

Combining Audio and Brain Activity for Predicting Speech Quality

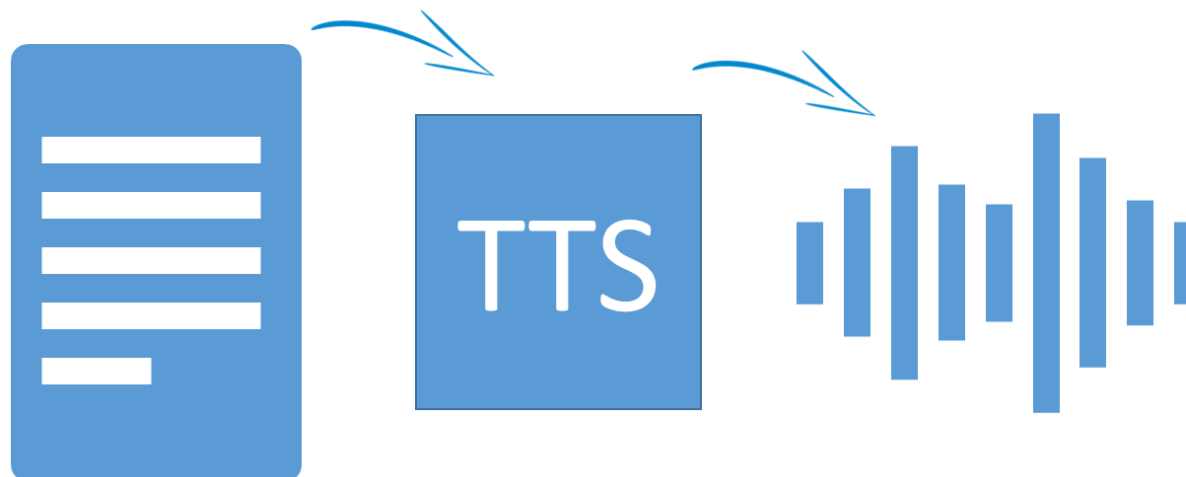
Ivan Halim Parmonangan¹, Hiroki Tanaka^{1,2}, Sakriani Sakti^{1,2},
Satoshi Nakamura^{1,2}

¹Division of Information Science, Nara Institute of Science and Technology,
Japan

²Center of Advanced Intelligence Project, RIKEN, Japan

Introduction

- Synthesized speech overview

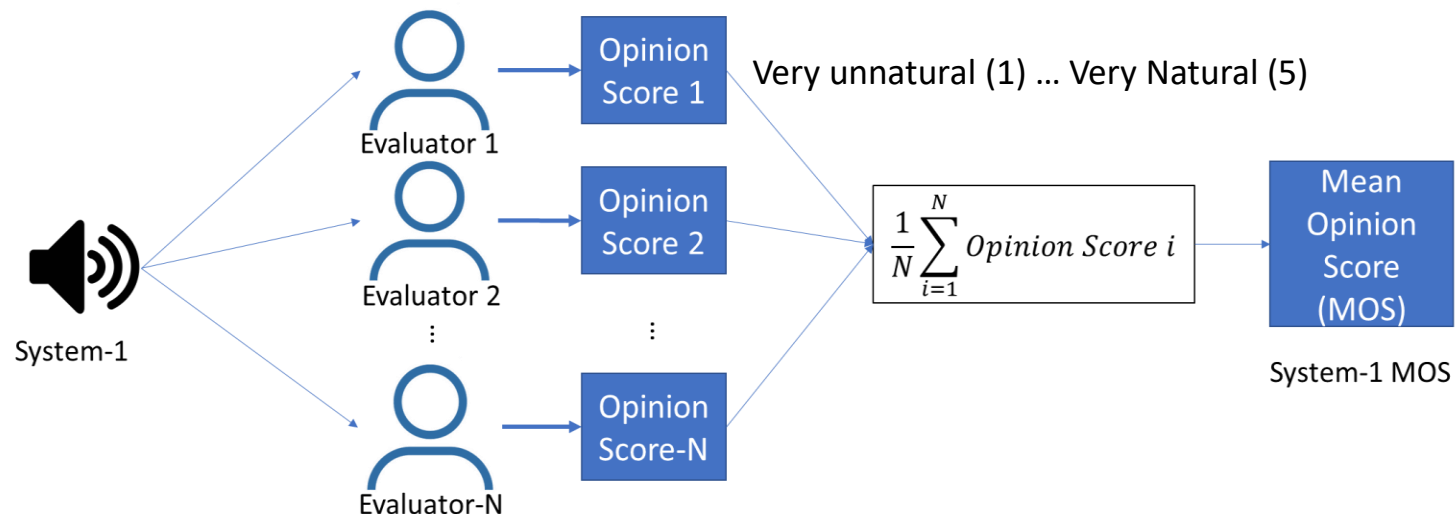


- A system that produce audible speech from a text input.
- One of many factors that determines its success:
 - Overall impression audio quality



Synthesized Speech Evaluation

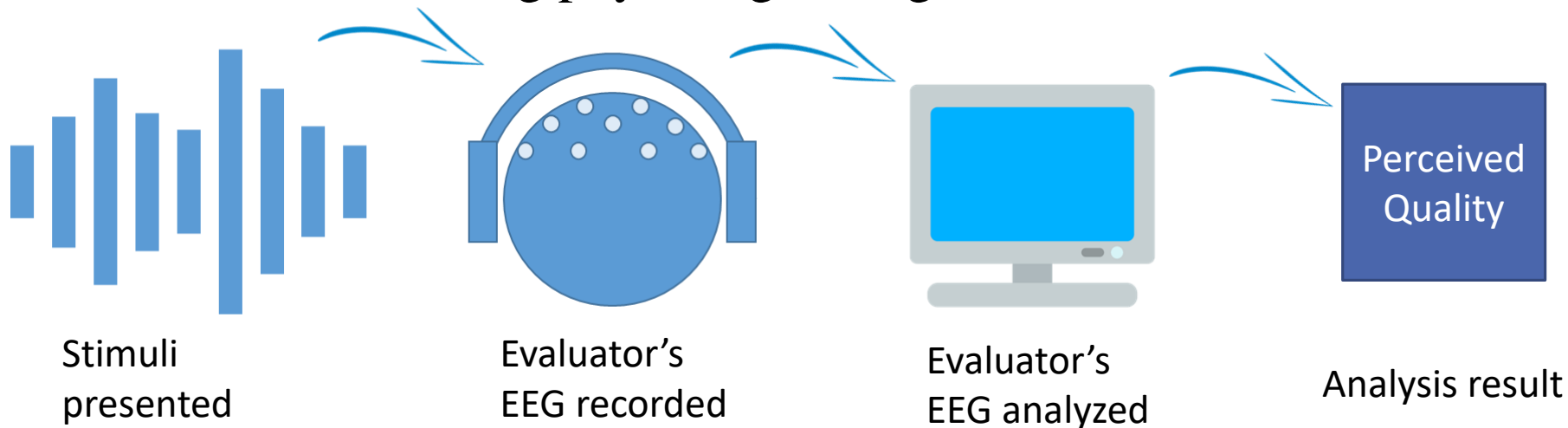
- Subjective evaluation (e.g. naturalness, intelligibility, etc.)
 - Usually done by calculating mean opinion score (MOS) or preference test (e.g. ABX test)
 - No insight about subject or evaluator's cognitive state [Maki et al., 2018]



- Objective evaluation: Analyze audio features (e.g. mel-distortion etc.)
 - No human evaluator involved
 - Fast & efficient
 - Relationship to human perceived quality is still unclear [Mayo et al, 2011]

Physiological Signals for Synthesized Speech Evaluation

- Physiological approach (e.g. brain activity, heart rate, skin conductance, etc.)
 - Not easy to conceal
 - Characterize evaluators' cognitive state (e.g. mental and emotional) [Gupta et al., 2016]
 - Brain is where judgement process and quality formation takes place [Antons et al., 2014]
- Typical workflow of utilizing physiological signal:





Related Works

[Maki et al., 2018]

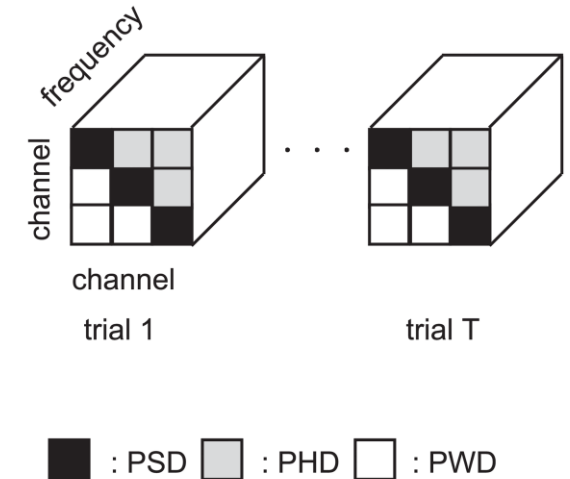
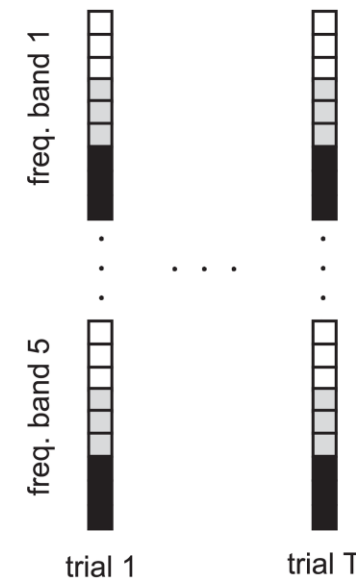
- Evaluated TTS with EEG (*electroencephalograph*)
- Regression method:
 - Partial Least Square (PLS) with linear vector [average RMSE: 1.098 ± 0.088]
 - High-order PLS (HOPLS) with tensor structure [average RMSE: 0.987 ± 0.104]
- Did not use audio features

(a) Vectorized features (Matrix)

(b) Tensorial features (4th order tensor)

EEG frequency band range:

- Delta (δ) <4Hz
- Theta (θ) 4-8Hz
- Alpha (α) 8-15 Hz
- Beta (β) 15-32 Hz
- Gamma (γ) >32Hz





Related Works

[Gupta et al., 2016]

- Evaluated TTS using mixed audio with EEG
- Using multiple linear regression
 - Showed how audio features (MFCC & F0) and EEG features* are correlated to the perceived quality
- Modelled to fit each subjects' data

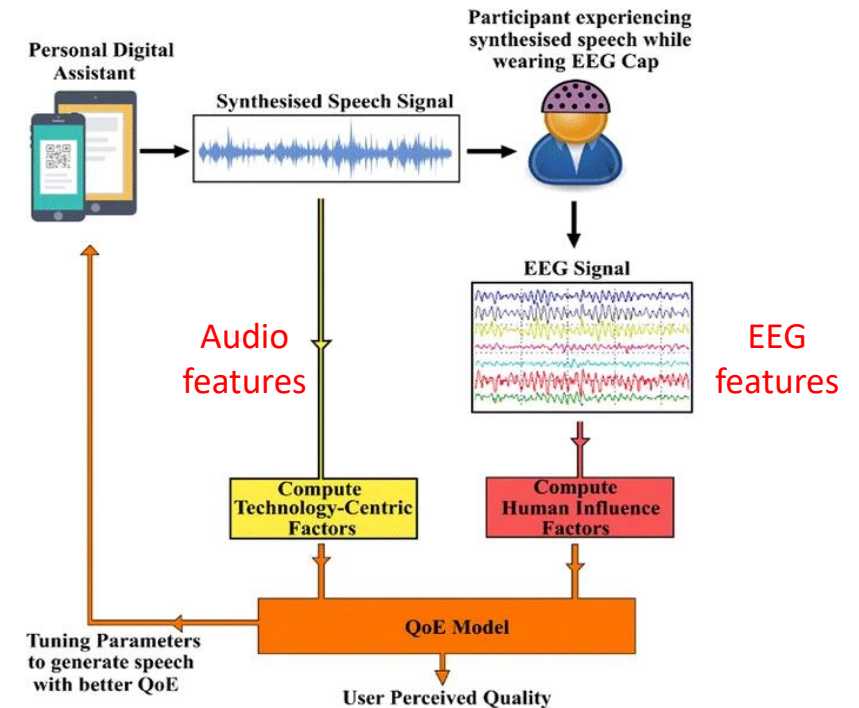
$$y_i = \epsilon_i + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_N x_{iN}$$

Opinion Score

Error

EEG/audio Features

Coefficient



*(Asymmetric Index & Medial Prefrontal Beta Power)



Proposed Method

1. Neural network based MOS regression

- Robustness in processing noisy data such as EEG signal processing [Subasi and Ercelebi, 2005]
- Previous work used PLS to perform regression [Maki et al., 2018]
- This work used Convolution Neural Network (CNN)
 - Ability to extract features with minimal feature engineering

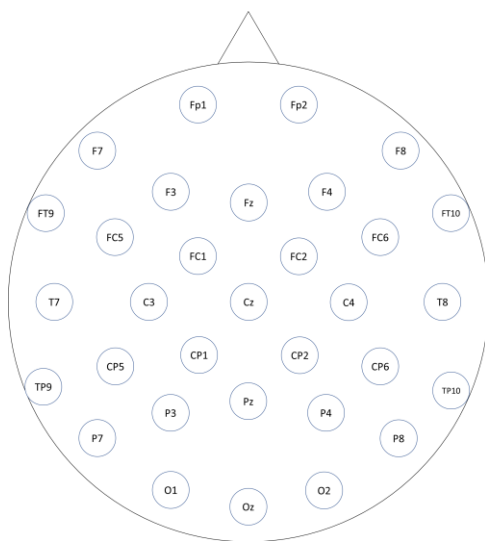
2. Combining brain activity and audio features to perform regression

- Multi-source input improved prediction performance [Kwon et al., 2018; Oramas et al., 2018]
- Previous work combined the features using multiple linear regression without performing regression to unseen data [Gupta et al., 2016]
- This work combines the features using deep learning to perform regression.

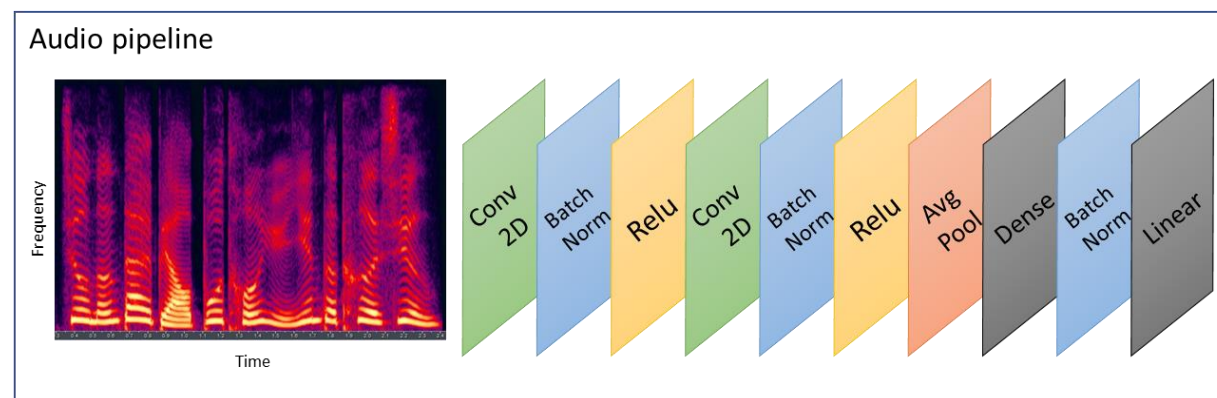
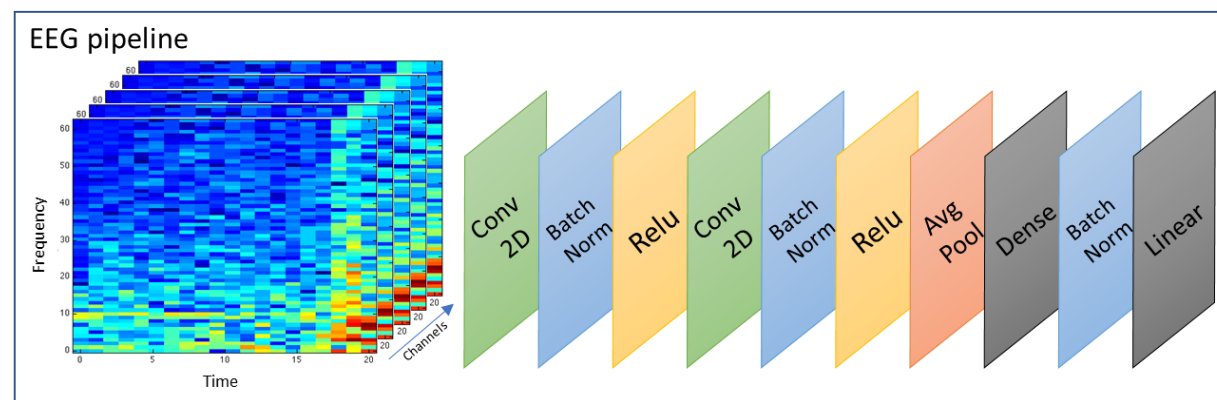


CNN Pipeline for Brain Activity and Audio

- 2D Convolution Layer (2 layers)
 - Kernel design adapted from [Kwon et al., 2018]
- Input:
 - 64 channels EEG spectrogram
 - 1 channel audio mel-spectrogram



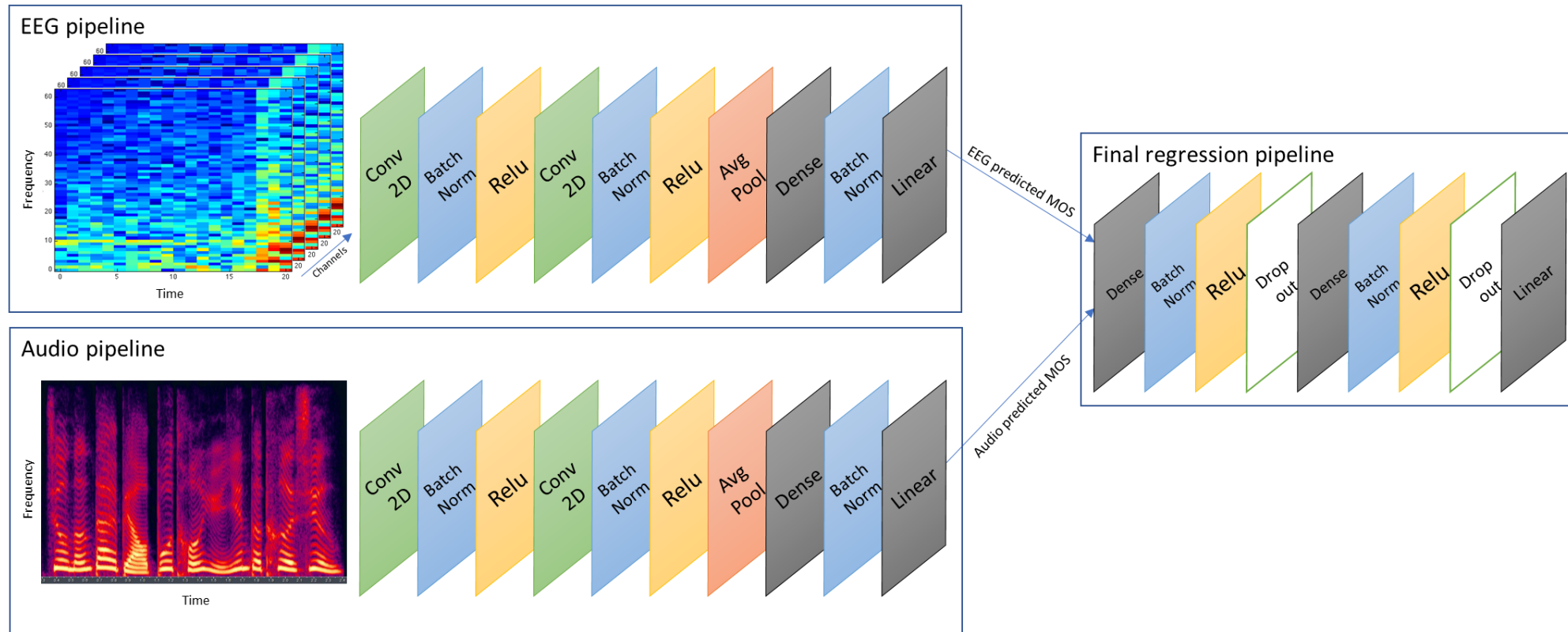
Example: 32 channels EEG topography





Combining Brain Activity & Audio

- Late integration approach
- Final regression pipeline: Two fully connected layers














Experiment Setup

- Dataset:
 - English TTS and EEG data: PhySyQX [Gupta et al., 2015]
- The baseline [Maki et al., 2018]:
 - Input: power spectrum density (PSD), channel paired phase and power spectrum density (PHD & PWD)
 - Used Partial Least Square regression (PLS)
 - Objective function: MOS (Mean Opinion Score) [very unnatural (1) ... very natural (5)]
- Metric:
 - Root Mean Squared Error (RMSE)
 - Significance test: Wilcoxon signed-rank test
 - ($\alpha = 0.01, N = 21, T = 42$)
- Compare:
 1. (baseline) PLS_{EEG} vs. CNN_{EEG}
 2. CNN_{EEG} vs. $CNN_{aud+EEG}$



PhySyQx - Audio Dataset

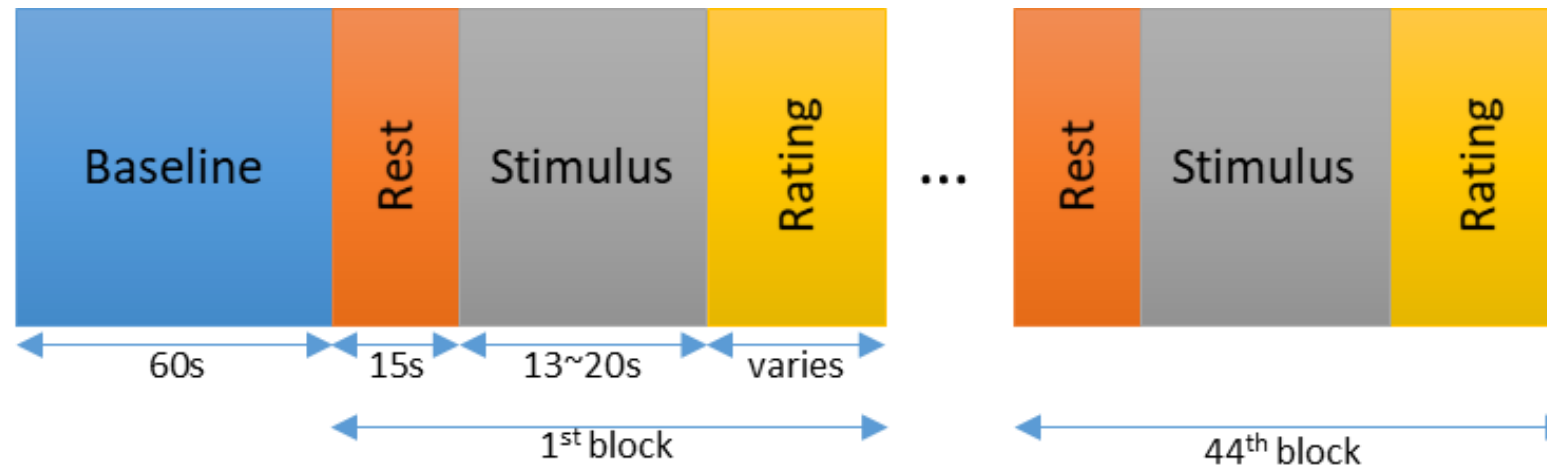
- Speech Audio (36 total samples)
 - Language: English
 - Natural & synthesized
 - Male & female
 - Synthesized using commercially available TTS systems

| Types | Audio sample |
|-------|---|
| 1 |  |
| 2 |  |
| 3 |  |
| 4 |  |
| 5 |  |
| 6 |  |
| 7 |  |
| 8 |  |
| 9 |  |



PhySyQx - Physiological Signal Dataset

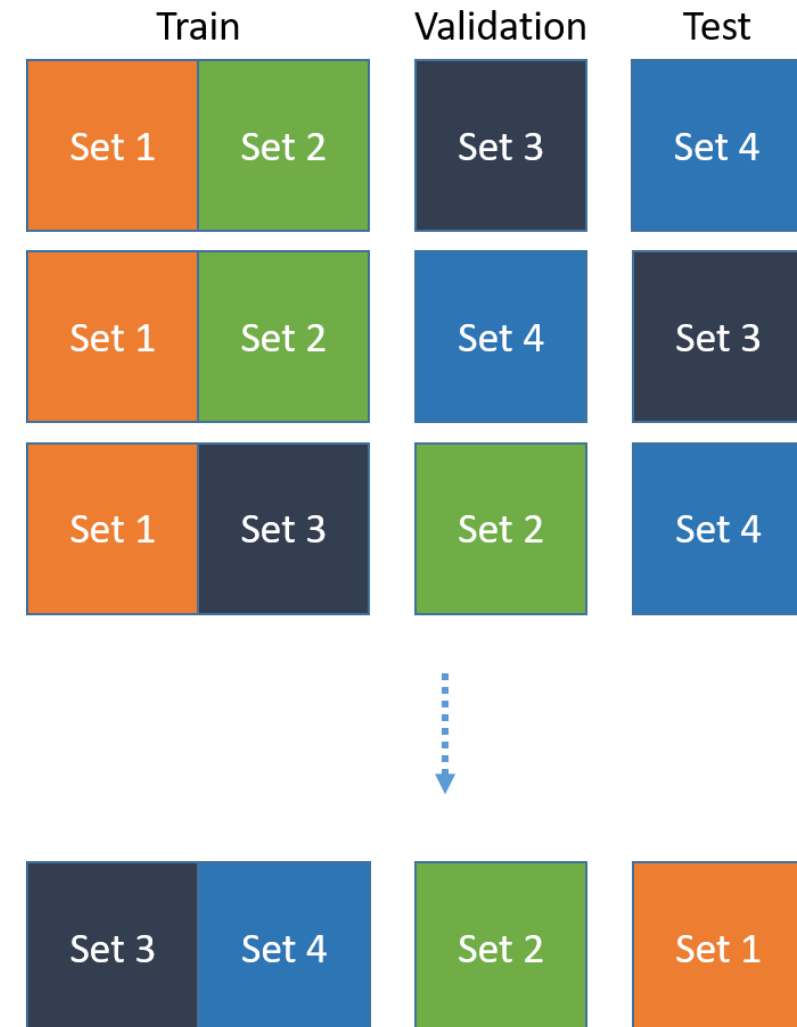
- EEG & fNIRS
 - 21 evaluators
 - Each listened to 44 speech audio stimuli
 - This work used only the EEG
- Stimuli presentation





Cross Validation Setup

- Audio data
 - 36 samples available
 - Separated into 4 sets
 - Train-Validation-Test : 18-9-9 audio samples
- EEG data:
 - 21 evaluators
 - Subject dependent
 - Same person : 18-9-9 EEG records





Result

- CNN_{EEG} has significantly lower RMSE than PLS_{EEG} ($W = 27, W < T$)
- CNN_{aud} has lower RMSE than CNN_{EEG}
- $\text{CNN}_{\text{aud+EEG}}$ has significantly lower RMSE than CNN_{EEG} ($W = 0, W < T$)
- Combining the audio and EEG improved the result significantly

| Sbj. | PLS_{EEG} | CNN_{EEG} | $\text{CNN}_{\text{aud+EEG}}$ | CNN_{aud} |
|--------|---------------------------|---------------------------|-------------------------------|---------------------------|
| 1 | 1.102 | 1.084 | 0.752 | 0.862 |
| 2 | 0.990 | 0.974 | 0.767 | - |
| 3 | 0.948 | 1.019 | 0.737 | - |
| 4 | 0.997 | 1.010 | 0.719 | - |
| 5 | 1.007 | 0.947 | 0.750 | - |
| ... | | | | |
| 18 | 1.075 | 1.010 | 0.742 | - |
| 19 | 0.971 | 1.034 | 0.709 | - |
| 20 | 0.937 | 0.927 | 0.714 | - |
| 21 | 1.091 | 0.936 | 0.694 | - |
| Mean | 1.122 | 0.984 | 0.732 | 0.862 |
| Stdev. | 0.275 | 0.037 | 0.017 | - |

Conclusion

- Physiological signals for Text-to-Speech audio quality evaluation
- Proposed methods:
 - Neural network based MOS regression
 - Combining EEG and audio features
- The proposed method results:
 - The proposed NN-based MOS regression has significantly lower RMSE than the PLS baseline
 - Combined method has significantly lower RMSE than single source input

Future Work

- Investigate the performance on subject-independent case
- Explore different fusion method such as early-fusion or tensor fusion [Zadeh, Amir et al, 2017]
- Investigate which EEG features could further improve the performance.
- Experiment with other audio features such as mel-cepstrum or LF0.
- Investigate different model to handle each brain activity and audio features such as combining CNN and BiLSTM [Lo, Chen-Chou et al., 2019].

Thank You

References

1. C. Mayo, R. A. Clark, and S. King, “Listeners weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis,” Speech Communication, 2011
2. J.-N. Voigt-Antons, S. Arndt, R. Schleicher, and S. Moller, Brain Activity Correlates of Quality of Experience, 2014.
3. 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
4. Y.-H. Kwon, S.-B. Shin, and S.-D. Kim, “Electroencephalography based fusion two-dimensional (2d)-convolution neural networks (cnn) model for emotion recognition system,” Sensors, 2018
5. H. Maki, S. Sakti, H. Tanaka, and S. Nakamura, “Quality prediction of synthesized speech based on tensor structured eeg signals,” 2018
6. R. Gupta, H. J. Banville, and T. H. Falk, “Phsyq: A database for physiological evaluation of synthesized speech quality-of-experience,”in 2015 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), 2015, pp. 1–5.
7. Oramas S, Barbieri F, Nieto O, Serra X. Multimodal Deep Learning for Music Genre Classification. Transactions of the International Society for Music Information Retrieval. 2018
8. Lo, Chen-Chou et al. “MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion.” Interspeech (2019).
9. Abdulhamid Subasi and Ergun Ercelebi, ”Classification of EEG Signals using Neural Network and Logistic Regression,”_Computer Methods and Programs in Biomedicine, Vol.78, May 2005
10. Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis.CoRR, abs/1707.07250, 2017