EEG Analysis Towards Evaluating Synthesized Speech Quality

Ivan Halim Parmonangan[1], Hiroki Tanaka[1], Sakriani Sakti[12], Shinnosuke Takamichi[3], Satoshi Nakamura[12]

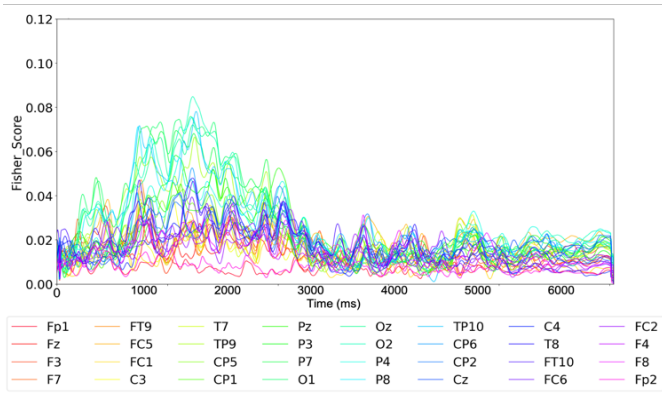
***Abstract*—This study proposes using electroencephalogram (EEG) to evaluate the perceived synthesized speech quality. Contrary to the N400 and P600 Event Related Potential (ERP) method which are popular in finding semantic and syntactic anomalies, it is not well known how fast human can detect differences in speech quality. Using generalized Fisher score, this study shows that there are certain channels even before one second in alpha band that show different activities during listening to natural and analysis-synthesized speech. This indicates that human can actually detect the difference in speech quality almost instantly.**

I. INTRODUCTION

There are two approaches to measure the synthesized speech quality; subjective and objective. Subjective measurement usually involves calculating opinion scores (e.g., mean opinion score (MOS) and preference tests). However, it can only provide an overall impression without any further detailed information about the speech. Objective measurement calculates the errors of synthesized speech features, i.e., Mel-cepstral Distance (MCD) [1]. However, the exact relationship between acoustic features and perceived quality is yet to be understood [2]. Therefore, even though the distortion is minimal, the overall naturalness might not meet human expectations. Furthermore, examining the speech quality usually use the entire length of the speech which is time-consuming. An existing study by [3] used full length of the EEG record to predict speech quality. Furthermore, their study did not investigate when human realized the difference. This study used single-trial EEG because we need to find out how quick human can recognize the unnaturalness in synthesized speech.

II. METHODS

We collected MOS of natural, analysis-synthesis, synthesized LF0, synthesized MCC, and synthesized LF0 and MCC. We asked 10 participants to listen to 53 sentences for each mentioned speech types. The EEG was recorded with 32 electrodes while the participants were listening to the sentences. The participants were instructed to evaluate the naturalness of the listened speech. The recorded EEG transformed into the time-frequency domain using Morlet wavelet and split into theta, alpha, beta, and gamma frequency bands. Finally, we used the generalized Fisher score [4] in order to analyze EEG activity from each electrode, frequency band, and time.



III. RESULTS

The collected MOS was (4.8) for natural speech, (4.3) for the analysis-synthesis speech, (2.6) for both synthesized LF0 and synthesized MCC, and (1.8) for synthesized LF0 and MCC. In this study, we firstly focus on finding the difference between natural and analysis-synthesis type because their MOS has the closest gap. In Fig. 1, Fisher score of P7 and O1 of alpha frequency band show that there are differences between 600ms and 700ms between natural and analysis-synthesis type.

IV. CONCLUSION

Based on the Fisher score result, the left parietal and occipital regions show distinct activities prior to the first second after the speech stimuli indicating that human can detect the difference in speech quality almost instantly. In the future, we will also assess the quality of other synthesized speech with different synthesized speech features.

REFERENCES

[1] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proc. of IEEE PacRim,*1993.

[2] C. Mayo, R.A. Clark, and S. King, "Listeners weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis," *Speech Communication*, 2011.

[3] H. Maki, S. Sakti, H. Tanaka, and S. Nakamura, "Quality prediction of synthesized speech based on tensor structured EEG signals," *PLOS ONE*, 2018.

[4] Q. Gu, Z. Li, and J. Han, "Generalized Fisher score for feature selection," in *Proc. of UAI,* USA, 2011.

*Part of this research is supported by JSPS KAKEN number JP17H06101, JP17K00237, and JP16K16172, and MIC/SCOPE #152307004.

[1]Ivan Halim Parmonangan, Hiroki Tanaka are with Division of Information Science, Nara Institute of Science and Technology, Japan; e-mail: ivan.halim_parmonangan.ia4@is.naist.jp.

[2]Sakriani Sakti and Satoshi Nakamura are with Center of Advanced Intelligence Project, RIKEN, Japan

[3]Shinnosuke Takamichi is with Graduate School of Information Science and Technology, The University of Tokyo, Japan