Continuous Speech Prediction by Segmentation of Auditory EEG

Tomoaki Mizuno
Graduate School of Informatics and Engineering
The University of Electro-Communications
Tokyo, Japan
t.mizuno2301@uec.ac.jp

Natsue Yoshimura
School of Computing
Institute of Science Tokyo
Kanagawa, Japan

Toru Nakashika
Graduate School of Informatics and Engineering
The University of Electro-Communications
Tokyo, Japan

***Abstract*—Synthesizing full speech from electroencephalography (EEG) signals is a challenging task. In this study, we investigate the reconstruction of continuous speech that a listener has perceived from non-invasive EEG using a Transformer-based deep learning model. Rather than utilizing the entire brain signal, we examine which brain regions yield more accurate extraction of speech features related to the perceived speech. To determine the localization of underlying information related to voice characteristic differences and phoneme prediction from the auditory EEG, electrodes are segmented by region, and speech features are inferred for each. The results imply that selective electrode placement may enable the concurrent extraction of speaker-identifying and linguistic information.**

***Index Terms*—Non-invasive electroencephalography (EEG) , brain-machine interface (BMI) , speech synthesis, speech recognition, Transformer**

I. Introduction

Brain machine interface (BMI) research encompasses the development of technologies that enable the control of computer systems solely through thought. A notable subset of this research involves the extraction of linguistic information from neural activity and its subsequent synthesis into speech. In this study, our primary focus is on the reconstruction of continuous speech that a listener has perceived from non-invasive EEG. The potential of this approach lies in its promise to provide a means of vocal communication for individuals who have lost the ability to speak due to medical conditions or other impairments. Specifically, this study focuses on reconstructing speech that a listener has perceived from their non-invasive EEG signals. This line of research is crucial as it not only helps us understand the neural correlates of speech perception but also lays foundational groundwork for future advancements, such as neural hearing aids that could provide real-time feedback of perceived sounds, or silent speech brain-computer interfaces aimed at vocalizing internal thoughts or unvoiced utterances for individuals with communication impairments.

BMI are generally categorized into two types, invasive and non invasive based on the techniques used to record neural activity. Electrocorticography (ECoG) is an invasive modality that entails positioning electrodes on the surface of the brain, and it has been utilized to extract neural signals for high-precision speech synthesis [1], [2]. Previous studies have demonstrated the synthesis of intelligible speech by combining neural activity with articulatory movements; however, complete speech synthesis using solely neural signals, without incorporating mouth movement information, has not yet been achieved. Moreover, the surgical requirements for electrode implantation in invasive techniques contribute to their higher cost.

Conversely, electroencephalography (EEG) is a non-invasive approach that records electrical signals via electrodes placed on the scalp, offering a more convenient and less risky means of monitoring brain activity. Nevertheless, synthesizing comprehensive speech sounds from EEG data remains more challenging than using ECoG. Despite these challenges, recent research has successfully synthesized two short vowels with high intelligibility from EEG recordings by employing a simple neural network designed to infer phonetic features [3]. This suggests that more complex network architectures may improve the performance of EEG-based speech synthesis.

We hypothesized that by utilizing the Transformer [4] model, which is known as a powerful tool for processing sequential signals, we could extract more complex information from EEG recordings obtained during continuous speech perception [5]. To examine this hypothesis, we conducted experiments in which we attempted to convert EEG signals into speech using a Transformer-based model. Our findings revealed that the EEG data indeed contained speaker-specific information, since it was possible to synthesize speech that preserved the original voice characteristics from EEG signals corresponding to two different speakers. However, the extraction of linguistic information, such as phoneme sequences and textual representations, proved to be challenging. Although one advantage of using EEG is the ability to capture whole-brain data, this may also lead to an overabundance of information. In light of this, we posited that narrowing the focus to specific EEG regions could potentially yield more accurate information extraction. Additionally, one of the specific objectives of this study was to examine whether there exists neural specificity in brain regions that distinguish differences in speech quality. In this study, we divided the EEG

This work was supported by Hirokazu Matsuura, Kaya Chin, and Keito Nakamori of Institute of Science Tokyo for measuring and providing EEG data for the execution of this study. This research was supported by JSPS Grant-in-Aid for Scientific Research 24H00715. This work was supported by JST FOREST Program (Grant Number JPMJFR216W).

electrodes into distinct regions and predicted speech features from each region. This approach enabled us to investigate which electrodes retain the most critical information when inferring speech features from EEG data.

II. Related work

In our previous work, we adopted the voice transformer network (VTN) [6] based on the Transformer architecture [4] to reconstruct speech from EEG signals recorded during auditory perception [5]. Given that voice conversion techniques transform the speech of one speaker into that of another while preserving linguistic content and speaking style, we anticipated that such a model would be effective for handling time-series signals like EEG. This method employs a Transformer-based architecture that comprises an encoder and a decoder to map an input EEG sequence $\boldsymbol{x}_{1:n} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \in \mathbb{R}^{d_x \times n}$ to a target speech sequence $\boldsymbol{y}_{1:m} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m) \in \mathbb{R}^{d_y \times m}$. Here, d_x and d_y denote the dimensions of the respective features. Note that the EEG sequence has a length of n and the target speech sequence has a length of m, and these lengths are not necessarily identical.

Additionally, similar to previous studies [5], We used guided attention [9] to reduce the risk of using brain wave information from periods when the subject was not listening during the speech reconstruction. The guided attention can make the attention matrix A, whose rows are attention weights $\boldsymbol{a}_t$, be approximately diagonal. In this approach, it is applied through diagonal weighting that penalizes deviations from the expected temporal correspondence. By imposing this constraint, the model is encouraged to associate EEG signals only with the vicinity of the speech segments that temporally align with the listener's auditory experience. This strategy effectively mitigates the risk of incorporating irrelevant brain wave information from periods when the subject was not actively listening, thereby preventing extreme contradictory associations (e.g., pairing the last second of the EEG with the first second of the audio). The loss function of guided attention is defined as follows:

$$L_{att}(A) = \mathbb{E}_{pt}[A_{pt} W_{pt}], \tag{1}$$

with

$$W_{pt} = 1 - \exp\left(-\frac{(\frac{p}{P} - \frac{t}{T})^2}{2g^2}\right). \tag{2}$$

In this context, let p denote the position of a character or word in the text, with P representing the total number of positions. Similarly, let t be the index of a time step in the audio sequence, and T its total duration. The element A_{pt} in the attention matrix A quantifies the correspondence between the text at position p and the audio at time t; higher values indicate a stronger correlation, while lower values indicate a weaker one. Moreover, W_{pt} is an entry in the weight matrix employed to compute the guided attention loss, as specified in Eq. (2). The parameter g governs the variance of the associated Gaussian function, thereby modulating the penalty based on the distance between text position p and time step t.

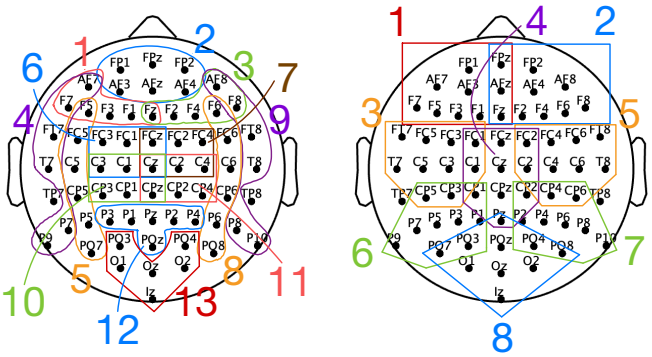

Fig. 1. Left: EEG divided into 6 electrodes, Right: EEG divided into 10 electrodes.

III. Speech prediction by segmentation of EEG electrode sets

To investigate which EEG electrode signals contain the most critical information for inferring speech features, we partitioned whole-brain auditory EEG data into groups of 6 or 10 electrodes. These subsets were paired with corresponding speech signals, and a VTN model was trained accordingly. An objective evaluation was then conducted on the speech generated from the EEG data.

A. EEG measurements and dataset preparation

In this study, we followed the methodology of our previous work by recording and processing EEG signals during speech perception, thereby creating a dataset of EEG responses to auditory stimuli. Speech materials were selected from the ATR Phoneme Balanced 503-sentence corpus [10], comprising utterances by both a male speaker (MMY) and a female speaker (FTK). To create the EEG and speech dataset, utterances from both speakers were presented to a single experimental participant at one-second intervals, and the corresponding EEG responses were recorded using the Biosemi ActiveTwo system (Biosemi). Triggers marking the onset of speech for each sentence were also recorded. The EEG signals were then sampled at a rate of 8192 Hz, band-pass filtered (1–40 Hz), and downsampled to 256 Hz. Independent component analysis (ICA) was employed to remove ocular artifacts. However, due to the presence of residual artifacts in TP8, this electrode was discarded, and TP7 was also omitted to maintain bilateral symmetry, resulting in a total of 62 electrodes. For each of the 503 sentences, the EEG data were segmented from the onset trigger to one second after speech ended and standardized per electrode. Subsequently, electrodes were grouped into regions comprising 6 or 10 electrodes. The electrodes divisions are illustrated in Figure 1 and the electrodes in each region are listed in Tables 1 and 2.

Hereinafter, the regions and electrode sets are referred to as Set 1, Set 2, etc. For the training set, speech signals were resampled to 16 kHz and subsequently transformed into log-amplitude Mel-spectrograms using an FFT with a size of 1024, a hop size of 256, and a Hann window. The Mel filter bank consisted of 80 filters spanning the frequency range from

TABLE I
SEGMENTED ELECTRODES REGION (6 ELECTRODES)

Number	electrodes List
1	Fz, F1, F3, F5, F7, AF7
2	Fpz, Fp1, Fp2, AFz, AF3, AF4
3	Fz, F2, F4, F6, F8, AF8
4	AF7, F7, FT7, T7, P7, P9
5	F5, FC5, C5, CP5, P5, PO7
6	FCz, FC1, FC3, Cz, C1, C3
7	FCz, FC2, FC4, Cz, C2, C4
8	F6, FC6, C6, CP6, P6, PO8
9	AF8, F8, FT8, T8, P8, P10
10	CPz, CP1, CP3, Pz, P1, P3
11	CPz, CP2, CP4, Pz, P2, P4
12	Pz, P1, P2, P3, P4, P5, POz
13	PO3, PO4, Oz, O1, O2, Iz

TABLE II
SEGMENTED ELECTRODES REGION (10 ELECTRODES)

Number	electrodes List
1	Fpz, Fp1, AFz, AF3, AF7, Fz, F1, F3, F5, F7
2	Fpz, Fp2, AFz, AF4, AF8, Fz, F2, F4, F6, F8
3	FC1, FC3, FC5, FT7, C1, C3, C5, C7, CP3, CP5
4	FCz, FC1, FC2, Cz, C1, C2, CPz, CP1, CP2, Pz
5	FC2, FC4, FC6, FT8, C2, C4, C6, C8, CP4, CP6
6	CP1, CP3, CP5, P1, P3, P5, P7, P9, PO3, PO7
7	CP2, CP4, CP6, P2, P4, P6, P8, P10, PO4, PO8
8	Pz, POz, PO3, PO4, PO7, PO8, Oz, O1, O2, Iz

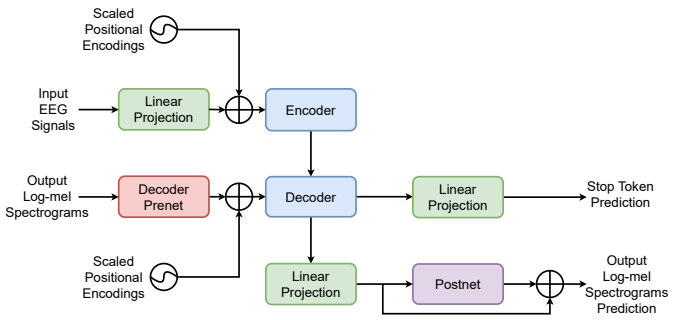

Fig. 2. Architecture of the Voice Transformer Network with EEG as input.

80 Hz to 7,600 Hz. The complete dataset was divided into training, validation, and test sets in a ratio of 480:12:11. Since separate datasets were constructed for the MMY and FTK recordings and then integrated, the final dataset comprised 960 training sentences, 24 validation sentences, and 22 test sentences.

B. Model detail

By applying VTN's architecture to EEG signal processing, we aimed to synthesize continuous speech from EEG signals. The model that extracts speech features from EEG data is illustrated in Figure 2. This architecture is a combination of the Transformer framework and Tacotron2 [11]. The Transformer, originally designed for machine translation, has been adapted for speech synthesis. In this configuration, the PreNet from Tacotron2 is added to the decoder. The PreNet, composed of multiple fully connected layers, dropout, and linear transformations, serves as a preprocessing network that converts the input mel spectrograms into dimensions more suitable for the decoder. In addition, position embeddings with learnable weights to adapt to the scales of the text and acoustic feature spaces, a linear projection to predict the output acoustic features, a linear projection to predict the Stop Token, and a PostNet consisting of a 5-layer CNN to predict the residual are incorporated. With the exception of the introduction of guided attention loss, the hyperparameters—including the learning rate and model architecture details (e.g., the number of layers and hidden dimensions)—were retained almost unchanged from the original VTN[1]. The input dimension and convolution kernel were changed in order to use EEG as input. Originally, the model was designed for 80 dimensional input, and the convolution for temporal subsampling in the encoder was designed to use a (3,3) kernel twice in the time direction. However, since this convolution is not possible when the input is 6 electrodes, the kernel size of the convolution was changed to (1,3). During training, we employed an overall loss function that integrates the L1-based reconstruction loss L_{L1} and the binary cross-entropy loss L_{BCE}—the two losses originally utilized in VTN—along with the previously described guided attention loss L_{att} [9]. This loss function is defined as follows:

$$L = L_{L1} + \alpha L_{BCE} + \beta L_{att}. \tag{3}$$

Here, α and β denote the weights applied to the binary cross-entropy loss and the guided attention loss, respectively. In our experiments, we set $\alpha = 10$ and $\beta = 1$, and we set the hyperparameter g in L_{att} to 0.4.

For Parallel WaveGAN [12] used as a vocoder, a two-speaker model of MMY and FTK was trained and used for the experiment. The sampling rate was 16 kHz, and the default parameters of Parallel WaveGAN [2] were used otherwise.

IV. RESULT AND DISCUSSION

A. Validation Items

In order to evaluate the naturalness of the synthesized speech and the phonemic content contained therein, objective evaluation metrics were applied to the speech generated from the test dataset. The following metrics were employed:

- PER (Phoneme Error Rate)
- Phoneme set matching rate (Dice coefficient, hereafter abbreviated as PSMR)
- SVM-based speech classification accuracy (hereafter abbreviated as SVM)
- UTMOS (UTokyo-SaruLab MOS prediction system) [13]

PER is intended to measure the preservation of linguistic information by quantifying phoneme errors, where a lower value indicates fewer errors. Let S denote the number of phoneme substitutions, D the number of phoneme deletions, I the number of phoneme insertions, and N the total number of phonemes in the reference transcript. S, D, I are computed by

[1]https://github.com/unilight/seq2seq-vc/tree/main/egs/arctic/vc1
[2]https://github.com/kan-bayashi/ParallelWaveGAN

counting the number of substitutions, deletions, and insertions in the transcription texts of the synthesized speech and the corresponding ground truth speech. PER is then calculated as follows:

$$\text{PER} = \frac{S + D + I}{N}. \tag{4}$$

PER is used as a metric to assess the extent to which the ground truth speech retains linguistic information.

In contrast, PSMR is utilized to evaluate the matching rate of the phoneme sequence. By disregarding the temporal order, this measure assesses whether the phonemes have been correctly reconstructed on a phoneme-by-phoneme basis, addressing issues observed in previous experiments where generated speech failed to retain linguistic information and temporal alignment. To compute this, we employed an acoustic model fine-tuned from Wav2Vec 2.0 [14]—which outputs Japanese hiragana for speech recognition[3]—to convert the speech into hiragana, and subsequently into phonemes for evaluation. PSMR was computed by extracting the phonemes from both the synthesized and ground truth speech (using the same method as for PER), removing duplicates, and forming sets without considering their sequential order. The matching rate is then determined using the Dice coefficient:

$$\text{Dice} = \frac{2 \times |\mathbf{X} \cap \mathbf{Y}|}{|\mathbf{X}| + |\mathbf{Y}|}, \tag{5}$$

where $\mathbf{X}$ and $\mathbf{Y}$ denote the two sets being compared, $|\mathbf{X} \cap \mathbf{Y}|$ is the number of common elements between sets $\mathbf{X}$ and $\mathbf{Y}$, and $|\mathbf{X}|$, $|\mathbf{Y}|$ represent the number of elements in each set, respectively. A higher Dice coefficient is expected when the correct phonemes are captured. Evaluating both PER and PSMR simultaneously provides a comprehensive assessment of the linguistic fidelity in synthesized speech—if both metrics yield favorable outcomes, namely a high PSMR combined with a low PER, it confirms that the synthesized speech not only captures the complete set of phonemes but also preserves their proper temporal order, thereby ensuring a high degree of linguistic fidelity and naturalness.

The SVM was employed as a metric to evaluate whether the synthesized speech accurately reproduced the voice of one of the two speakers. The mel-frequency cepstral coefficients (MFCC) of the correct speech were computed using 20-frame segmentation, and a speaker identification model (with an accuracy of 94%) was trained for this purpose. This SVM model was then used to classify the synthesized speech as either male or female, and its classification accuracy was assessed.

UTMOS serves as a metric for evaluating speech quality, analogous to MOS, by assigning scores on a 5-point scale (from 1 to 5), with values closer to 5 indicating higher speech quality. Moreover, as UTMOS has been trained on speech data of varying quality, it is utilized to assess the overall impressions of naturalness and clarity of the synthesized speech.

[3]https://huggingface.co/vumichien/wav2vec2-large-xlsr-japanese-hiragana

Furthermore, to comprehensively evaluate these metrics, we introduce the **Score** for each model defined in Eq. (6). Note that, since lower PER values indicate better performance, the PER metric was transformed by subtracting it from 1 prior to applying. UTMOS was processed to have a maximum value of 1 and a minimum value of 0 in the calculation.

$$\text{Score} = \frac{(1 - \text{PER}) + \text{PSMR} + \text{SVM} + \frac{\text{UTMOS}-1}{4}}{4} \tag{6}$$

B. Effects of electrode selection

Tables III and IV present the evaluations of PER, PSMR, SVM, UTMOS, and Score for the generated speech of the models trained using the EEG data from the respective electrode sets.

Regarding PSMR and PER, although values varied depending on the electrode set, statistical tests revealed no significant differences in any electrode area when compared to the full-electrode configuration. This lack of statistical significance is likely attributable to the large variance within the data and the limited test dataset of 22 items. Nevertheless, even without reaching statistical significance, numerical trends towards improvement were observed in some electrode sets. For example, in the 10ch configuration, PER for Set 4 (p-value 0.315) and Set 5 (p-value 0.299), and in the 6ch configuration, PSMR for Set 2 (p-value 0.239) and Set 9 (p-value 0.064), as well as PER for Set 5 (p-value 0.254) and Set 9 (p-value 0.2739), showed lower p-values compared to the full-electrode configuration, suggesting a trend towards improvement. In specific regions, PER showed a numerical tendency to improve when compared to the full-electrode configuration. However, PSMR did not show a similar improvement trend, particularly in the 10ch configuration. This observation can be interpreted as suggesting an ability to approximate the ground truth with higher precision in terms of the number of phonemes extracted from the predicted speech. Conversely, in Set 9 of the 6ch configuration, PER showed an improvement trend while PSMR showed a decline, with the p-values indicating a relatively strong tendency for these trends. This result suggests that while there might be challenges in the variety or comprehensiveness of phonemes generated by the model (indicated by decreased PSMR), the sequencing of the generated phonemes (indicated by decreased PER) is relatively accurate, demonstrating a limited precision. Specifically, this can be interpreted as the model attempting to reliably extract and generate only certain, for example, more recognizable phonemes from the EEG signals, thereby avoiding the generation of uncertain phonemes and consequently keeping PER (phoneme substitutions, deletions, and insertions) low. This suggests a state where the model prioritizes precision, potentially at the expense of comprehensiveness. Further comparison with SVM classification accuracy revealed that electrode sets with higher SVM accuracy tended to have lower PER values. Notably, for both the 6-electrode and 10-electrode configurations, the top three electrode sets in terms of SVM accuracy consistently demonstrated numerically better PER performance compared to the full-electrode configuration. UTMOS values generally

TABLE III
COMPARISON OF ELECTRODE SET PERFORMANCE (6 ELECTRODES).

No.	PER(↓)	PSMR(↑)	SVM(↑)	UTMOS(↑)	Score(↑)
All	0.596	0.732	0.832	1.31	0.511
1	0.615	0.736	0.660	1.30	0.464
2	0.574	0.702	0.742	1.29	0.485
3	0.591	0.737	0.758	1.31	0.496
4	0.591	0.745	0.695	1.30	0.481
5	0.560	0.738	0.671	1.29	0.481
6	0.591	0.724	0.636	1.30	0.461
7	0.593	0.723	0.615	1.30	0.455
8	0.590	0.722	0.729	1.32	0.485
9	0.563	0.683	0.744	1.30	0.485
10	0.575	0.739	0.571	1.29	0.452
11	0.597	0.711	0.672	1.29	0.465
12	0.579	0.730	0.669	1.30	0.474
13	0.590	0.729	0.636	1.30	0.463

TABLE IV
COMPARISON OF ELECTRODE SET PERFORMANCE (10 ELECTRODES).

No.	PER(↓)	PSMR(↑)	SVM(↑)	UTMOS(↑)	Score(↑)
All	0.588	0.733	0.876	1.31	0.524
1	0.576	0.733	0.864	1.31	0.525
2	0.601	0.719	0.776	1.31	0.493
3	0.569	0.727	0.818	1.29	0.512
4	0.561	0.725	0.821	1.33	0.517
5	0.563	0.734	0.831	1.31	0.520
6	0.569	0.734	0.758	1.28	0.498
7	0.592	0.735	0.692	1.29	0.477
8	0.578	0.728	0.773	1.30	0.500

tended to be lower, likely due to the diminished linguistic content in the generated speech and the tendency of its acoustic features to differ from natural human speech.

Additionally, the Score metric varied among electrode sets; while the full-brain configuration achieved the highest overall score, specific electrode sets (namely, sets 2, 3, 8, and 9 in the 6-electrodes configuration and sets 1 and 5 in the 10-electrodes configuration) also obtained high scores. This spatial pattern suggests that electrodes positioned in the central frontal region (e.g., FPz, AFz, Fz) and the right lateral temporal region (e.g., FC6, C6, CP6, FT8, T8) likely carry critical information for both speaker identification and the extraction of linguistic features. These results imply that it is possible to extract both speaker identification and language information at the same time with the choice of electrodes.

V. CONCLUSION

The present study indicates that the strategic selection of electrode sets has the potential to facilitate the extraction of speaker-specific and linguistic information in a simultaneous manner. Notably, this study observed a tendency towards numerically improved results for specific electrode configurations across evaluation metrics such as PER and PSMR. Subsequent studies will integrate the weighted signal combinations to further optimize speech production and maintain speaker identity while enriching the linguistic content of the synthesized speech. Additionally, since our electrode selection approach was broadly partitioned across brain regions, developing precise electrode combination methods remains an important future challenge. One limitation of EEG data is its relatively low spatial resolution; however, improving spatial resolution with the effect of electrode selection would enable more precise analysis of localized brain regions. Consequently, we aim to explore methodologies that integrate strategies for enhancing EEG spatial resolution with our prevailing Transformer-based approach.

REFERENCES

[1] Anumanchipalli, G. K., et al. (2019). Speech synthesis from neural decoding of spoken sentences. Nature, 568, 493–498.

[2] Metzger, J. C., et al. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. Nature, 620, 1037–1046.

[3] Akashi, W., et al. (2021). Vowel sound synthesis from electroencephalography during listening and recalling. Advanced Intelligent Systems, 2000164, 1-9.

[4] Zhang, L., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

[5] Mizuno, T., et al. (2024). An investigation on the speech recovery from EEG signals using transformer. In APSIPA ASC 2024, 1–6.

[6] Huang, W.-C., et al. (2020). Voice Transformer Network: Sequence-to-Sequence Voice Conversion Using Transformer with Text-to-Speech Pretraining. In Proc. Interspeech, 4676–4680.

[7] Bahdanau, D., et al. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[8] Luong, T., et al. (2015). Effective approaches to attention-based neural machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, Sep. 2015, 1412–1421.

[9] Tachibana, H., et al. (2018). Efficiently Trainable Text-to-Speech System Based on Deep Convolutional Networks with Guided Attention. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4784–4788.

[10] Akira, K., et al. (1990). ATR Japanese speech database as a tool of speech recognition and synthesis. Speech Communication, 9(4), 357–363.

[11] Shen, J., et al. (2018). Natural TTS Synthesis by Conditioning WaveNet on MEL Spectrogram Predictions. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 4779–4783.

[12] Yamamoto, R., et al. (2019). Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. arXiv preprint arXiv:1910.11480.

[13] Saeki, T., et al. (2022). UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022. In Proceedings of Interspeech 2022, 386–390.

[14] Baevski, A., et al. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems, 33, 12449–12460.