INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Speech Quality Evaluation of Synthesized Japanese Speech using EEG

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
[2] RIKEN, Center for Advanced Intelligence Project AIP, Japan
[3] Graduate School of Information Science and Technology, The University of Tokyo, Japan

{ivan.halim_parmonangan.ia4, hiroki-tan, ssakti, s-nakamura}@is.naist.jp,
shinnosuke_takamichi@ipc.i.u-tokyo.ac.jp

Abstract

As synthesized speech technology becomes more widely used, the synthesized speech quality must be assessed to ensure that it is acceptable. Subjective evaluation metrics, such as mean opinion score (MOS), can only provide an overall impression without any further detailed information about the speech. Therefore, this study proposes predicting speech quality using electroencephalographs (EEG), which are more objective and have high temporal resolution. In this paper, we use one natural speech and four types of synthesized speech lasting two to six seconds. First, to obtain ground truth of MOS, we gathered ten subjects to give opinion score on a scale of one to five for each recording. Second, another nine subjects were asked to measure how close to natural speech each synthesized speech sounded. The subjects' EEGs were recorded while they were listening to and evaluating the listened speech. The best accuracy achieved for classification was 96.61% using support vector machine, 80.36% using linear discriminant analysis, and 59.9% using logistic regression. For regression, we achieved root mean squared error as low as 1.133 using SVR and 1.353 using linear regression. This study demonstrates that EEG could be used to evaluate the perceived speech quality objectively.

Index Terms: EEG; synthesized speech; text-to-speech; quality assessment

1. Introduction

Rapidly spreading modern systems such as mobile assistants, smart devices, and navigation systems mostly have one thing in common: they can synthesize speech. Unlike human speech, machine synthesized speech sometimes includes unnaturalness such as missing pitches, mispronunciations, and strange pauses. Hence, the synthesized speech quality must be evaluated.

The naturalness of synthesized speech is usually measured objectively or subjectively. Subjective measurement usually involves calculating opinion scores (e.g., mean opinion score (MOS) and preference tests) [1]. Although this approach might be the most natural method, it can only provide an overall impression without any further detailed information about the speech.

Objective measurement is usually done by a computer. There are several method to evaluate the synthesized speech quality such as calculating the RMSE of F0, unvoiced/voiced (U/V) prediction errors, and several other methods. For example, by using linear regression, a study by [2] shows decent correlation between subjective Diagnostic Acceptability Measure (DAM) with cepstral distance (CP) and Mel-cepstral distance (MCD). However, the exact relationship between acoustic features and perceived quality is yet to be understood [3]. Therefore, even though the predicted quality is high, the naturalness of the synthesized speech might not meet human expectations.

This study proposes utilizing neurophysiological signals such as brain activity [4, 5, 6], to understand how listeners react towards synthesized speech of differing quality. A previous study showed that electroencephalographs (EEG) could reveal the correlation between pitch emphasis and brain activity [7]. In [8], they 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. In addition, by using tensor representation of all channels and all frequency bands, a study conducted by [9] shows that EEG signals could be used to predict MOS, valence, and arousal within the same subject. We also previously examined which EEG electrodes, frequency bands, and time length significantly represent perceived speech quality in Japanese using the generalized fisher scores [10].

In this paper, we focused more on whether machine learning algorithms could objectively evaluate different speech qualities from EEG. To do so, we compared several traditional classification and regression models with different frequency bands within 4-38 Hz. We used support vector machine (SVM), linear discriminant analysis (LDA), and logistic regression (LR) for the classification. For the regression, we used linear regression and support vector regression (SVR).

2. Synthesized Speech Materials

To construct synthesized speech, we generally need to extract speech parametric representation (i.e., mel-cepstrum) and excitation parameters (i.e., log F0) from an original speech database. Then, by using a set of generative models, those parameters are modelled. This study used a deep neural network (DNN) as a generative model that learns the correspondence between text and speech in an attempt to generate speech parameters given a word sequence. Finally, speech waveforms are reconstructed from the parametric representations of speech (mel-cepstrum and log F0) using a vocoder. LF0 is a measure of fundamental frequency in speech or the voice pitch. Mel-cepstrum representation is a parametric model for the spectral envelope of speech in which frequency resolution mimics the human auditory system. The content of the speech is daily conversation sentences, for example:

"Tokaide wa deau hito no hotondo ga mishiranu hitodearu." (Romaji)

"Most of the people we meet in the city are strangers." (English translation)

http://dx.doi.org/10.21437/Interspeech.2019-2059

2.1. Speech Stimuli

This study used 53 single-speaker Japanese sentences recorded in natural speech and four types of synthesized speech with 16 kHz sampling rate. The used speech length ranged from two to six seconds with an average of 3.43 seconds and a standard deviation of 1.35 seconds.

2.1.1. Natural speech

Natural speech was the original recording of human speaker.

2.1.2. Analysis-synthesized speech

Analysis-synthesized speech was not generated using a generative model but rather reconstructed using original features of natural speech. Speech features (mel-cepstrum and LF0) were extracted from the natural recordings and then transformed back into speech waveforms. It was used as the baseline of generated DNN-based synthesized speech.

2.1.3. DNN-based speeches

Three types of speech are generated using predicted features from the generative model that generates both LF0 and mel-cepstrum. In the last step, the waveform is generated using the vocoder on the basis of the predicted features. In this study, we prepared DNN-based speech synthesis based on based on [11].

- Synthesized LF0. This is speech waveforms reconstructed using generated LF0 and natural mel-cepstrum assuming that the model predicts the mel-cepstrum perfectly.
- Synthesized MCC. This is speech waveforms reconstructed using generated mel-cepstrum and natural LF0 assuming that the model predicts the LF0 perfectly.
- Synthesized LF0 and MCC. This is speech waveforms reconstructed using both generated LF0 and mel-cepstrum. The real Text-to-Speech (TTS) system usually uses both generated parameters.

2.2. Mean Opinion Score

In this study, MOS is used to subjectively observe how the quality of each speech type is distributed. For MOS, we collected opinion scores from ten subjects: nine males and one female aged 24 to 27 years old. All subjects had normal hearing without hearing aids. We asked the subjects to give scores ranging from (1) very bad to (5) very good. The speech used in both the MOS collection and EEG recording session was the same. Figure 1 represents how subjects differentiated the quality among the categories. Both synthesized LF0 and synthesized MCC were scored similarly while natural and analysis-synthesized were scored significantly higher. We found a statistical significance for all comparisons using paired t-tests with Bonferroni correction except for Synthesized LF0 and Synthesized MCC (α=0.05).

3. Methods

We collected the subjects' EEGs while they were listening to natural and synthesized speech. We used MNE python library [12, 13] to pre-process the recorded EEGs and finally used the Scikit-Learn [14] python library to perform classification and regression.

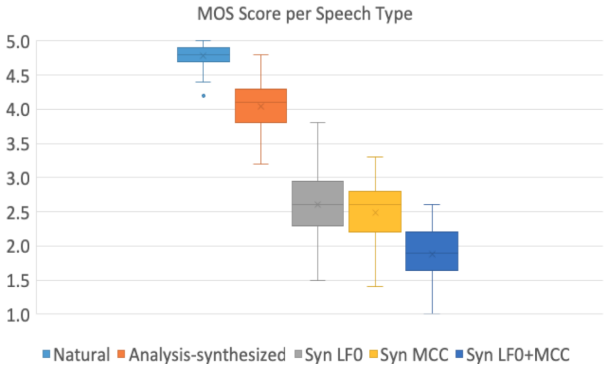

Figure 1: *Boxplot of MOS on every speech type.*

3.1. EEG Data Collection

3.1.1. Subjects

This research was approved by the ethical committee of the Nara Institute of Science and Technology. We collected EEGs from another ten subjects, seven males and three females, aged 23 to 35 years old. Subjects were provided with a brief explanation about the experiment. All subjects were right-handed native Japanese speakers without any medical history of brain injuries or severe brain trauma.

3.1.2. Experiment Procedure

During the EEG recording session, all 53 sentences were played in randomized order. For each loop, the natural speech was firstly played, followed by four randomly ordered types of synthesized speech of the same sentence as shown in Figure 2. While listening to the natural speech, subjects were asked to remember how it sounds and when the synthesized speeches started to play, subjects were asked to remember how it sounded, and when the synthesized speech started to play, the subjects were asked to press a button every time they heard parts that sounded bad.

EEG experiment was done in a dimly lit soundproof room. The subjects listened to the speech recordings using an in-ear-monitor earphone setup. Finally, the subjects were told to avoid blinking and making excessive body movement while listening to the speech to minimize unnecessary noises in the recordings. EEG data were then recorded throughout the experiment using ActiCAP with 32 scalp channels and a BrainAmp DC, both from Brain Products. Figure 3 shows the diagram of the full recording and pre-processing procedure.

3.1.3. Pre-processing

The recorded EEG was referenced on the FCz channel during recording session and re-referenced using common average reference. The band-pass filter from 0.5 to 40 Hz was applied followed by down-sampling from 1000 to 250 Hz. The down-sampled EEG was then split per epoch to remove breaks and pauses during the recording, leaving only epochs when the subjects listened to the speech. Epochs of amplitudes above 400 μV or below -400 μV were rejected assuming they were contaminated with large amplitude artifacts. One subject was removed from the analysis process due to more than 10% of his recorded samples being removed during this procedure. Each epoch aligns to the corresponding speech record played. Since

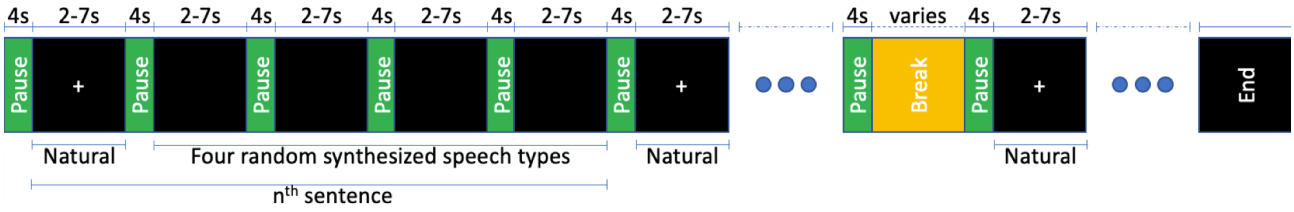

Figure 2: *EEG Scenario layout. 53 sentences are played in random order. Each sentence will began with a natural speech marked with a '+' sign on the screen followed by four types of randomly ordered synthesized speech with a blank screen. Every five sentences, the subject were given a break time of which the length depended on the subject and therefore varied.*

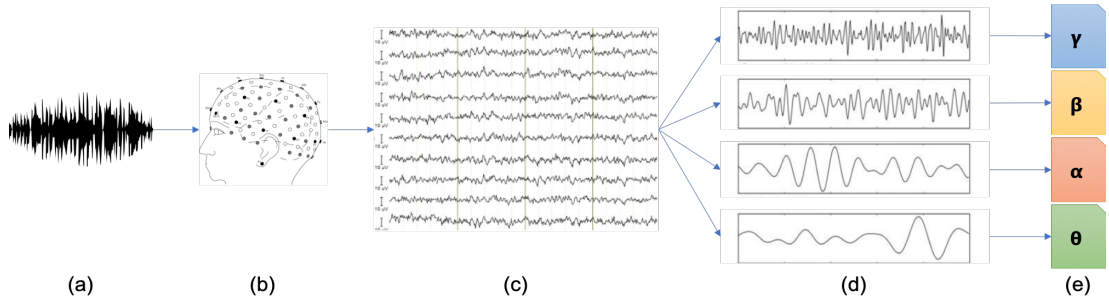

Figure 3: *Complete EEG processing diagram. (a) is speech stimuli, (b) is recording process, and (c) is recorded EEG from each electrodes. The recorded EEG was then pre-processed with re-referencing, band-pass filtering, and down-sampling and then split per epoch. Every epoch, the EEG was divided into four frequency bands (d) and saved as a dataset (e).*

the lengths of the epoch varied, we zero-padded the EEG to the maximum record length.

3.2. Wavelet transform and feature extraction

The epochs were converted from the time-amplitude domain into time-frequency representation using Morlet wavelet transform [15] and then separated on the basis of four frequency bands; theta (4-8 Hz), alpha (8-13 Hz), beta (13-31 Hz), and gamma (31-38 Hz). The final product of this procedure is four datasets based on the previously mentioned frequency bands. Each dataset contains five types of speech with 53 epochs each. Each record has six second data from 32 EEG electrodes. The exsample of EEG comparison at one channel before frequency band separation is shown in Figure 4. This shows visualization of synthesized speeches in terms of time-frequency analysis of the EEG. For example, we can see that strong EEG responses to specific frequency (around 10-20 Hz) are observed in the synthesized LF0.

3.3. Classification

The classification was done for every comparison between natural speech and the four types of synthesized speech: natural and analysis synthesis (reconstructed LF0 and MCC), natural and synthesized LF0, natural and synthesized MCC, natural and synthesized LF0 and MCC. For each frequency band, we compared the natural speech type with every synthesized speech type. This study is focused on finding out which classifier and regression method works best for evaluating synthesized speech quality. For the classification task, we used LDA, L1 regularized LR, and SVM with a 7-degree sigmoid kernel. We also tried using multi-layer perceptron (MLP), but the results were not better than the previously mentioned methods. Moreover, due to large input dimensions, MLP took tremendous amount of time to train. For the LR, L1 regularization was used because our EEG data is sparse which the important data is relatively small in comparison to the total input, therefore many input features should be reduced. We compared each model by 15-fold cross validation. Finally, we investigated which frequency band and the models produce highest overall accuracy.

3.4. Regression

In evaluating the synthesized speech quality, the classification is not sufficient enough because it can only categorize the speech quality. On the other hand, regression is able to predict slight differences in between the categories. On the basis of the classification results, we attempted to predict the actual MOS scores from EEG using the regression method. The regression was done per frequency band. For each frequency band, we tried to predict each record's MOS. Each record was scored on the basis of the MOS of ten people as described in section 2.2. We used L1 regularized multiple linear regression and SVR with a 7-degree sigmoid kernel. For each method, we compared each frequency band and models to find a best combination which produces lowest RMSE score.

4. Results

4.1. Classification

Table 1 show accuracy results of four frequency bands for LDA, SVM, and LR respectively. The results show that for the classification task, in general, alpha band is best for differentiating between natural speech and the four types of synthesized speech.

Table 1: *Classification results comparison.*

Type	Nat vs. ana-syn			Nat vs. syn LF0			Nat vs. syn MCC			Nat vs. syn LF0 and MCC		
	LR	LDA	SVM	LR	LDA	SVM	LR	LDA	SVM	LR	LDA	SVM
Theta	54.13	71.50	49.66	59.44	73.75	**96.61**	59.99	75.79	20.99	59.30	73.27	35.00
Alpha	53.28	76.87	73.66	53.56	78.99	76.27	53.64	**79.39**	67.55	62.29	79.36	54.77
Beta	49.73	**80.36**	25.44	51.33	75.38	95.70	51.08	78.22	18.77	56.28	76.71	**85.00**
Gamma	58.99	70.26	80.00	55.03	75.78	47.90	50.59	68.94	29.11	50.66	69.46	16.33

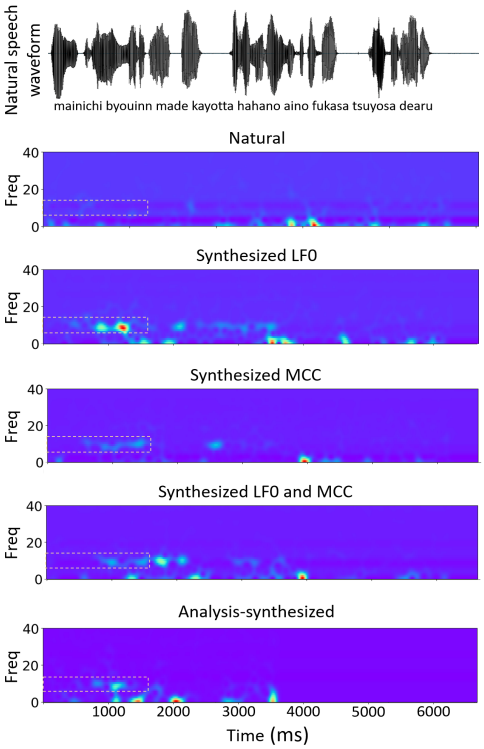

Figure 4: *Example of EEG wavelet comparison of the same sentence on left parietal electrode according to [10] from one subject for each speech quality. The marked regions show the difference in brain activity around 10Hz during the first 1500ms after the stimuli starts.*

By a one-tailed binomial test, we compared the chance rate (around 50%) and the model that achieved highest average accuracy for each comparison as shown in bold values in 1. From the result, we confirmed a statistical significance compared to the chance rate on all four comparisons ($p < 0.01$).

4.2. Regression

Table 2 compares linear regression and SVM based regression. From the table, we can see that RMSE of the linear regression reached as low as 1.353 at theta band. While, SVM has better RMSE and the results are distributed almost equally among frequency bands. Therefore, from this result, theta band is the best frequency band in order to do regression of the speech quality.

Table 2: *RMSE comparison between linear regression and SVR between predicted and actual MOS.*

Band	Linear Regression	SVR
Theta	1.353	1.142
Alpha	1.726	1.143
Beta	2.114	**1.133**
Gamma	4.229	1.139

5. Discussion

This research showed that EEG can be utilized to predict perceived speech quality. Both classification and regression results showed that EEG is effective for objectively predicting MOS. This study also found which frequency band is useful in order to reduce the complexity of models which will shorten the processing time. In comparison with [9], our study tried to generalize the approach across the subjects while the previous work was done within subject. Therefore, our approach may reduce the prediction performance. In the classification case, some frequencies perform better when used as input of certain comparisons. In the regression, we can conclude that SVM performs better than linear regression, however, the RMSE results were almost equal among frequency bands. Therefore, by comparing the results, theta band seems to perform better than any other frequency bands, followed by the alpha band.

6. Conclusion

In this study we proposed to evaluate speech qualities from EEG signals. We prepared MOS scores and four types of Japanese synthesized speeches. We collected EEG data from nine subjects during listening the speech samples, and extracted frequency band features using the Morlet wavelet transform. We constructed several methods to classify and predict the speech qualities. The results showed at least 79% accuracy and 1.133 of RMSE. For future work, we consider doing the evaluation within subject and using more sophisticated algorithms with higher dimensional data representation such as tensor representation. [16]

7. Acknowledgements

Part of this research is supported by JSPS KAKEN number JP17H06101, JP17K00237, and JP18K11437, and the Ministry of Internal Affairs and Communications.

8. References

[1] I. T. U. T. Recommendation, *A Method for Subjective Performance Assessment of Quality of The Speech Voice Output Devices*. Geneva: International Telecommunication Union, 1996.

[2] R. Kubichek, "Mel-cepstral distance measure for objective speech

quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, 06 1993, pp. 125 – 128 vol.1.

[3] C. Mayo, R. A. Clark, and S. King, "Listeners weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis," *Speech Communication*, vol. 53, no. 3, pp. 311 – 326, 2011.

[4] J. Antons, R. Schleicher, S. Arndt, S. Mller, and G. Curio, "Too tired for calling? a physiological measure of fatigue caused by bandwidth limitations," in *2012 Fourth International Workshop on Quality of Multimedia Experience*, July 2012, pp. 63–67.

[5] S. Arndt, J. Antons, R. Schleicher, S. Moller, S. Scholler, and G. Curio, "A physiological approach to determine video quality," in *2011 IEEE International Symposium on Multimedia*, Dec 2011, pp. 518–523.

[6] S. Arndt, K. Brunnstrm, E. Cheng, U. Engelke, S. Mller, and J.-N. Voigt-Antons, "Review on using physiology in quality of experience," *Electronic Imaging*, vol. 2016, pp. 1–9, 02 2016.

[7] C. Tang, L. S. Hamilton, and E. F. Chang, "Intonational speech prosody encoding in the human auditory cortex," *Science*, vol. 357, no. 6353, pp. 797–801, 2017.

[8] 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*, vol. 6, no. 1, p. 5, 2016.

[9] H. Maki, S. Sakti, H. Tanaka, and S. Nakamura, "Quality prediction of synthesized speech based on tensor structured eeg signals," *PLOS ONE*, vol. 13, no. 6, pp. 1–13, 06 2018.

[10] 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.

[11] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7962–7966.

[12] A. Gramfort, M. Luessi, E. Larson, D. Engemann, D. Strohmeier, C. Brodbeck, R. Goj, M. Jas, T. Brooks, L. Parkkonen, and M. Hmlinen, "Meg and eeg data analysis with mne-python," *Frontiers in Neuroscience*, vol. 7, p. 267, 2013.

[13] A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and M. S. Hmlinen, "Mne software for processing meg and eeg data," *NeuroImage*, vol. 86, pp. 446 – 460, 2014.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[15] A. Grossmann and J. Morlet, "Decomposition of hardy functions into square integrable wavelets of constant shape," *SIAM Journal on Mathematical Analysis*, vol. 15, no. 4, pp. 723–736, 1984.

[16] Q. Gu, Z. Li, and J. Han, "Generalized fisher score for feature selection," in *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, ser. UAI'11. Arlington, Virginia, United States: AUAI Press, 2011, pp. 266–273. [Online]. Available: http://dl.acm.org/citation.cfm?id=3020548.3020580