noise of electrical activity, the signal-to-noise ratio in a single EEG trial is very low. One common solution to increase the ratio is to average the EEG signal over many presentations of the stimulus [13].

In this study, we presented four kinds of sentences as visual stimuli: (1) completely correct sentences; (2) sentences with substitution errors; (3) sentences with deletion errors; and (4) sentences with insertion errors. Examples of substitution, deletion, and insertion errors are shown in Fig. 1. On each stimulus (correct, substitution, deletion, and insertion), several trials were performed. Those sentences were presented visually, one word phrase segment at a time. A standard approach would present each segment at a presentation time of 500 ms (Fig. 2). The intervals of the segments were 500 ms, and the pause intervals between the sentences were also 500 ms.

The results of the ERP components can be defined in terms of peaks with characteristic polarities and latency ranges.

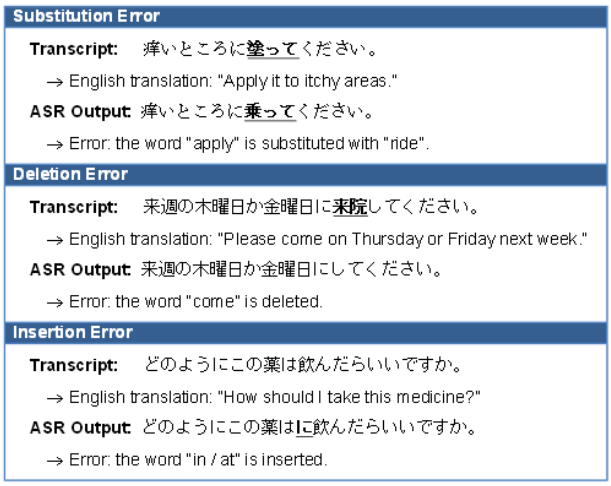

Fig. 1. Examples of sentences with mistaken words including substitution, deletion, and insertion errors.

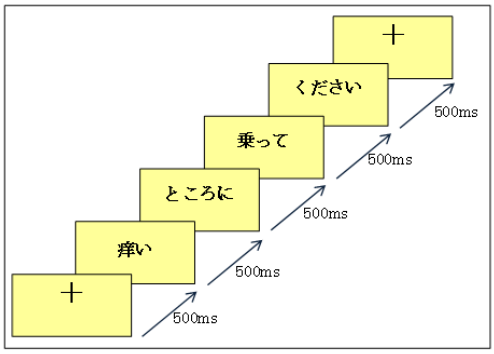

Fig. 2. Presentation of visual stimuli with substitution error (described in Fig. 1).

Thus, an ERP measurement involves the assessment of peak amplitude (in μV) and/or peak latency (in milliseconds). In this preliminary study, only the base-to-peak amplitude, which defines the distance between the peaks of the correct and violation conditions, is taken into account to formulate our new evaluation metric.

IV. Formulation of the New ERP-WWER Metric

The standard WER metric is calculated as

$$\texttt{WER} = \frac{S + D + I}{N} * 100, \tag{1}$$

where S is the number of substitution errors, D is the number of deletion errors, I is the number of insertion errors, and N is the number of words in the correct transcription. Based on ERP studies, we formulated the weighted word error rate, in which all the substitution, deletion, and insertion errors have different weights to represent their impact on human language processing. ERP-WWER is defined as follows:

$$\texttt{ERP-WWER} = \frac{\alpha S + \beta D + \gamma I}{N} * 100, \tag{2}$$

where α, β, and γ are the weights of the substitution, deletion, and insertion errors, respectively. In this preliminary study, we

simply apply the difference in the base-to-peak amplitudes of the ERP results to define the weight parameters. If all of the substitution, deletion, and insertion weights are set to 1, then ERP-WWER is equivalent to the standard WER.

V. Experimental Set-Up

A. ASR Developments

We used our Japanese ASR system that was built for speech-to-speech translation in the medical domain. It was trained based on the Kaldi toolkit [14]. A spontaneous Japanese speech corpus (CSJ) [15] with 518 hours of speech was used for the acoustic model training. For the language model training, we used medical phrasebooks and conversations, TED[1] Talk transcripts, and ATR BTEC [16] text data for a total of 519k sentences. For the ASR evaluation, 200 sentences of medical dialog conversation development and test sets were selected and recorded. There were 27 speakers with a balance of genders and ages in which each speaker uttered either 100 sentences from the development or test sets, resulting in 27,000 utterances.

We trained the systems with a front-end based on widely used mel-frequency cepstral coefficients (MFCC). The front-ends provided features every 10 ms with a 25-ms width. To incorporate the temporal structures and dependencies, nine adjacent frames of MFCCs were stacked into one single feature vector and projected to an optimum 40 dimensions by applying linear discriminant analysis (LDA). After that, the resulting features were further de-correlated using maximum likelihood linear transformation (MLLT), and speaker adaptive training (SAT) was also performed. All the models were context-dependent cross-word triphones with a standard three-state, left-to-right HMM topology without skip states.

The ASR development details can be found in a previous work [17].

B. ERP Materials and Subjects

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. 300 sentences were presented in each experiment.

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 300 sentences were presented visually, one word phrase segment at a time (Fig.2). On average, the experiment lasted about two hours.

[1]http://www.ted.com/talks

VI. Experiment Results

To produce various different mistaken words, we used three types of ASRs based on different acoustic features: (a) MFCC-DELTA, (b) LDA-MLLT, and (c) LDA-MLLT-SAT. Table I (5th column) shows the WER of our Japanese ASR systems on the medical development and test sets (the relative word error rate reduction is also included). The ASR with LDA-MLLT features outperformed the standard MFCC with delta features. Optimum performance was achieved using LDA-MLLT-SAT features with WERs of 16.79% and 15.87% on the medical development and test sets, respectively.

TABLE I
Comparison performance of our Japanese ASR systems based on WER and the new ERP-WWER metric.

Development Set (N=6163)					
Features	S	D	I	WER	ERP-WWER
MFCC-DELTA	844	291	105	20.12%	26.72%
LDA-MLLT	813	275	80	18.95%	25.53%
				(-5.81%)	(-4.45%)
LDA-MLLT-SAT	749	198	88	16.79%	22.51%
				(-11.39%)	(-11.82%)
Test Set (N=5766)					
Features	S	D	I	WER	ERP-WWER
MFCC-DELTA	818	223	103	19.84%	26.46%
LDA-MLLT	730	195	100	17.78%	23.57%
				(-10.40%)	(-10.91%)
LDA-MLLT-SAT	680	115	120	15.87%	20.71%
				(-10.73%)	(-12.14%)

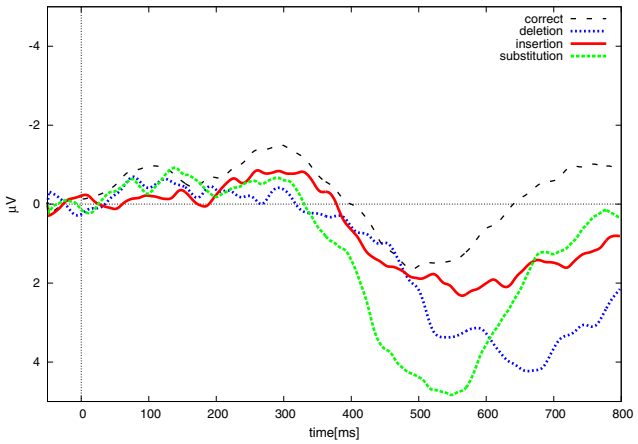

Fig. 3. ERP waveforms for correct and three violation conditions: substitutions, deletions, and insertions.

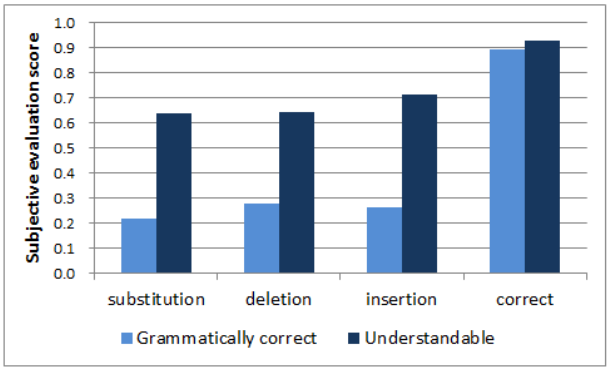

Fig. 4. Subjective evaluations during ERP experiments.

Based on the existing errors, we selected 300 sentences and randomly presented them during the ERP experiments, as described in Sections III and V-B. For each sentence, after the brain activities were recorded, we also requested the subjects to evaluate whether the sentences are grammatically correct and whether they remain understandable (1 denotes maximum score and 0 denotes minimum score). Fig. 3 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. 4 shows the subjective evaluation results. 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.4, 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.

Applying the base-to-peak amplitudes of the ERP results to the ERP-WWER metric while keeping the same total weights as in the standard WER, we defined $\alpha = 1.5$, $\beta = 1.2$, and $\gamma = 0.3$ as the substitution, deletion, and insertion weight parameters, respectively. The new ERP-WWER results of our Japanese ASR systems are also shown in Table I (last column). With the standard WER metric, the relative reductions on the test set from MFCC-DELTA to LDA-MLLT (10.40%) and LDA-MLLT to LDA-MLLT-SAT (10.73%) are almost the same. This is because the total number of reduced errors is similar. However, the reduction of the substitution and deletion errors from LDA-MLLT to LDA-MLLT-SAT is actually bigger than the reduction from MFCC-DELTA to LDA-MLLT, and therefore the impact of these reductions is different. This phenomenon can be seen more clearly with the ERP-WWER metric, where the ERP-WWER relative reduction from LDA-MLLT to LDA-MLLT-SAT is 12.14%, which is significant bigger than the relative reduction from MFCC-DELTA to LDA-MLLT (10.91%). Overall, the ASR performance with ERP-WWER is lower than the performance with WER, due to the high amount of substitution errors. Nevertheless, it closely reflects how the human brain processes and perceives the effect of mistaken words from ASR outputs.

VII. Conclusion

This paper presents ERP studies on the impact of mistaken words. ERP results reveal that a positive shift (P600 of the ERP component) appeared around 600 ms after the error words were presented. The amplitudes of the positive shift after the substitution and deletion violations were much bigger than the insertion violation. Our new ERP-WWER metric is formulated to closely reflect human language perception, in which the impact of substitution, deletion, and insertion errors is perceived differently depending on the base-to-peak amplitudes of the ERP results. In the future, we will further investigate the impact of mistaken words, especially such word functions as nouns, verbs, and particles. We must also investigate whether similar impacts appear in different languages.

Acknowledgment

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] G. Riccardi and A.-L. Gorin, "Stochastic language models for speech recognition and understanding," in *Proc. ICSLP*, Sydney, Australia, 1998.

[2] A. Morris, V. Maier, and P. Green, "From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition," in *Proc. of ICSLP*, Philadelphia, PA, USA, 2004.

[3] H. Nanjo and T. Kawahara, "A new asr evaluation measure and minimum bayes-risk decoding for open-domain speech understanding," in *Proc. of ICASSP*, Philadelphia, PA, USA, 2005, pp. 1053–1056.

[4] T. Mishra, A. Ljolje, and M. Gilbert, "Predicting human perceived accuracy of ASR systems," in *Proc. INTERSPEECH*, Florence, Italy, 2011, pp. 1945–1948.

[5] S. Regel, *The comprehension of figurative language: Electrophysiological evidence on the processing of irony*, MPI Series 111. Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, 04 2009.

[6] M. Kutas and S.A. Hillyard, "Event-related brain potentials to semantically inappropriate and surprisingly large words," *Biological Psychology*, vol. 11, pp. 99–116, 1980.

[7] L. Osterhout, "On the brain response to syntactic anomalies: Manipulations of word position and word class reveal individual differences," *Brain and Language*, vol. 59, pp. 494–522, 1997.

[8] M. Kutas and S.A. Hillyard, "Brain potentials during reading reflect word expectancy and semantic association," *Nature*, vol. 307, pp. 161–163, 1984.

[9] P. Hagoort, "Interplay between syntax and semantics during sentence comprehension: Erp effects of combining syntactic and semantic violations," *Cognitive Neuroscience*, vol. 15, pp. 883–899, 2003.

[10] S. Takazawa, N. Takahashi, K. Nakagome, O. Kanno, H. Hogiwara, 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.

[11] Z. Ye, Y.-J. Luo, A.D. Friederici, and X. Zhou, "Semantic and syntactic processing in chinese sentence comprehension: Evidence from event-related potentials," *Brain Research*, vol. 1071, pp. 186–196, 2006.

[12] M. Kutas and A.M. Dale, *Electrical and magnetic readings of mental functions*, Cognitive Neuroscience. MIT Press, Cambridge, MA, 1997.

[13] M. Fabiani, G. Gratton, and K.D. Federmeier, *Psychophysiology*, chapter Event-Related Brain Potentials: Methods, Theory, and Applications, pp. 85–119, Cambridge University Press, 3rd edition, 2007.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Moticek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, Hawaii, USA, 2011.

[15] K. Maekawa, H. Koiso, S. Furui, and H. Isahara, "Spontaneous speech corpus of Japanese," in *Proc. LREC*, 2000, pp. 947–952.

[16] G. Kikui, S. Yamamoto, T. Takezawa, and E. Sumita, "Comparative study on corpora for speech translation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1674–1682, 2006.

[17] S. Sakti, K. Kubo, S. Matsumiya, G. Neubig, T. Toda, S. Nakamura, F. Adachi, and R. Isotani, "Towards multilingual conversations in the medical domain: Development of multilingual medical data and a network-based asr system," in *Proc. of LREC*, Reykjavik, Iceland, 2014, pp. 2639–2643.