PLOS ONE

RESEARCH ARTICLE

Quality prediction of synthesized speech based on tensor structured EEG signals

Hayato Maki*, **Sakriani Sakti, Hiroki Tanaka, Satoshi Nakamura**

Graduate School of Information Sciences, Nara Institue of Science and Technology, Ikoma, Nara, Japan

* maki.hayato.lt3@is.naist.jp

OPEN ACCESS

Citation: Maki H, Sakti S, Tanaka H, Nakamura S (2018) Quality prediction of synthesized speech based on tensor structured EEG signals. PLoS ONE 13(6): e0193521. https://doi.org/10.1371/journal.pone.0193521

Editor: Christos Papadelis, Boston Children's Hospital / Harvard Medical School, UNITED STATES

Received: August 14, 2017

Accepted: February 13, 2018

Published: June 14, 2018

Data Availability Statement: The Physyqx data, which was used in this study, is third-party data from the following published paper: Gupta R, Banville HJ, Falk TH. PhySyQX: A database for physiological evaluation of synthesised speech quality-of-experience. Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA'15). 2015;1–5. DOI: 10.1109/WASPAA.2015.7336888. Researchers interested in accessing this data may contact Dr. Rishabh Gupta (grishabhg@gmail.com).

Abstract

This study investigates quality prediction methods for synthesized speech using EEG. Training a predictive model using EEG is challenging due to a small number of training trials, a low signal-to-noise ratio, and a high correlation among independent variables. When a predictive model is trained with a machine learning algorithm, the features extracted from multi-channel EEG signals are usually organized as a vector and their structures are ignored even though they are highly structured signals. This study predicts the subjective rating scores of synthesized speeches, including their overall impression, valence, and arousal, by creating tensor structured features instead of vectorized ones to exploit the structure of the features. We extracted various features to construct a tensor feature that maintained their structure. Vectorized and tensorial features were used to predict the rating scales, and the experimental result showed that prediction with tensorial features achieved the better predictive performance. Among the features, the alpha and beta bands are particularly more effective for predictions than other features, which agrees with previous neurophysiological studies.

Introduction

Text-to-Speech (TTS) systems, which convert a written text into speech, and are becoming more widely implemented in mobile phones, car navigation systems, and other consumer electronics. Such systems play a critical role in many applications because speech is the most fundamental and easiest communication tool for human beings. Therefore, synthesized speeches must sound natural for good machine-to-human communications.

The research of TTS systems needs reasonable criteria that evaluate the qualities of synthesized speeches. Several current evaluation methods have their own advantages and disadvantages: (1) subjective ratings [1–3], (2) analyzing a speech signal itself [4–6], and (3) measuring the physiological responses of listeners to speech [7–14].

In the first approach, the two most common aspects for quality judgment are naturalness and intelligibility. Naturalness describes how close synthesized speech is to human speech, and intelligibility reflects how well the speech content can be heard. The former is usually measured by a mean opinion score (MOS) test [1], and the latter is gauged by semantically unpredictable sentences (SUS) [3]. In addition, valence and arousal are often used to evaluate the

Funding: Part of this work was supported by JSPS KAKENHI (Grant Numbers JP17H06101 to SN, JP17K00237 to SS, and JP16K16172 to HT). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. No additional external funding received for this study.

Competing interests: The authors have declared that no competing interests exist.

subjective impressions of speech [11, 13, 15, 16] and to model emotions [17–20]. Valence reflects a positive or a negative emotion. Arousal reflects the degree of intensity or activation. In a MOS test, subjects listen to speech and rate its relative perceived quality on some kind of a scale, for example, "excellent," "good," "fair," "poor," "bad." Then the scores are averaged across subjects. This is well established method for which references on how to perform it are available [2], making it the only standard way to evaluate the naturalness quality of synthesized speech. However, their appropriateness has not been fully proven because high inter- and intra-subject inconsistencies are often observed in the ratings, resulting in poor reproductivity [21].

In the second approach, speech quality is automatically evaluated at its signal level by software that inputs a speech file and outputs the estimated speech quality. Advantages of these methods include complete reproductivity and less time consumption after such software is developed. However, appropriateness is difficult to prove because the exact relationship between the acoustic features and the perceived quality of speech by a listener is not well understood [21]. In fact, speech quality must be evaluated not only physically but also psychologically because it is commonly defined as an assessment result within which a listener compares his/her perceptions with expectations [22, 23].

Last, quality estimation methods are emerging that measure the physiological responses of a listener [24]. Even though these methods have not been established yet, they are worth investigating because physiological signals can be recorded automatically and continuously to provide insight about listener's cognitive states without interruptions caused by directly asking him/her to answer questions. Among existing non-invasive physiological response measures, electroencephalography (EEG) has especially great potential to estimate a listener's perceived speech qualities for the following reasons. EEGs can be recorded at a higher temporal resolution, e.g., a millisecond range, than hemodynamic measures, including functional magnetic resonance imaging (fMRI) and functional near-infrared spectroscopy (fNIRS), both of which analyze the changes in blood flow that inherently take a few seconds until a brain response can be recorded. Temporal resolution is important to evaluate speech quality since the temporal structure of speech largely affects its perceived quality. In addition, an EEG recording equipment is relatively small, less expensive than other brain recording equipments, and can be even wireless, which allows us to use it in daily environments, whereas fMRI and magnetoencephalography (MEG) can only be used in experimental rooms because of the lack of portability. Measuring physiological responses to speech in daily environments is critical because speech is everywhere. Despite the above advantages, the main disadvantage of physiological measures is the difficulty of data gathering. The amount of data that can be collected from a subject is limited for practical and ethical reasons. Conducting experiments is usually time-consuming and labor-intensive. In addition, physiological data are generally noisy and easily contaminated by artifacts. Furthermore, multi-channel EEG signals are usually highly correlated to each other, which makes the features extracted from them less informative compared to the height of their dimensions. These aspects of EEG (limited amount of data, noise, and high correlation and dimension) complicate training a predictive model with EEG data and require a sophisticated dimension reduction or regularization techniques [25].

Existing researches have analyzed EEG responses to speech stimuli using event-related potentials (ERP), which are time-locked responses to external or internal events in terms of a voltage change that are usually visualized and quantified after synchronous averaging of multiple epochs [7–9]. Due to its definition, measuring ERP need the instantaneous time-locking points at which an event occurs, complicating the use of ERP if stimuli onsets are gradual or unclear [26]. Therefore, ERP is not suitable for our purpose of the predicting perceived quality of speech whose length exceeds a second because it is usually unclear which time points affect

a listener's perceived quality. Other research used power spectral density [14, 27] and their difference between EEG channels [11, 13] at multiple frequency bands. Neuroscience studies reported that EEG spectral changes in distinct regions and between hemispheres are related to emotions [28–31]. Other studies used EEG phase synchronization between EEG channel pairs and found a correlation to emotions [32, 33].

The purpose of this research is to predict the perceived qualities of synthesized speeches using only EEG. Interest is growing in the development of a machine learning algorithm that uses an input/output data structure as tensor formats [34–36]. Such tensor structured features were investigated in this study because EEG signals can have structures in time, frequency, space, experimental condition, and other modalities.

Materials

We used the PhySyQX data set [10], which consists of speech files, their subjective rating scores from 21 subjects, and EEG signals from the same subjects recorded while they listened to the speech. The data recording protocol was approved by the INRS Research Ethics Office, and participants gave informed consent for their participation and to make their data anonymous and freely available online. The details of the data set and the experimental procedures are available in [10]. We obtained it by an e-mail request.

Speech stimuli

The speech stimuli presented to the subjects in the data set consist of speech collected from four humans and seven commercially available TTS systems. From each human and each TTS system, four English sentences were collected, whose durations ranged from 13 to 22 seconds. The 44 human and synthesized speeches were presented to each subject in random order.

Experimental procedure

The experiment's timeline is shown in Fig 1. A 15-second rest period was provided before each stimulus presentation. It is followed by a subjective rating period during which the subjects evaluated the speech to which they had just listened. The subjective rating scales used in this study are shown in Table 1 and include overall impression (MOS), valence (VAL), and arousal (ARL). MOS was evaluated with a 5-scale rating and the others with a 9-scale using self-assessment manikin [37].

EEG recording and preprocess. EEG data were recorded throughout the experiment with 64 scalp channels. The sampling rate was 512 Hz, which was down-sampled to 256 Hz. All the channels were placed on scalp according to the modified 10/20 system [38]. Some channels were removed from further analysis because they were noisy. A band-pass filter was

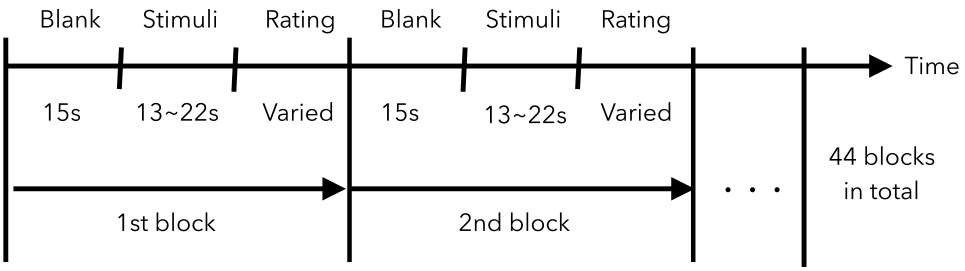

Fig 1. Timeline of EEG and subjective evaluation data recording experiment.

https://doi.org/10.1371/journal.pone.0193521.g001

Table 1. Subjective rating scales.

Rating Scale	Abbreviation	Description
Overall Impression	MOS	1 (Bad) to 5 (Excellent)
Valence	VAL	1 (Negative) to 9(Positive)
Arousal	ARL	1 (Unexcited) to 9 (Excited)

https://doi.org/10.1371/journal.pone.0193521.t001

applied to all the data between 0.5–50 Hz and applied an independent component analysis based semi-artifact removal technique using the ADJUST toolbox [39]. After these preprocessing, the EEG signal of each subject was cut into 44 epochs corresponding to the stimuli listening periods.

Methods

Feature extraction

All features were extracted at five frequency bands from a channel or a channel pair. The frequency bands include delta (δ: 1–4 Hz), theta (θ: 4–8 Hz), alpha (α: 8–12 Hz), beta (β: 12–30 Hz), and gamma (γ: 30–45 Hz). Let us denote the Fourier transformation at the frequency of f_k of the n-th trial recorded by the m-th channel by $\boldsymbol{x}_{n,m}(f_k)$. An estimator of the power spectrum density and a phase spectrum denoted by p_k and h_k can be calculated using the periodogram method as follows:

$$p_{n,m}(f_k) = \frac{1}{T}|\boldsymbol{x}_{n,m}(f_k)|^2 \tag{1}$$

$$h_{n,m}(f_k) = \text{angle}(\boldsymbol{x}_{n,m}(f_k)), \tag{2}$$

where T is the number of time samples within a trial. Then, we averaged the power spectrum density over the frequency bins within the range of each frequency band to define channel-based features $\text{PSD}_n(m,f)$ as follows:

$$\text{PSD}_n(m,f) = \frac{1}{|D_f|}\sum_{f_k \in D_f} p_{n,m}(f_k), \tag{3}$$

where D_f is the index set of the frequency bins included in the range of the f-th frequency band and $|D_f|$ is the number of the elements in D_f. The channel-pair-based features are also defined using the averaged power spectrum density and the phase spectrum as follows:

$$\text{PWD}_n(m_1,m_2,f) = \text{PSD}(m_1,f) - \text{PSD}_n(m_2,f) \tag{4}$$

$$\text{PHD}_n(m_1,m_2,f) = \frac{1}{|D_f|}\sum_{f_k \in D_f} h_{n,m_1}(f_k) - h_{n,m_2}(f_k). \tag{5}$$

If M EEG channels and F frequency bands are used ($F = 5$ in this study), $I = F(M(M-1) + M)$ features are calculated. The feature matrix X can be expressed as:

$$\text{X} = (\boldsymbol{x}(1), \boldsymbol{x}(2), \ldots, \boldsymbol{x}(N))^\top \in \mathbb{R}^{N \times I}, \tag{6}$$

where N is the number of training trials and $\boldsymbol{x}(n)$ is a feature vector of the n-th trial and has all the features PSD_n, PWD_n, and PHD_n.

To exploit structures of the features, we organized the features as a tensor $\mathcal{X} \in \mathbb{R}^{N \times M \times M \times F}$ as follows:

$$\mathcal{X}(n, m_1, m_2, f) = \begin{cases} \mathrm{PWD}_n(m_1, m_2, f) & (m_1 > m_2) \\ \mathrm{PHD}_n(m_1, m_2, f) & (m_1 < m_2) \\ \mathrm{PSD}_n(m_1, f) & (m_1 = m_2) \end{cases} \quad (7)$$

The feature matrix and tensor are depicted in Fig 2.

Regression analysis

Higher order partial least square (HOPLS) [34] and standard partial least square (PLS) [40, 41] simultaneously perform dimension reduction and regression, which were used in this study. The former is a natural extension of the latter so that tensor-format features can be used.

Let us denote the response matrix by Y that has all the response variables of all training trials:

$$\mathrm{Y} = (\boldsymbol{y}(1), \boldsymbol{y}(2), \ldots, \boldsymbol{y}(N))^{\top} \in \mathbb{R}^{N \times J}, \quad (8)$$

where $\boldsymbol{y}(n)$ is the $J = 3$ dimensional response vector of the n-th trial. All response variables were normalized to have zero mean and unit variance.

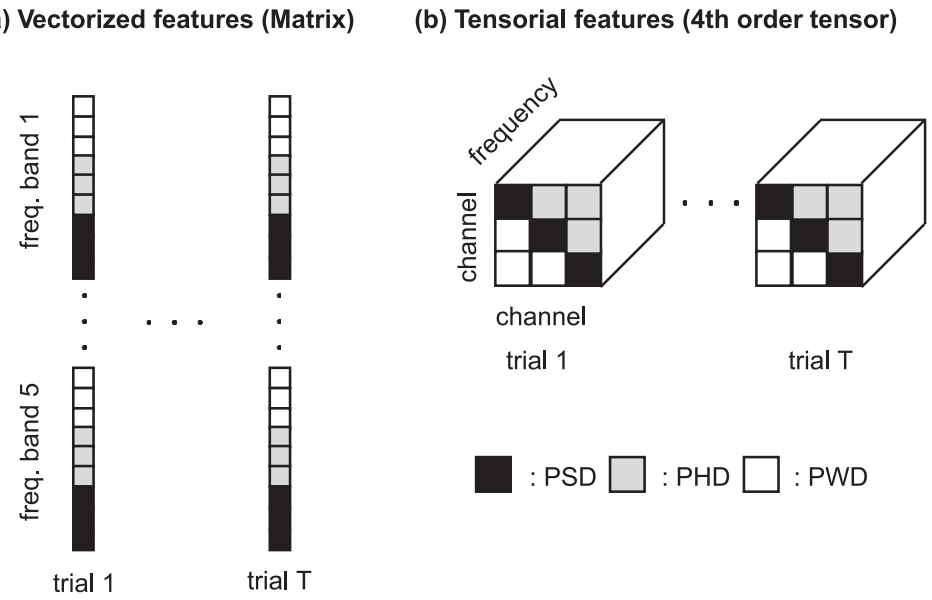

Fig 2. Schematic image of vectorized and tensorial features. (a) Vectorized feature is included in a matrix whose first and second modes are trials and features. PWD, PHD, and PSD at each frequency band are lined as a vector in each trial. (b) Tensor structured dependent variable with four modes: trials, channel-1, channel-2, and frequency bands. Tensor elements with a larger channel-1 index than channel-2 are PWD, and a smaller channel-1 index than channel-2 are PHD, and identical channel indexes are PSD.

https://doi.org/10.1371/journal.pone.0193521.g002

PLS performs a simultaneous decomposition of X and Y to find common latent variables $\boldsymbol{t}_r \in \mathbb{R}^N$ as:

$$\mathrm{X} = \sum_{r=1}^{R_1} \boldsymbol{t}_r \boldsymbol{p}_r^\top + \mathrm{E} \tag{9}$$

$$\mathrm{Y} = \sum_{r=1}^{R_1} \boldsymbol{t}_r \boldsymbol{q}_r^\top + \mathrm{F}, \tag{10}$$

where E and F are the residual matrices, and R_1 is called the number of the components.

On the other hand, HOPLS can be similarly formulated as the problem to find latent variables as follows:

$$\mathcal{X} = \sum_{r=1}^{R_2} \mathcal{G}_r \times_1 \boldsymbol{t}_r \times_2 \mathrm{P}_r^{(1)} \times_3 \mathrm{P}_r^{(2)} \times_4 \mathrm{P}_r^{(3)} + \mathcal{E} \tag{11}$$

$$\mathcal{G}_r \in \mathbb{R}^{1 \times M \times M \times F}, \quad \mathrm{P}_r^{(k)} \in \begin{cases} \mathbb{R}^{M \times L_k} & (k = 1, 2) \\ \mathbb{R}^{F \times L_k} & (k = 3) \end{cases} \tag{12}$$

$$\mathrm{Y} = \sum_{r=1}^{R_2} \boldsymbol{t}_r \boldsymbol{q}_r^\top + \mathrm{V}, \tag{13}$$

where $\mathcal{G}_r$ is called the core tensor, $\mathcal{E}$ and V are the residuals, R_2 is the number of the components, and $\times_k$ denotes the k-mode product [42]. $\mathrm{P}_r^{(n)}$ is called the loading matrix of the r-th component, and L_k is called the number of the k-mode loadings.

If data are plentiful, which is rare in EEG studies, the best approach for training and evaluating the performance of a predictive model is to randomly divide the dataset into three parts: training, validation, and test sets, which are respectively used to train a model, tune hyper-parameters or select a model, and evaluate the generalization error [43]. However, since the amount of data in this study is too small to exploit such an ideal protocol, we instead used leave-one-out cross-validation for each subject. The hyper-parameter R_1 of PLS varied from 1 to 43, loadings of the channel-1 L_1 and the channel-2 L_2 ranged from 1 to 7. The loading gs of the frequency band L_3 and the number of components R_2 of HOPLS ranged from 1 to 5. The result of the models that achieved the best performance was reported in Results.

Evaluation metrics

Root mean squared error (RMSE) was used to quantify the predictability of the regression models for each subject, which are formulated as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2}, \tag{14}$$

where N is the number of test samples, $\hat{y}_i$ is the predicted value for the i-th test data, and y_i is the actual value.

Table 2. Prediction results and the number of latent factors.

subject	RMSE (vector)			RMSE (tensor)			R_1	L_1	L_2	L_3	R_2
	MOS	VAL	ARL	MOS	VAL	ARL					
1	1.091	1.090	1.086	0.961	0.957	0.962	1	3	4	5	3
2	1.042	1.057	1.058	0.987	0.958	0.975	1	4	4	4	3
3	0.997	1.018	1.020	0.907	0.936	0.931	2	7	3	4	2
4	1.148	1.110	1.043	0.949	0.944	0.994	2	7	5	2	2
5	1.187	1.225	1.229	1.082	1.177	1.150	1	6	7	4	5
6	1.260	1.292	1.151	1.000	2.176	1.941	4	7	4	4	4
7	0.979	0.981	1.007	0.869	0.935	1.167	1	2	3	5	5
8	1.155	1.186	1.125	0.970	0.989	1.017	1	1	5	1	1
9	1.215	1.221	1.160	0.996	0.994	1.023	2	7	7	1	2
10	1.111	1.112	1.022	0.957	0.985	1.144	1	5	4	3	4
11	1.125	1.243	1.0.19	1.013	1.047	0.920	3	7	7	1	1
12	0.996	1.193	1.177	0.641	0.912	0.680	29	4	2	3	4
13	1.258	1.227	1.234	1.051	1.050	1.035	3	4	2	5	4
14	0.991	1.102	1.040	0.940	0.980	0.929	12	7	1	2	3
15	1.022	0.989	0.969	0.965	0.927	0.934	1	7	3	2	5
16	1.196	1.206	1.087	1.058	1.047	1.027	4	7	6	2	5
17	1.021	1.083	1.087	0.884	0.886	0.924	3	4	7	4	3
18	1.055	1.027	1.092	0.915	0.920	0.969	2	1	3	4	4
19	1.130	1.142	1.126	1.021	1.081	1.020	1	5	1	5	4
20	0.995	1.028	0.944	0.887	0.990	1.057	1	4	5	2	5
21	1.121	1.157	1.103	0.900	0.969	0.997	1	1	1	4	2

https://doi.org/10.1371/journal.pone.0193521.t002

Results

Table 2 summarizes RMSE, and the numbers of latent factors identified by PLS and HOPLS, respectively. Predictions with tensorial features generally made smaller errors than the vectorized ones for all the rating scales. Fig 3 reports the one hundred features that contributed to the prediction the most greatly, where feature contributions were calculated by taking the

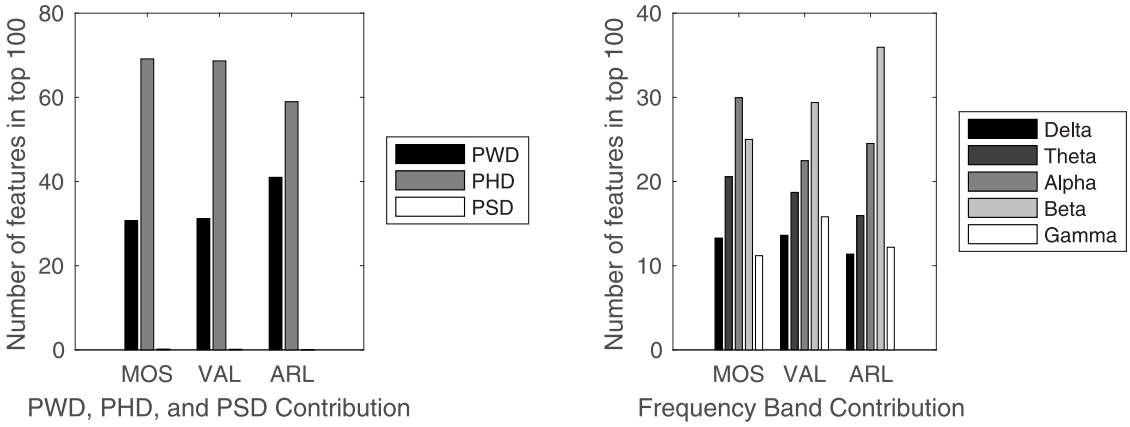

Fig 3. Contribution of features. Feature contributions were calculated by taking magnitude of their regression coefficients. (Left) Numbers of PWD, PHD, and PSD among the one hundred features that most greatly contributed to the prediction of each rating scale among all features. (Right) Number of features of each frequency band among the one hundred features that most greatly contributed to each rating scale.

https://doi.org/10.1371/journal.pone.0193521.g003

magnitude of the regression coefficients and PSD, PWD, and PHD are separately shown. PSD rarely appeared in the list of the top one hundred features list for all of the rating scales. Among the five frequency bands, the alpha band contributed the most to the MOS prediction, followed by the beta band. For the VAL and the ARL predictions, the beta band contributed the most, followed by the alpha band. The top ten channel pairs, which contributed the most to the MOS prediction extracted from subjects 1, 2, 3, and 4, are shown in Fig 4.

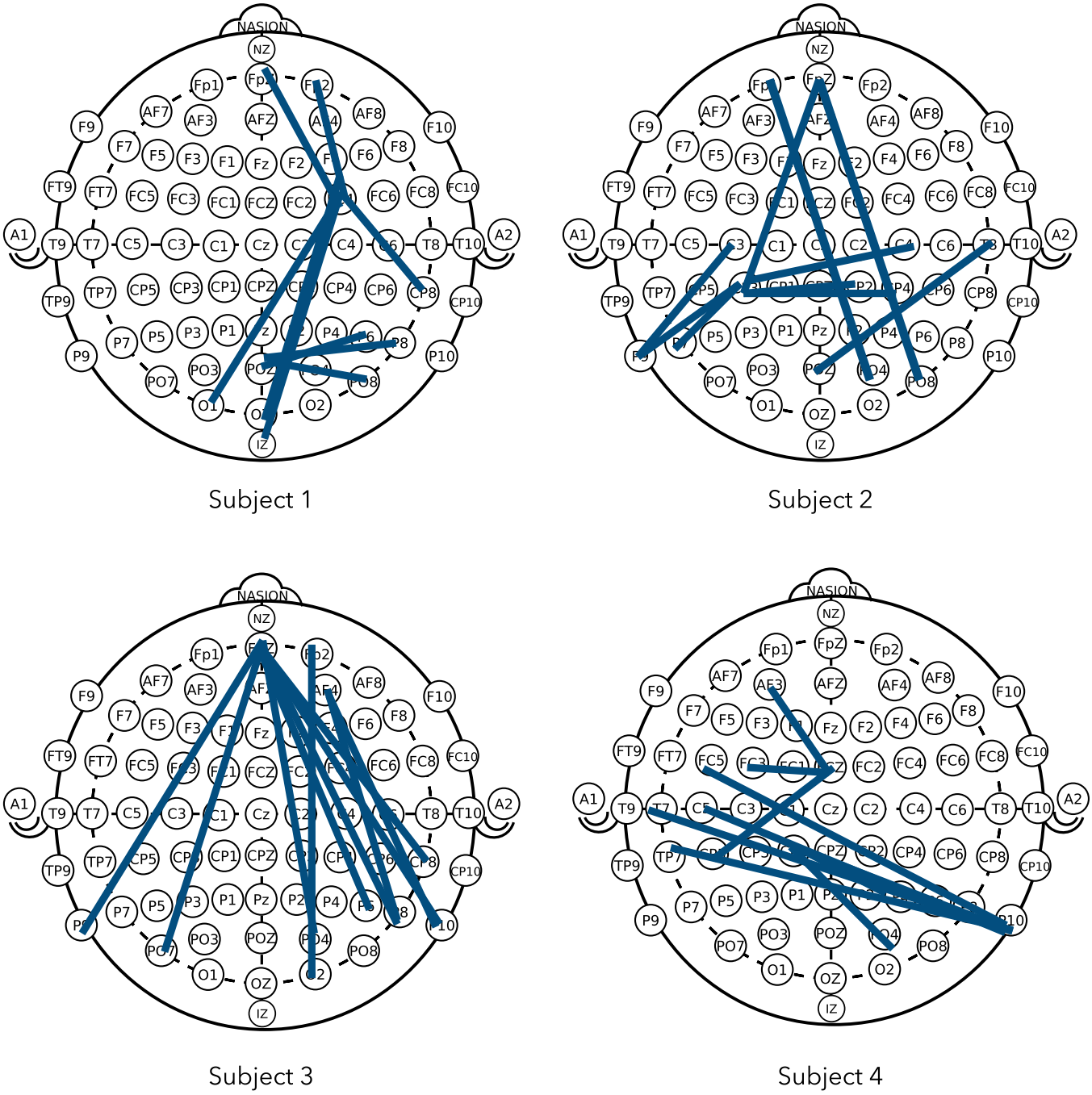

Fig 4. Contribution of channel pairs. Top ten channel pairs that contributed the most to MOS predictions of subjects 1, 2, 3, and 4. This figure was made by modifying the original one, which is distributed under the public domain dedication [44].

https://doi.org/10.1371/journal.pone.0193521.g004

Discussion

Channel-pair-based features (PWD and PHD) contributed more to the predictions than channel-based ones (PSD), which agrees with a previous study [31] and suggests the importance of considering scalp EEG dynamics between brain regions, and that graph theory based features or functional connectivity analysis can be effective [45, 46]. The importance of spectral differences in caudality (DCAU) between the anterior and posterior [12, 47] or the front-posterior brain regions [31] as well as the lateral (left-right) spectral difference (DLAT) have been documented [28, 30]. In this study, both of DLAT and DCAU contributed to the predictions (Fig 4) although their effectiveness was dependent on the subjects.

Quality prediction models were independently trained for each subject in this study because emotion regulation is reportedly dependent on individuals [48]. The commonality of the channels/channel pairs, which greatly contributed to the predictions, was actually rather small (Fig 4). Therefore, creating subject-independent features is an interesting future work. However, note that the alpha and beta bands commonly contributed to the predictions, whereas the effective channels/channel pairs differed depending on the subjects. The alpha and beta bands contributed more largely to the predictions than the other frequency bands, which is in line with previous neurophysiological studies. The relationship between alpha band asymmetry and the withdrawal or disengagement from a stimulus or negative valence has been well documented in response to a variety of stimuli, including pictures [49, 50], music [31, 47, 51], movies [52], and speech [11, 13]. The beta band, which contributed the most to the ARL predictions, is reportedly associated with arousal and emotional experiences [53, 54].

Gupta et al. [13] predicted MOS values using the same data set that we used in this study. Their study used a simple linear regression model with not only EEG but also speech features. They reported the RMSE of their model was 0.117, which is much lower than our model, and suggests that speech features are much more informative than EEG features to predict subjective quality ratings.

Although we predicted the response values of MOS, VAL, and ARL, other perpetual dimensions were also proposed recently to model emotions or perceived quality-of-experiences [55, 56], which should be investigated in future research.

Neither previous work nor our current study advocate that physiological assessment methods of speech quality should replace subjective rating methods or signal analysis methods because, as stated in Introduction, each method has its own advantages and disadvantages and they can complement each other.

Several open questions remain. First, features were extracted and constructed as tensors as described in Feature Extraction and Regression Analysis, but other features and construction ways are also possible. For example, if time-frequency analysis is employed, times frames can be treated as one of the tensor modes. Second, this study analyzed the overall quality of each speech stimulus longer than ten seconds. However, parts of speech can affect much more largely its overall perceived quality. Therefore, analysis methods to specify such parts need to be studied.

Conclusion

This study predicted the subjective quality ratings of synthesized speech solely based on EEG. We created vectorized and tensorial features for the regression that include channel-based and channel-pair-based features at multiple frequency bands. The experimental result showed that tensorial features more effectively predicted the subjective ratings than the other, and the trained predictive models were neurophysiologically plausible.

Author Contributions

Conceptualization: Hayato Maki, Sakriani Sakti, Hiroki Tanaka, Satoshi Nakamura.

Data curation: Hayato Maki.

Formal analysis: Hayato Maki.

Funding acquisition: Sakriani Sakti, Hiroki Tanaka, Satoshi Nakamura.

Investigation: Hayato Maki, Sakriani Sakti, Hiroki Tanaka.

Methodology: Hayato Maki, Hiroki Tanaka.

Project administration: Hayato Maki.

Software: Hayato Maki.

Supervision: Sakriani Sakti, Hiroki Tanaka, Satoshi Nakamura.

Validation: Hayato Maki, Sakriani Sakti.

Visualization: Hayato Maki.

Writing – original draft: Hayato Maki.

Writing – review & editing: Hayato Maki.

References

1. CCITT. Absolute category Rating (ACR) method for subjective testing of digital processors. Red Book. 1984.
2. ITU-T Recommendation P.85. A Method for Subjective Performance Assesment of the Quality of the Speech Voice Output Devices. International Telecommunication Union. 1996.
3. Benoît C, Grice M, Hazan V. The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. Speech Communication, 1996; 18(4): 381–392. https://doi.org/10.1016/0167-6393(96)00026-X
4. Mariniak A. A global framework for the assessment of synthetic speech without subjects. Proceedings of the third European Conference on Speech Communication and Technology. 1993.
5. Norrenbrock C, Hinterleitner F, Heute U, Möller S. Quality prediction of synthesized speech based on perceptual quality dimensions. Speech Communication. 2015; 66: 17–35. https://doi.org/10.1016/j.specom.2014.06.003
6. ITU-T Recommendation P.862. Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs. International Telecommunication Union. 2001.
7. Antons J N. Neural correlates of Quality Perception for Complex Speech Signals. Springer, 2015.
8. Antons J N, Schleicher R, Arndt S, Möller S, Porbadnigk A K, Curio G. Analyzing Speech Quality Perception Using Electroencephalography. IEEE Journal of Selected Topics in Signal Processing. 2012; 6(6): 721–731. https://doi.org/10.1109/JSTSP.2012.2191936
9. Antons J N, Blankertz B, Curio G, Möller S, Porbadnigk A K, Schleicher R. Subjective Listening Tests and Neural Correlates of Speech Degradation in Case of Signal-Correlated Noise. Audio Engineering Society Convention 129. 2010.
10. Gupta R, Banville H J, Falk T H. PhySyQX: A database for physiological evaluation of synthesised speech quality-of-experience. Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics. 2015;1–5.
11. Gupta R, Laghari K, Banville H, Falk T. H. Using affective brain-computer interfaces to characterize human influential factors for speech quality-of-experience perception modeling. Human-centric Computing and Information Sciences. 2016; 6(5).
12. Sarlo M, Buodo G, Poli S, Palomba D. Changes in EEG alpha power to different disgust elicitors: the specificity of mutilations. Neuroscience Letters. 2005; 382(3): 291–296. https://doi.org/10.1016/j.neulet.2005.03.037 PMID: 15925105

13. Arndt S, Antons J N, Gupta R, Schleicher R, Möller S, Falk T H. The effects of text-to-speech system quality on emotional states and frontal alpha band power. Proceeding of the sixth International IEEE Engineering in Medicine and Biology Society Conference on Neural Engineering.
14. Antons J N, Schleicher R, Arndt S, Möller S, Curio G. Too tired for calling? A physiological measure of fatigue caused by bandwidth limitations. Proceedings of the fourth International Conference on Quality of Multimedia Experience. 2017;63–67.
15. Asgari M, Kiss G, van Santen J, Shafran I, Song X. Automatic measurement of affective valence and arousal in speech. Proceedings of the 39th IEEE International Conference on Acoustics, Speech and Signal Processing. 2014.
16. Nicolaou M A, Gunes H, Pantic M. Continuous Prediction of Spontaneous Affect from Multiple Cues and Modalities in Valence-Arousal Space. IEEE Transactions on Affective Computing. 2011; 2(2):92–105. https://doi.org/10.1109/T-AFFC.2011.9
17. Russell J A, Mehrabian A. Evidence for a three-factor theory of emotions. Journal of Research in Personality. 1977; 11(3):273–294. https://doi.org/10.1016/0092-6566(77)90037-X
18. Russell J A. A circumplex model of affect. Journal of Personality and Social Psychology. 1980; 39 (6):1161–1178. https://doi.org/10.1037/h0077714
19. Russell J A, Barrett L F. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. Journal of Personality and Social Psychology. 1999; 76(5):805–819. https://doi.org/10.1037/0022-3514.76.5.805 PMID: 10353204
20. Russell J A. Core affect of and the phychological construction of emotion. Psychological Review. 2003; 110(1):145–172. https://doi.org/10.1037/0033-295X.110.1.145 PMID: 12529060
21. Mayo C, Clark R A, King S. Listeners' weighting of acoustic cues to synthetic speech naturalness: A multidimensional scaling analysis. Speech Communication. 2011; 53(3):311–326. https://doi.org/10.1016/j.specom.2010.10.003
22. Jekosch U. Voice and speech quality perception: assessment and evaluation. Springer Science & Business Media. 2006.
23. Möller S, Hinterleitner F, Falk T H, Polzehl T. Comparison of approaches for instrumentally predicting the quality of text-to-speech systems. Proceedings of eleventh annual conference of the International Speech Communication Association. 2010.
24. Arndt S, Brunnström K, Cheng E, Engelke U, Möller S, Antons J N. Review on using physiology in quality of experience. Electronic Imaging. 2016; 16: 1–9. https://doi.org/10.2352/ISSN.2470-1173.2016.16.HVEI-125
25. Mwangi B, Tian T S, Soares J C. A review of feature reduction techniques in neuroimaging. Neuroinformatics. 2014; 12(2):229–244. https://doi.org/10.1007/s12021-013-9204-3 PMID: 24013948
26. Luck S. An introduction to the event-related potential technique. MIT press. 2014.
27. Antons J N, Friedemann K, Arndt S, Möller S, Schleicher R. Changes of vigilance caused by varying bit rate conditions. Proceedings of the fifth International Conference on Quality of Multimedia Experience. 2013;148–151.
28. Davidson R J. Anterior cerebral asymmetry and the nature of emotion. Brain Cognition. 1992; 20:125–151. https://doi.org/10.1016/0278-2626(92)90065-T PMID: 1389117
29. Aftanas L I, Reva N V, Varlamov A A, Pavlov S V, Makhnev V P. Analysis of evoked EEG synchronization and desynchronization in conditions of emotional activation in humans: temporal and topographic characteristics. Neuroscience and Behavioral Physiology. 2004; 34(8):859–867. https://doi.org/10.1023/B:NEAB.0000038139.39812.eb PMID: 15587817
30. Coan J A, Allen J J. Frontal EEG asymmetry as a moderator and mediator of emotion. Biological Psychology. 2004; 67(1): 7–50. https://doi.org/10.1016/j.biopsycho.2004.03.002 PMID: 15130524
31. Lin Y P, Wang C H, Jung T P, Wu T L, Jeng S K, Duann J R, Chen J H. EEG-based emotion recognition in music listening. IEEE Transactions on Biomedical Engineering. 2010; 57(7): 1798–1806. https://doi.org/10.1109/TBME.2010.2048568 PMID: 20442037
32. Lee Y-Y, Hsieh S. Classifying Different Emotional States by Means of EEG-Based Functional Connectivity Patterns. PLoS ONE 2014; 9(4): e95415. https://doi.org/10.1371/journal.pone.0095415 PMID: 24743695
33. Costa T, Rognoni E, Galati D. EEG phase synchronization during emotional response to positive and negative film stimuli. Neuroscience Letters. 2006; 406(3): 159–164. https://doi.org/10.1016/j.neulet.2006.06.039 PMID: 16942838
34. Zhao Q, Caiafa C F, Mandic D P, Chao Z C, Nagasaka Y, Fujii N, Zhang L, Cichocki A. Higher Order Partial Least Squares (HOPLS): A Generalized Multilinear Regression Method. IEEE Transactions on

Pattern Analysis and Machine Intelligence. 2012; 35(7): 1660–1673. https://doi.org/10.1109/TPAMI.2012.254

35. Rabusseau G, Kadri H. Low-Rank Regression with Tensor Responses. Proceedings of the Advances in Neural Information Processing Systems. 2016;1867–1875.
36. Zhou H, Li L, Zhu H. Tensor regression with applications in neuroimaging data analysis. Journal of the American Statistical Association. 2013; 108(502): 540–552. https://doi.org/10.1080/01621459.2013.776499 PMID: 24791032
37. Bradley M M, Lang P J. Measuring emotion: the self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry. 1994.; 25(1): 49–59. https://doi.org/10.1016/0005-7916(94)90063-9 PMID: 7962581
38. American Clinical Neurophysiology Society. American Electroencephalographic Society Guidelines for Standard Electorode Position Nomenclature. Journal of Clinical Neurophysiology. 1996; 8(2): 200–202.
39. Mognon A, Jovicich J, Bruzzone L, Buiatti M. ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features. Psychophysiology. 2011; 48(2):229–240. https://doi.org/10.1111/j.1469-8986.2010.01061.x PMID: 20636297
40. World S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems. 2001; 58(2): 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1
41. Zhao Q, Zhang L, Cichocki A. Multilinear and nonlinear generalizations of partial least squares: an overview of recent advances. Wiley Interdisciplinary Reviews:f Data Mining and Knowledge Discovery. 2014; 4(2): 104–115.
42. Kolda T G, Bader B W. Tensor Decompositions and Applications. SIAM Review. 2009; 51(3):455–500. https://doi.org/10.1137/07070111X
43. Friedman J, Hastie T, Tibshirani R. The Elements of Statistical Learning. Springer Series in Statistics. 2001.
44. https://commons.wikimedia.org/wiki/File:International_10-20_system_for_EEG-MCN.svg
45. Gupta R, Falk T H. Relevance vector classifier decision fusion and EEG graph-theoretic features for automatic affective state characterization. Neurocomputing 2016; 174: 875–884. https://doi.org/10.1016/j.neucom.2015.09.085
46. Srinivasan R, Winter W R, Ding J, Nunez P L. EEG and MEG coherence: measures of functional connectivity at distinct spatial scales of neocortical dynamics. Journal of Neuroscience Methods. 2007; 166(1): 41–52. https://doi.org/10.1016/j.jneumeth.2007.06.026 PMID: 17698205
47. Schmidt L A, Trainor L J. Frontal brain electrical activity (EEG) distinguishes valence and intensity of musical emotions. Cognition and Emotion. 2001; 15(4): 487–500. https://doi.org/10.1080/02699930126048
48. Gross J J, John O P. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. Journal of Personality and Social Psychology. 2003; 85(2): 348–362. https://doi.org/10.1037/0022-3514.85.2.348 PMID: 12916575
49. Huster R J, Stevens S, Gerlach A L, Rist F. A spectralanalytic approach to emotional responses evoked through picture presentation. International Journal of Psychophysiology. 2009; 72(2): 212–216. https://doi.org/10.1016/j.ijpsycho.2008.12.009 PMID: 19135486
50. Balconi M, Guido M. Lateralisation effect in comprehension of emotional facial expression: a comparison between EEG alpha band power and behavioural inhibition (BIS) and activation (BAS) systems. Laterality: Asymmetries of Body, Brain and Cognition. 2010; 15(3): 361–384. https://doi.org/10.1080/13576500902886056
51. Altenmüller E, Schürmann K, Lim V K, Parlitz D. Trainorsic are reflected in cortical lateralisation patterns. Neuropsychologia. 2002; 40(13): 2242–2256.
52. Jones N A, Nathan A F. Electroencephalogram asymmetry during emotionally evocative films and its relation to positive and negative affectivity. Brain and Cognition. 1992; 20(2): 280–299. https://doi.org/10.1016/0278-2626(92)90021-D PMID: 1449758
53. Phan K L, Taylor S F, Welsh R C, Decker L R, Noll D C, Nichols T E, Liberzon I. Activation of the medial prefrontal cortex and extended amygdala by individual ratings of emotional arousal: a fMRI study. Biological Psychiatry. 2003; 53(3): 211–215. https://doi.org/10.1016/S0006-3223(02)01485-3 PMID: 12559653
54. Dan Glauser E S, Scherer K R. Neuronal processes involved in subjective feeling emergence: Oscillatory activity during an emotional monitoring task. Brain Topography. 2008; 20(4): 224–231. https://doi.org/10.1007/s10548-008-0048-3 PMID: 18340523

55. Rubin D C, Talarico J M. A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. Memory, 17(8), 802-808. Journal of Behavior Therapy and Experimental Psychiatry. 1994. 25.1: 49-59. Memory. 2009; 17(1): 49–59.

56. Gupta R, Falk T H, "Latent factor analysis for synthesized speech quality-of-experience assessment", Quality and User Experience. 2017; 2(1): https://doi.org/10.1007/s41233-017-0005-6