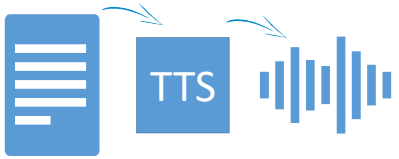
EEG Analysis Towards Evaluating Synthesized Speech Quality

Ivan Halim Parmonangan[1], Hiroki Tanaka[1], Sakriani Sakti[1,2], Shinnosuke Takamichi[3], Satoshi Nakamura[1,2]

[1]Division of Information Science, Nara Institute of Science and Technology, Japan; e-mail: ivan.halim_parmonangan.ia4@is.naist.jp.
[2]Center of Advanced Intelligence Project, RIKEN, Japan
[3]Graduate School of Information Science and Technology, The University of Tokyo, Japan

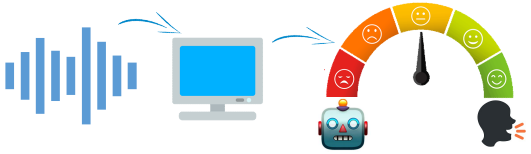
Background

- Text to speech system generates voice from text input automatically.



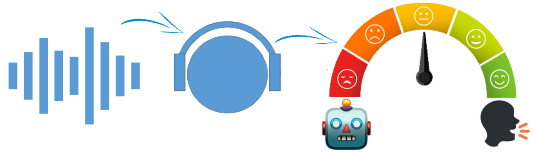
However, **sometimes** it sounds **unnatural.**
↳Therefore, it **needs evaluation**.

- One approach is to **use computer** program and **evaluate** the synthesized speech **acoustic features** objectively.



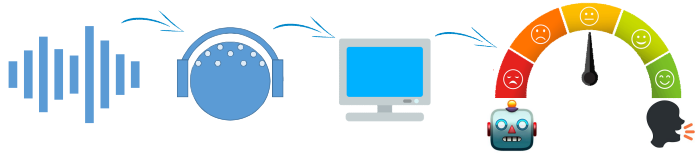
Despite of the **speed**, the **relation** between **acoustic features** and **human perception** is **yet to be understood[2].**

- Another way to evaluate the quality **subjectively** is to **collect Mean Opinion Score** (MOS) from subjects.



However, not only it is **time consuming**, but also only provides overall impression[1].

- Previously, a physiological approach was conducted **using EEG** to **predict** speech quality[3].
  - This approach used full length of English speech and all electrodes combined.



However, all of the aforementioned subjective evaluation approaches use **full length** of the speech.

Research Question

1. Are there **any electrodes** that show **difference in activity** when listening to different types of synthesized speech quality?
2. **How fast** human brain can **notice** the **difference** between synthesized speech quality?

Method

Using almost the same scenario, we tried to find out if the difference in speech quality can be seen at the beginning of the stimuli in certain electrodes.

53 Japanese sentences are prepared. Each sentence has original record and 4 synthesized types.

Record EEG while listening the speech

The recorded EEG then preprocessed and analyzed using generalized Fisher score[4].

Listen to the speech and give score 1~5

The collected opinion score from 10 people then averaged and used as the baseline.

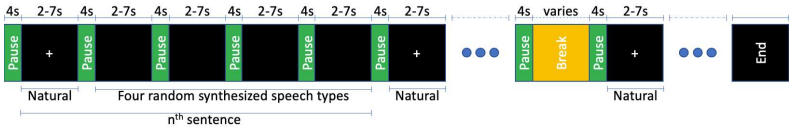


Figure of stimuli presentation during EEG recording

In this research, we want to see how human brain reacts towards different synthesized speech quality given that they already know how the original sounds like.

Fisher score of j^{th} feature can be calculated as:

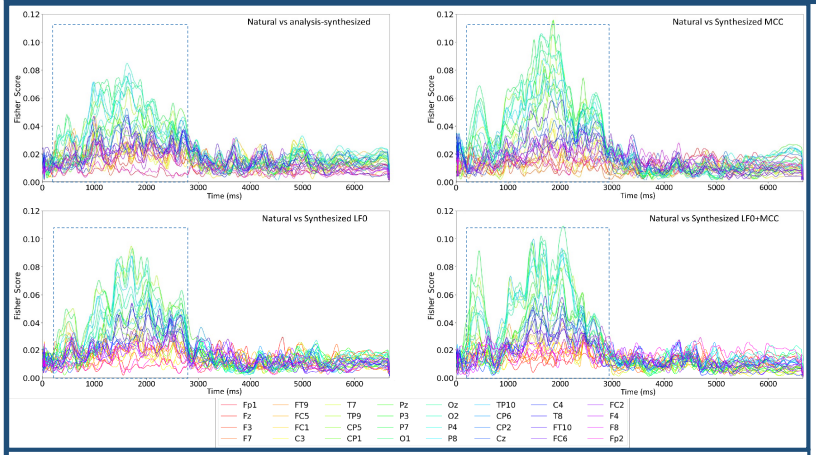
$$F(x^j) = \frac{\sum_{k=1}^{c} n_k (\mu_k^j - \mu^j)^2}{(\sigma^j)^2}$$

Where $(\sigma^j)^2 = \sum_{k=1}^{c} n_k \, (\sigma_k^j)^2$.

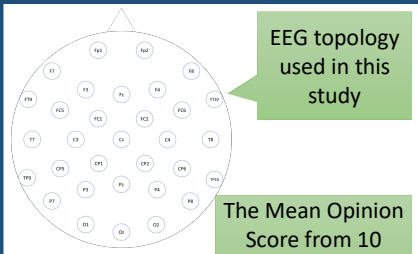
Let σ_k^j and μ_k^j be the standard deviation and mean of k^{th} class.

Let μ^j and σ^j be the mean and standard deviation of the whole dataset corresponding to the j^{th} feature.

Results

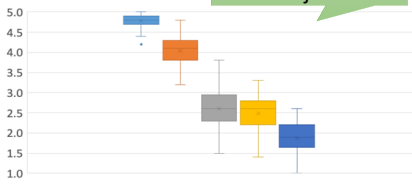


Figures of Fisher score result between natural and each synthesized speech type in alpha band.



EEG topology used in this study

The Mean Opinion Score from 10 subjects



Conclusion

- **Left parietal and occipital electrodes** show small but different neural activities between 300ms and 700ms.
- However, the largest difference happens right after 1000ms until 3000ms.

Discussion

- The result shows how fast human brain can distinguish the difference between natural and each synthesized speech in short time.
- It **needs further evaluation** whether human can notice the quality difference between all speech types with **total random order**.

[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.