Speech Quality Evaluation of Synthesized Japanese Speech using EEG

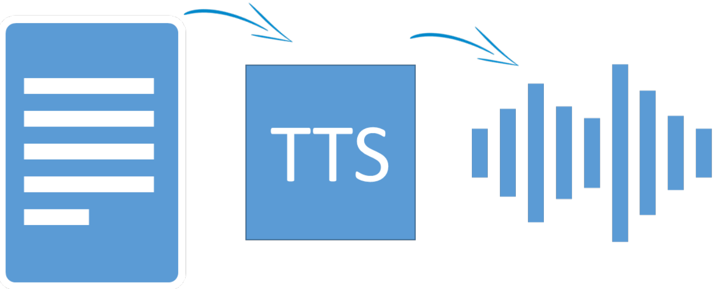
Ivan Halim Parmonangan[1], Hiroki Tanaka[1,2], Sakriani Sakti[1,2], Shinnosuke Takamichi[3],
Satoshi Nakamura[1,2]

[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

Nara Institute of Science and Technology NAIST®

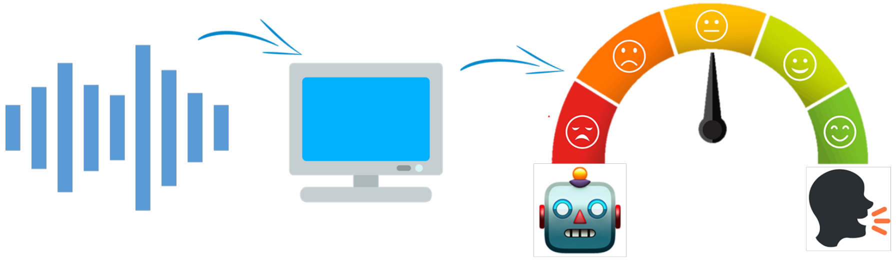
Objective Evaluation

- Synthesized speech overview.



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

- One way to evaluate is analyzing the speech features.



Despite of the speed, the **relation** between **acoustic features** and **human perception** is **yet to be understood** [Mayo et al., 2011]

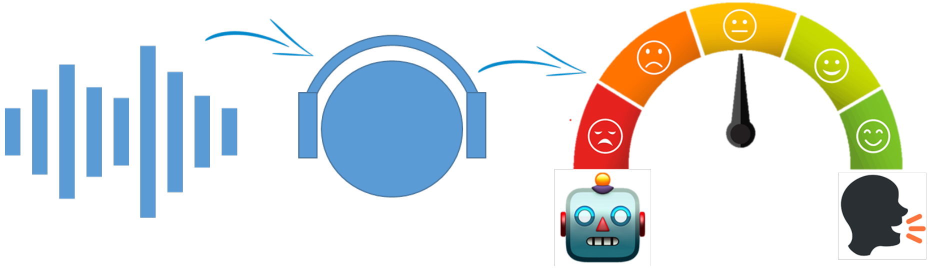
Therefore,

Even though the attained score using objective evaluation is high, it doesn’t always mean that it is also good in human perspective.

C. Mayo, R. A. Clark, and S. King, “Listeners weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis,” Speech Communication, 2011

Human perception

- Another way is to evaluate it subjectively. Meaning that we collect opinions from a group of people e.g., Mean Opinion Score (MOS).



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

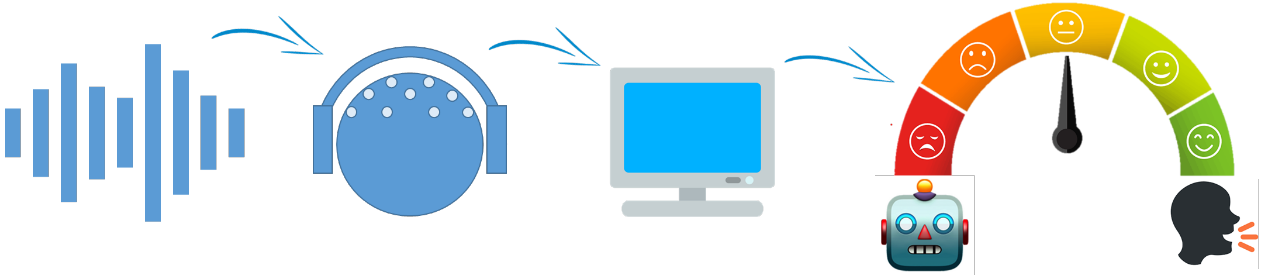
- Physiological approach reads the physical reaction when listening to different synthesized speech quality.

[Gupta et al., 2016]

- Proposed brain computer interface-based equation to predict quality of experience MOS, and achieved 1.00 of root mean squared error (RMSE) between actual and predicted MOS

[Tang et al., 2017]

- EEG could reveal the correlation between pitch emphasis and brain activity



C. Tang, L. S. Hamilton, and E. F. Chang, "Intonational speech prosody encoding in the human auditory cortex," Science, 2017

R. Gupta, K. Laghari, H. Banville, and T. H. Falk, "Using affective brain-computer interfaces to characterize human influential factors for speech quality-of-experience perception modelling, "Human-centric Computing and Information Sciences, 2016

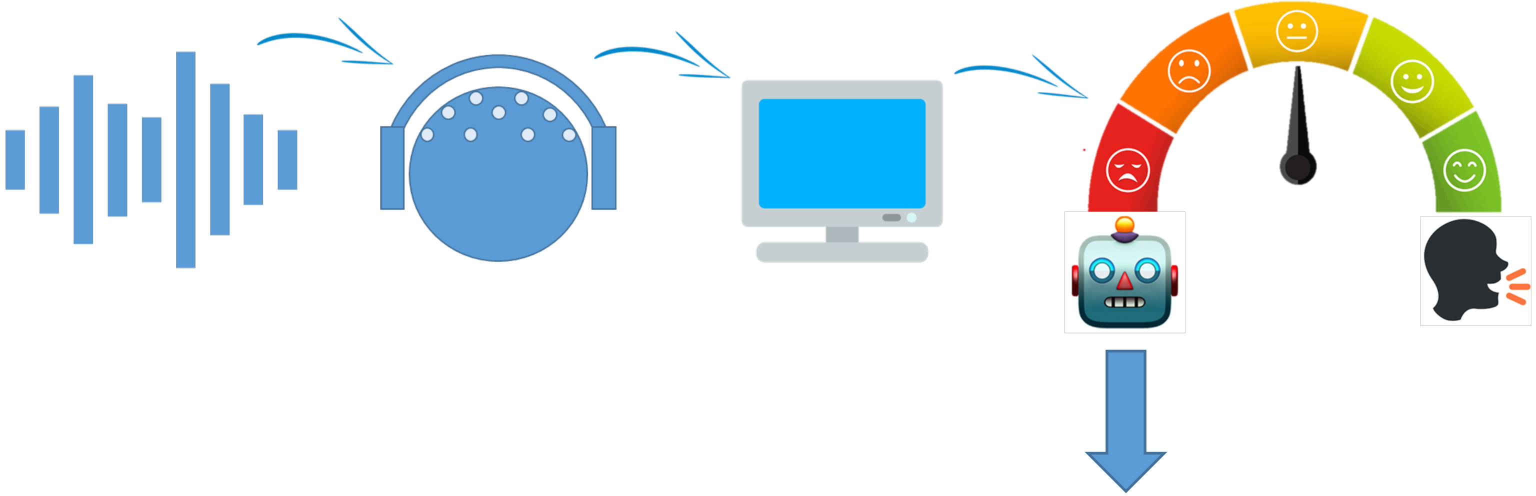
What This Study is Up To

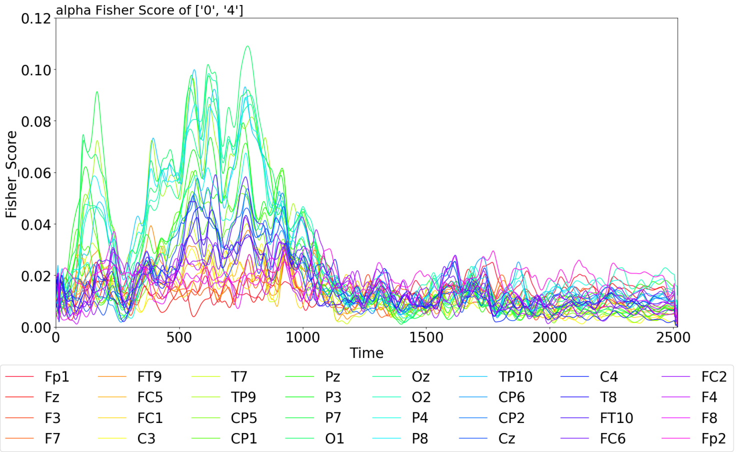
[Maki et al., 2018]

- EEG signals could be used to predict MOS, valence, and arousal using **tensor representation** of **all channels and frequency bands**, within the same subject.

[Parmonangan et al., 2019]

- Using generalized fisher score there are differences in certain channels at certain time period from each EEG frequency bands.





Can we use EEG to classify or predict MOS between synthesized type using only some of the frequency bands and simpler algorithm?

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

I. H. Parmonangan, H. Tanaka, S. Sakti, S. Takamichi, and S. Nakamura, "Subject response and eeg reaction analysis towards evaluating synthesized speech qualities," International Engineering in Medicine and Bioscience Conference, 2019.

Types of Speech Used

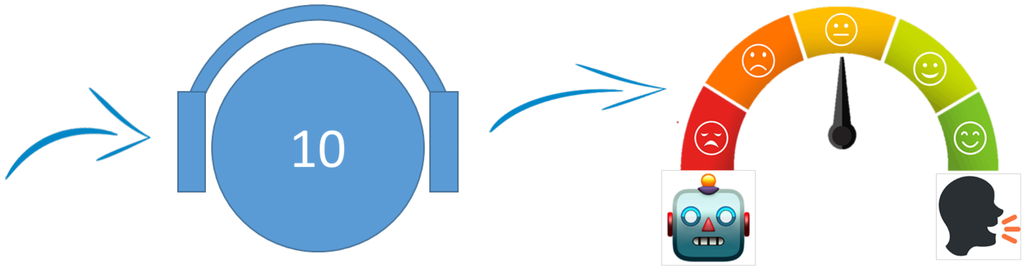
- 53 Japanese sentences
  - Neutral sentences from daily life
    - E.g. “Tokaide wa deau hito no hotondo ga mishinaru hito dearu”
      = “Most people we meet in the city are strangers.”
- @5 types:
  - Natural = The original recording
  - Analysis-synthesis = The reconstructed speech using natural MCC and LF0
  - DNN_synthesized :
    - LF0 = Only LF0 (pitch) feature is synthesized
    - MCC = Only Mel-Cepstrum feature is synthesized
    - LF0 + MCC = Both LF0 and MCC are synthesized

  - DNN : Feed forward (488 – 3*512 ReLU – 127 Linear), procedure minimizing Mean Squared Error
  - Train data : 2250 utterances
  - Output features : 40-dimensional Mel-Cepstrum, LF0, etc.

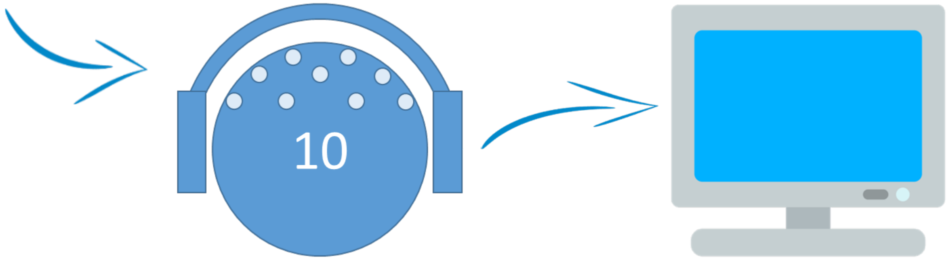
Natural	Ana-Syn	Syn LF0	Syn MCC	Syn LF0 & MCC
🔊	🔊	🔊	🔊	🔊

Used Stimuli and Experiment Scenario

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



Listen to the speech and give score 1~5



Record EEG while listening the speech

The collected opinion score from 10 people averaged (MOS) and used as the baseline.

The recorded EEG then preprocessed and used as the input for classification and regression step.

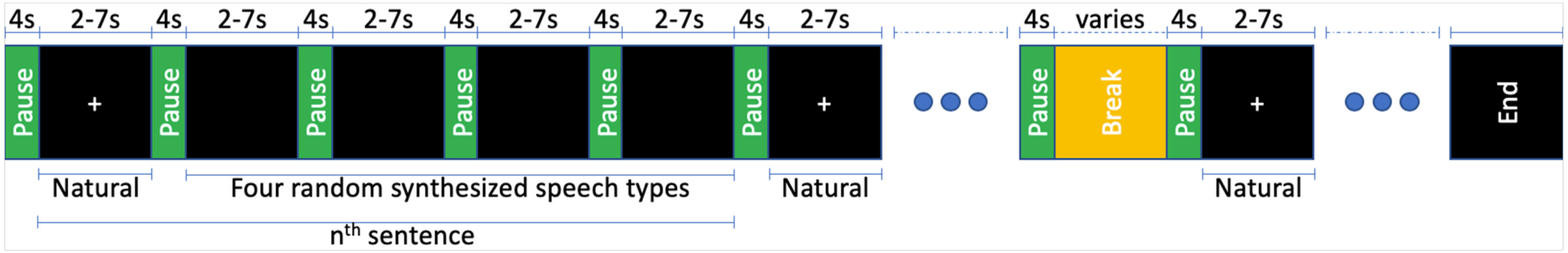


Figure of stimuli presentation during EEG recording

This research aims to find out whether EEG could be used to determine overall MOS of the synthesized speech.

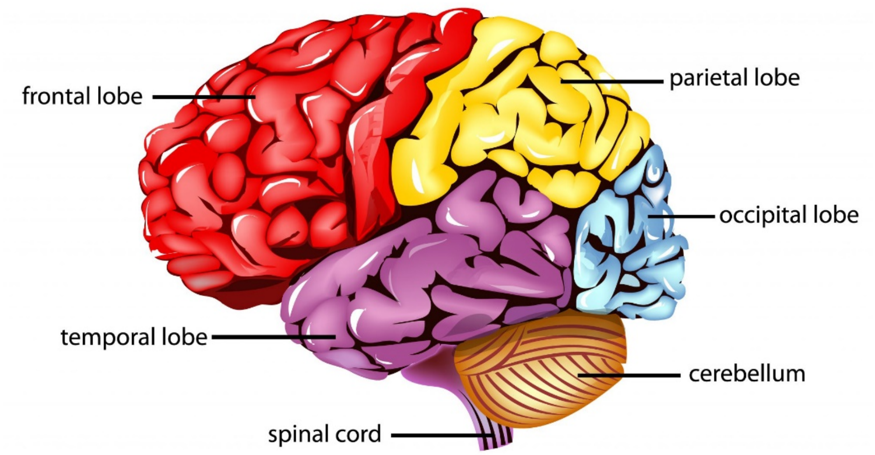
EEG Overview

- Recorder used : ActiCap, Brain Products GmbH
  - 32 channels

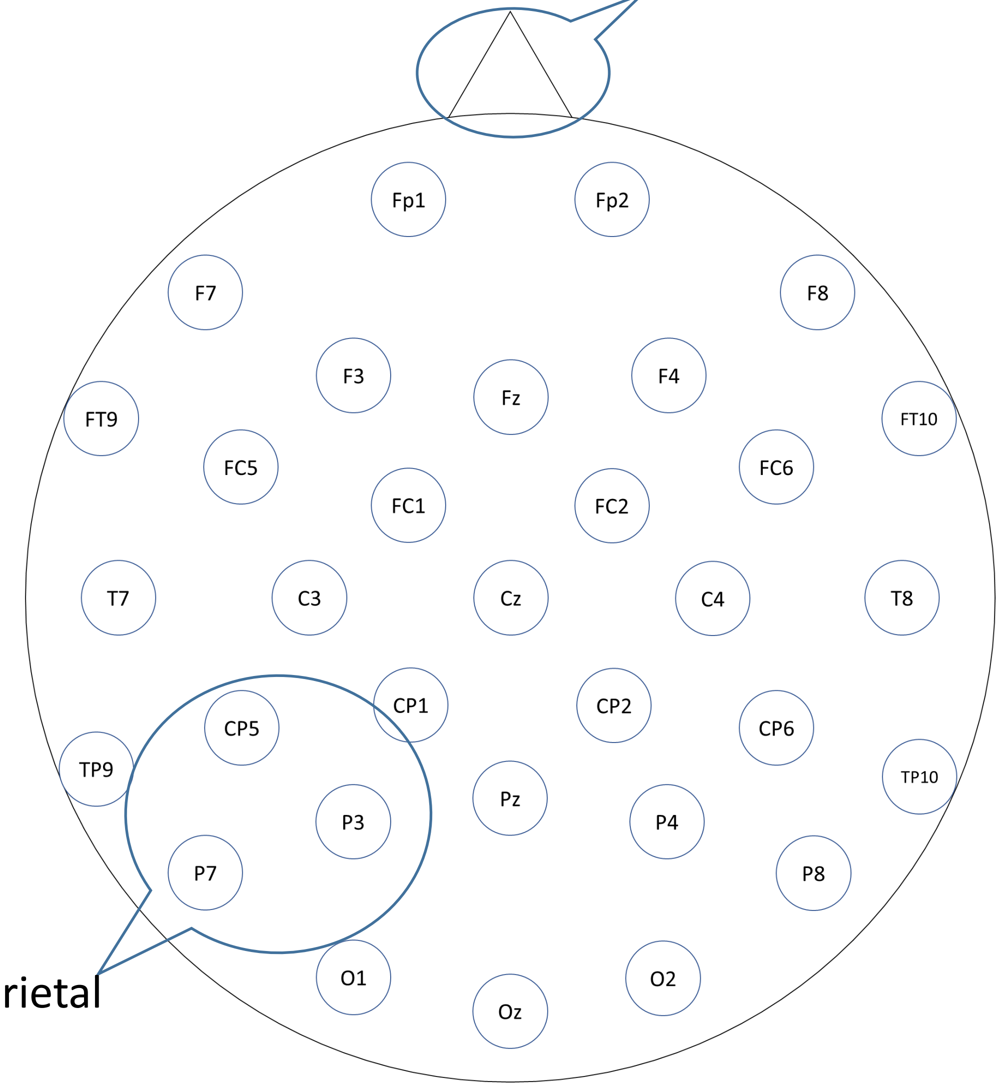
EEG frequency bands:

- Theta : 4-8 Hz
- Alpha : 8-13 Hz
- Beta : 13-32 Hz
- Gamma : 32-40 Hz

Parts of the Human Brain



URL: neeuro.com/prefrontal-cortex-exercises/

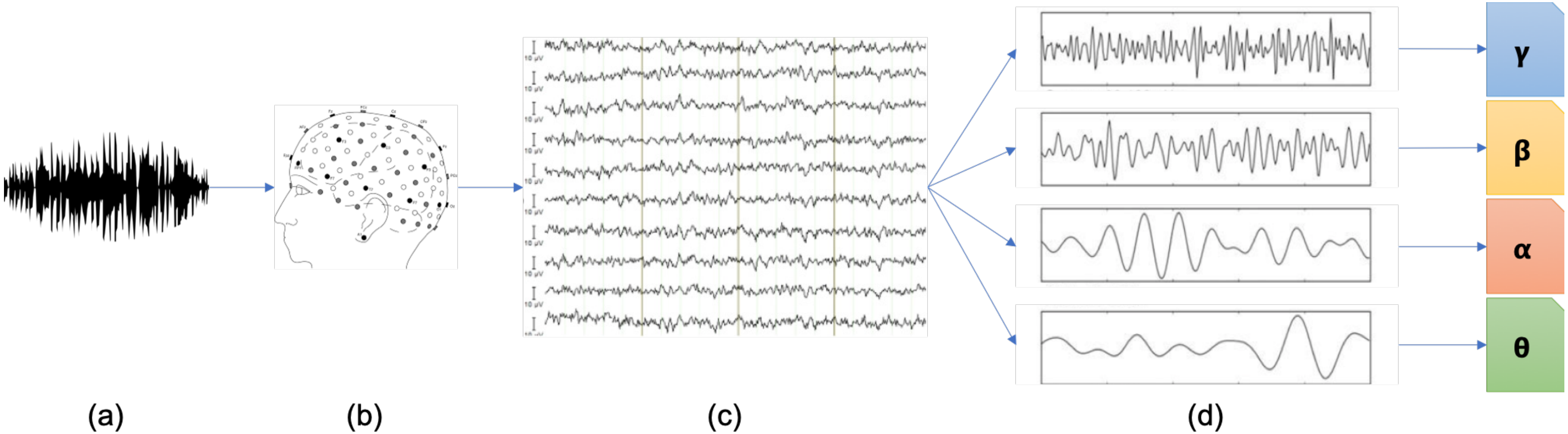


Nose

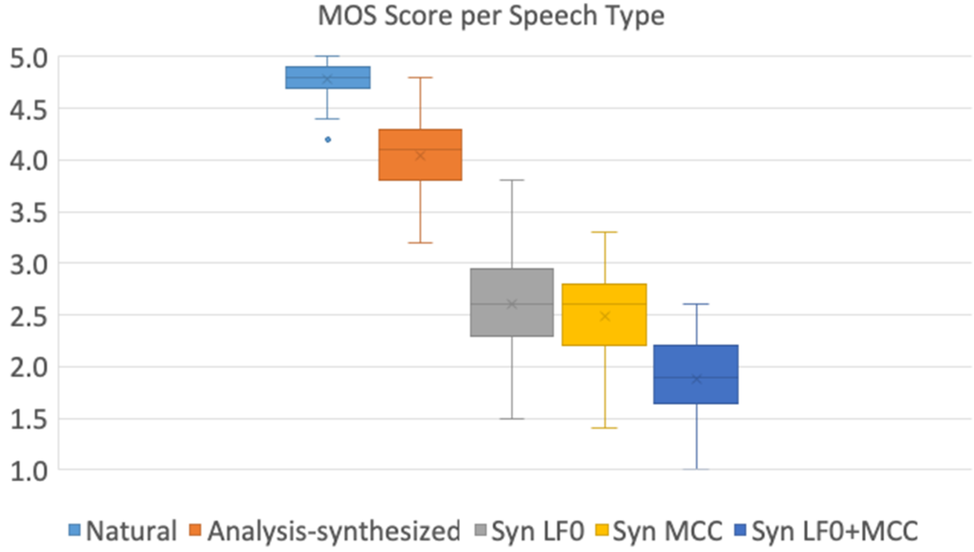
Left parietal

EEG Recording and Feature Extraction

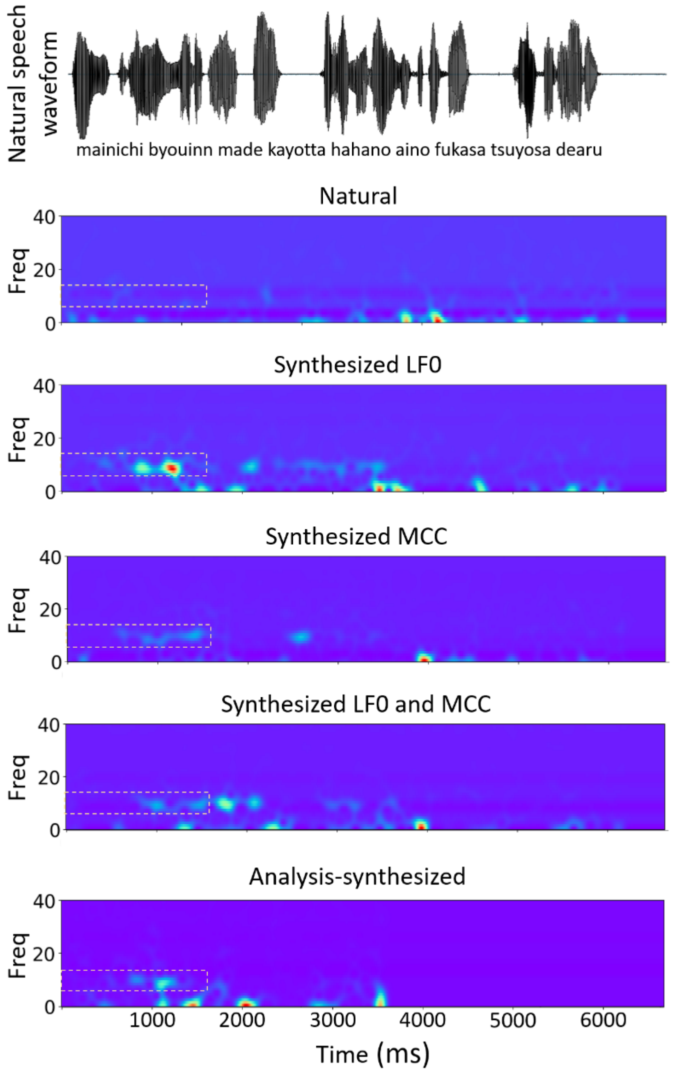
a) Stimuli presented to the subject
b) Subject listens to the stimuli
c) EEG recorded, and 0.5 Hz – 38 Hz band pass filtered
d) EEG down-sampled, separated into theta, alpha, beta, and gamma by wavelet transformation



Results: Collected MOS and EEG Response Comparison



MOS of the full dataset used in the experiment. Syn LF0 and Syn MCC are scored very close to each other.



Difference in EEG response of left parietal in time-frequency domain of the same sentence for each type of speech

Results: Classification and Regression

- Classification [Accuracy]

Nat vs. Ana-Syn		
	LDA	*SVM*
ϑ	71.50	49.66
α	76.87	73.66
ϐ	**80.36**	25.44
γ	70.26	80.00

Nat vs. Syn LF0		
	LDA	*SVM*
ϑ	73.75	**96.61**
α	78.99	76.27
ϐ	75.38	95.70
γ	75.78	47.90

Nat vs. Syn MCC		
	LDA	*SVM*
ϑ	75.79	20.99
α	**79.39**	67.55
ϐ	78.22	18.77
γ	68.94	29.11

Nat vs. Syn LF0 & MCC		
	LDA	*SVM*
ϑ	73.27	35.00
α	79.36	54.77
ϐ	76.71	**85.00**
γ	69.46	16.33

- Regression [RMSE]
  - 1: lowest quality, 5: highest quality

Band	***Linear Regression***	***SVR***
θ	1.353	1.142
α	1.726	1.143
β	2.114	**1.133**
γ	4.229	1.139

Conclusion and Future Work

- This work separated each frequency band as opposed to using full band range in the previous work.
- This study also found that Logistic regression is not well suited for this classification task.
- Possible to only use theta, alpha, and beta; reducing the input size
  - Gamma band is not useful for this task
- Possible to classify and predict MOS with EEG
  - Classification result achieves at least 79.39% accuracy
  - Regression result achieves 1.133 RMSE
- Future work:
  - Bias towards natural speech as the experimental task design
  - Need to use total random scenario for synthesized speech difference analysis

Thank you