Speech Synthesis from Electrocorticogram During Imagined Speech Using a Transformer-Based Decoder and Pretrained Vocoder

Shuji Komeiji[1], Kai Shigemi[1], Takumi Mitsuhashi[2], Yasushi Iimura[2], Hiroharu Suzuki[2], Hidenori Sugano [2], Koichi Shinoda[3], Kohei Yatabe[1], and Toshihisa Tanaka[1]

[1] Department of Electronic and Information Engineering, Tokyo University of Agriculture and Technology, 2-24-16 Naka-cho, Koganei-shi, Tokyo, Japan
[2] Department of Neurosurgery, Juntendo University School of Medicine, 2-1-1 Hongo, Bunkyo-ku, Tokyo, Japan
[3] Department of Computer Science, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo, Japan

E-mail: `komeiji@sip.tuat.ac.jp`, `tanakat@cc.tuat.ac.jp`

August 2024

Abstract. This study describes speech synthesis from an Electrocorticogram (ECoG) during imagined speech. We aim to generate high-quality audio despite the limitations of available training data by employing a Transformer-based decoder and a pretrained vocoder. Specifically, we used a pre-trained neural vocoder, Parallel WaveGAN, to convert the log-mel spectrograms output by the Transformer decoder, which was trained on ECoG signals, into high-quality audio signals. In our experiments, using ECoG signals recorded from 13 participants, the synthesized speech from imagined speech achieved a dynamic time-warping (DTW) Pearson correlation ranging from 0.85 to 0.95. This high-quality speech synthesis can be attributed to the Transformer decoder's ability to accurately reconstruct high-fidelity log-mel spectrograms, demonstrating its effectiveness in dealing with limited training data.

1. Introduction

Brain–Computer Interfaces (BCIs) are technologies that establish a direct pathway between the brain and the external world. The development of techniques to interpret meanings and messages from measured brain activity is of paramount importance in BCI research. Generally, the methods for measuring, processing, and analyzing recorded brain activity vary depending on the type of message intended to be extracted from the brain. These targets may include motor actions or motor intentions [1, 2, 3], internal speech [4, 5, 6], or psychological states [7, 8, 9]. In particular, speech-related BCIs (speech BCIs) hold great promise as assistive technologies for patients who have lost or significantly impaired speech function due to brain disorders such as aphasia [10],

stroke, or amyotrophic lateral sclerosis [11]. These technologies have the potential to greatly enhance communication capabilities for individuals affected by such conditions. To achieve such a BCI, one of the promising ways to extract speech is by measuring invasive electrocorticography (ECoG) with implanted subdural grid electrodes. Using ECoG is superior to surface electroencephalography (EEG) in terms of spatiotemporal resolution and signal-to-noise ratio, making it particularly suitable for analyzing brain activity related to speech in the high gamma band [12, 13].

A review of previous studies shows that the forms of estimated speech from ECoG or stereo EEG (sEEG) signals are mainly categorized into two: text [14, 15, 16] or synthesized speech [17, 18, 19]. Most studies on synthesized speech in the form of audio signals follow a two-step strategy: the first step is the estimation of speech representation, and the second is the synthesis of audio signals from that representation. Herff et al. [17] estimated subsequent units of an audio signal using the participant's voice directly from ECoG signals. The estimated units were concatenated to form a sequence of speech. Anumanchipalli et al. [18] used a bidirectional long short-term memory (BLSTM) [20] to decode articulatory movements as speech representations from cortical activity. These representations were then transformed into audio signals. Angrick et al. [19] estimated a log-mel spectrogram using linear discriminant analysis to predict the energy level for spectral bins. The spectrogram was converted into an audio signal using the Griffin–Lim algorithm [21].

The quality of synthesized speech has improved dramatically with the help of neural vocoders such as WaveNet [22], WaveGlow [23], and Parallel WaveGAN [24]. Angrick et al. [25] used WaveNet to reconstruct spoken audio from ECoG signals. They estimated a log-mel spectrogram from ECoG signals using DenseNets [26] and converted it into audio signals using WaveNet. Kohler et al. [27] used WaveGlow to convert log-mel spectrograms into synthesized speech. They estimated log-mel spectrograms using a gated recurrent unit (GRU) [28] that considers the temporal dependencies of cortical activity, achieving better performance than when using DenseNets [25]. Shigemi et al. [29] estimated log-mel spectrograms as speech representation from ECoG signals during the overt reading using Transformer [30] and used a pre-trained model of Parallel WaveGAN [24] to synthesize speech from the estimated log-mel spectrograms, achieving high-quality speech synthesis.

Although previous studies have mainly focused on speech synthesis from overt speech, which is the neural activity during vocalized speech, the ultimate goal of a speech-BCI is to decode or synthesize covert (imagined) speech. However, collecting covert speech data for training is challenging. Unlike overt speech or perception tasks, covert speech involves silently reading text displayed on a screen without audio output. Capturing this internal process without external speech cues poses a significant challenge.

Thus, we apply pre-trained neural vocoders, which enable high-quality speech synthesis even with limited training data, to speech synthesis from covert speech. We hypothesize that pre-trained neural vocoders can generate high-quality synthesized

Table 1. Clinical profiles for 13 participants. The age, sex, handedness, and electrode location are listed.

Participant	Age	Sex	Handedness	Electrodes location
js4	25	F	Right	Left
js5	36	M	Left	Left
js6	29	M	Right	Left
js7	41	F	Right	Left
js8	20	M	Right	Left
js9	34	F	Right	Left
js10	24	F	Right	Both
js11	41	M	Right	Left
js12	17	M	Right	Left
js13	19	F	Right	Left
js14	30	M	Right	Left
js15	15	F	Right	Left
js16	8	M	Right	Left

speech from covert speech despite being trained on limited neural signals. To validate this hypothesis, we combine a BLSTM-based or Transformer-based decoder and pre-trained Parallel WaveGAN. This combination follows the network architecture proposed by Shigemi et al. [29], which successfully synthesized high-quality audio signals from overt speech, and the previous study by Komeiji et al. [31], which clarified that the Transformer encoder and decoder can decode text from covert speech. For the experiment, we recorded the ECoG signals of overt and covert speech using Japanese sentences by utilizing a "Karaoke-like text highlighting" technique that aided participants in reading the text aloud or silently. The recorded ECoG signals from 13 participants were used to obtain synthesized speech by combining the BLSTM or Transformer decoder with the pre-trained Parallel WaveGAN. The synthesized speech from overt or covert speech was evaluated by mean square error (MSE) loss, dynamic time-warping (DTW) Pearson correlation, character error rate (CER) of automatic speech recognition (ASR) results, and token error rate (TER) of dictation test results. The results showed that covert speech could be decoded into synthesized speech with high quality (around 0.9 DTW Pearson correlation) and accurate content (42.2% TER, below the 50.0% chance level, ($p < 0.05$).

2. Methods

2.1. Participants

The 13 volunteer participants (js4 to js16; according to Table 1, eight males, eight females, mean age 26.1 years, range 8–41) in this study were undergoing treatment for drug-resistant temporal lobe epilepsy at the Department of Neurosurgery, the

Juntendo University Hospital. ECoG electrode arrays were surgically implanted on each participant's cortical surface to localize their seizure foci. ECoG and audio data in this study were obtained from all 13 human participants who gave informed consent. The Medical Research Ethics Committee of Juntendo University Hospital approved the experimental design to record the data. The Ethics Committee of the Tokyo University of Agriculture and Technology approved the data analysis. All research was performed according to relevant guidelines and regulations.

2.2. Data acquisition

The ECoG signals were recorded using intracranial electrodes implanted on the cortical surface. We set the ground (GND) as an epidural electrode. All signals other than the one with the epidural electrode, between their respective electrodes and the GND as a reference, were amplified using a biomedical amplifier (g.HIAMP, g.tec Medical Engineering GmbH, Austria). The electrooculograms were recorded simultaneously with the ECoG signals using four disposable electrodes to confirm eye movements. In addition to ECoG signals, audio signals, such as participants' utterances, were recorded using a microphone (ECM360, SONY, Japan) and digitized in sync with the ECoG signals. The ECoG and audio signals were digitized at sampling rates of 9,600 Hz.

Using a photodetector circuit, we recorded trigger signals indicating changes in the event status. The circuit combined a linear light sensor (LLS05-A, Nanyang Senba Optical and Electronic Co., Ltd., China) with an operational amplifier (NJM3404AD, New Japan Radio, Japan). The photodetector was placed in a monitor corner where the pixel colors changed in response to the event status, capturing and recording the changes. We used the Simulink toolbox (MathWorks, USA) to control the biomedical amplifier and record all observed signals.

2.3. Experimental design

We recorded the three conditions of ECoG signals (i.e., auditory speech perception, overt speech, and covert speech). To record these ECoG signals, we designed the task schedule depicted in Fig. 1a. According to the task schedule, participants performed perception, overt, and covert tasks. We refer to the series of three tasks as a "track." A single sentence, which was presented at random, was recorded for each track. More than 6 s intervals between tasks can prevent time-correlated artifacts. Participants repeated this track 80 times (if they made a mistake, it would take more than 80 times), and we obtained 80 segments of ECoG and audio signals for each task and participant. During the recording, the participants were seated on a chair with a desk equipped with a screen, loudspeaker, microphone, and keyboard. The monitor and loudspeaker were operated based on the task schedule.

The track consisted of the perception, overt, and covert tasks in that order. At the beginning of each task, the instruction and countdown periods were inserted to prepare the participants for the task (yellow and blue boxes depicted in Fig. 1a).

In the perception task, the participants were asked to listen to the sentence played from the loudspeaker and gaze at the cross symbol displayed on the monitor. In the covert and overt tasks, the participants were asked to read silently or aloud a sentence characterized as "Karaoke-like text highlighting" displayed on the monitor. The speed of text highlighting was at a constant rate (5 characters/s). To avoid reading faster or slower than the highlight speed at the beginning of a sentence, we inserted a dummy word, which does not exist in the Japanese words and has no meaning but can only be pronounced, at the beginning of the sentence (lime green boxes in Fig. 1a). At the end of the track, the participants operated the keyboard to judge whether this track had any mistakes (red box in Fig. 1a). The next track started if the participant judged there was no mistake; otherwise, this track was repeated.

During the experiments, we continuously recorded ECoG signals, an audio signal, and a trigger signal, which indicated the periods of the dummy words and sentences (lime green and green boxes in Fig. 1a). The trigger signal was used to trim both the ECoG and audio signals.

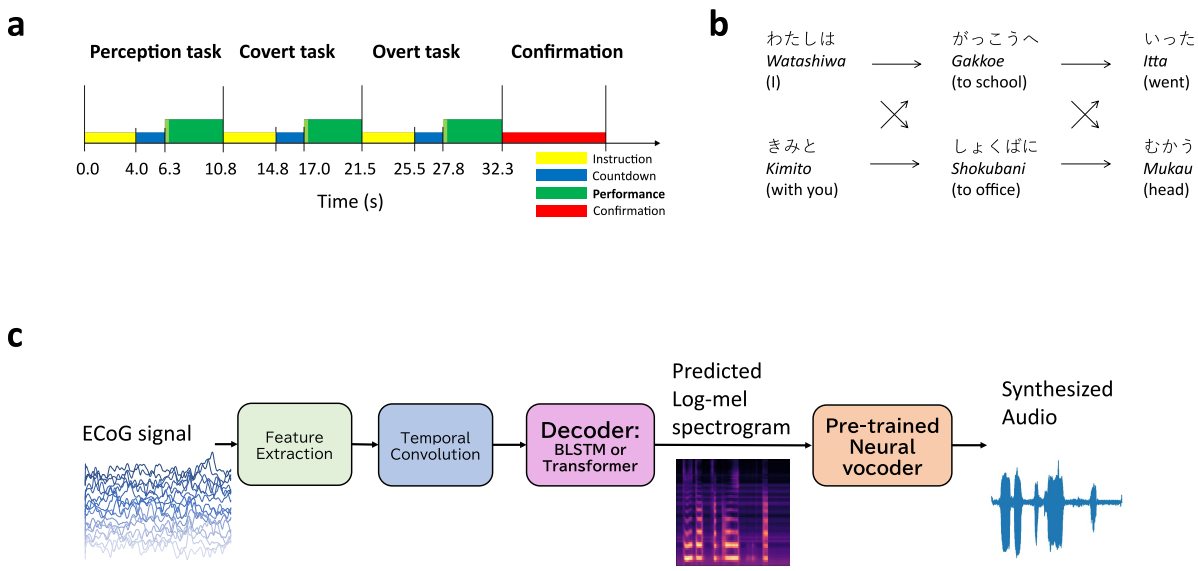

Figure 1. Study overview a Experimental design. A single track includes over 33 seconds to record sentences for the three tasks. The yellow box indicates task instructions, the blue indicates a countdown for preparation, and the green indicates task performance: perception, covert, or overt. The red box is the prompt status for error judgment. **b** Sentence generation. Sentences comprising three Japanese phrases, each with two options. This results in a total of eight sentences. The figure depicts the structure of each sentence, with Hiragana representation at the top, Romaji (pronunciation) in the middle, and English translation in parentheses below. **c** The decoding pipeline. The pipeline comprises the feature extraction and neural network, which comprises the temporal convolution, log-mel spectrogram decoder, and pre-trained neural vocoder. ECoG signals are input into the feature extraction. Synthesized audio signals are output from the pre-trained neural vocoder.

2.4. Sentences

The sentences displayed to the participants comprised three Japanese phrases (Fig. 1b). The total number of sentences was eight. Thus, the same sentence appeared ten times for all 80 trials of each task. Each phrase had two candidates to generate one sentence. The first phrase was either "watashiwa" (I) or "kimito" (with you), the second phrase was either "gakko:e" (to school) or "shokubani" (to the office), and the third phrase was either "itta" (went) or "mukau" (head). Consequently, we generated eight ($= 2 \times 2 \times 2$) sentence patterns.

2.5. Decoding pipeline

We synthesized audio signals from each segment of the ECoG signals by employing the decoding pipeline shown in Fig. 1c. The pipeline comprises the feature extraction and neural network parts, which comprise the temporal convolution, log-mel spectrogram decoder, and pre-trained neural vocoder.

In the feature extraction part, the segments of the ECoG signals were converted to envelopes of high gamma bands with a sampling rate of 200 Hz. A detailed explanation of the feature extraction part is presented in Section 2.6. The envelopes were input to the temporal convolution part to downsample the time axis effectively. The output of the temporal convolution was input to the log-mel spectrogram decoder. The predicted log-mel spectrograms were transformed into speech waveforms using a pre-trained neural vocoder.

In this study, we implemented the log-mel spectrogram decoder in two types, BLSTM-based and Transformer-based. The architectures for both decoders are the same as Shigemi et al. [29]. Section 2.8 provides a detailed explanation of the architecture.

2.6. ECoG preprocessing

Preprocessing, which obtains the ECoG features from the ECoG signals used for training and decoding in a machine learning network, entails the following steps [32]. In the first step, electrodes with artifacts or excessive noise were removed through visual inspection by an epileptologist. Consequently, the numbers of electrodes used for the participants in the analysis were as follows: 54 for js4, 54 for js5, 54 for js6, 53 for js7, 55 for js8, 71 for js9, 42 for js10, 45 for js11, 39 for js12, 56 for js13, 58 for js14, 65 for js15, and 24 for js16. The valid electrode coverage across all participants is depicted in Fig. S1. In the second step, we trimmed each segment from both the ECoG and audio signals based on the start and end times of sentences from the trigger signal; specifically, we discarded the periods of the dummy words. The length of each trimmed segment was the maximum duration of all sentences throughout the experiment for each participant. This operation prevented the decoder from predicting log-mel spectrograms based on their trimmed length. The length was approximately 3.5 seconds. The trim operation before signal processing, such as filtering, can prevent time-correlated artifacts that

are the effects of the previous segments. In the third step, the trimmed ECoG signals were anti-aliased (low-pass filtered at 200 Hz) and downsampled to 400 Hz. Infinite impulse response notch filters of 50 and 100 Hz were applied to suppress the power supply noise and the harmonic. Finally, the amplitudes of the analytic signals at each electrode were extracted in each band of eight adjacent frequency bands, between 70 and 150 Hz, filtered by eight finite impulse response (FIR) bandpass filters, averaged across bands, and downsampled to 200 Hz. The eight FIR bandpass filters were designed with passbands of 68–78, 74–84, 82–92, 91–102, 101–112, 112–124, 124–136, and 137–150 Hz. The amplitudes of the analytic signals were then z-scored, and the ECoG features were obtained.

2.7. Audio preprocessing

The log-mel spectrograms to train a decoder were calculated from the trimmed audio using the "Python speech features package [33]" via the following procedure. First, the audio data were upsampled to 24,000 Hz. After upsampling, we calculated the log-mel spectrograms in the frequency range of 80–7,600 Hz. We used 80 mel frequency bins, 1,200 window lengths, and 300 hop sizes (12.5 msec) in the calculation. These settings were adjusted based on the training data of the pre-training neural vocoder model for speech waveform (see Section 2.9).

2.8. Neural network architecture

We applied the BLSTM and Transformer to a log-mel spectrogram decoder. BLSTM [34] is a type of recurrent neural network (RNN) explicitly designed to handle both long-term and short-term memory efficiently. It finds common applications in ASR and natural language processing (NLP). On the other hand, the Transformer [30] is a neural network model distinguished by its attention layers and is well-suited for capturing long-term dependencies. Since 2018, the Transformer has emerged as the de facto standard in machine learning technology, in ASR and NLP, and in diverse fields, including image processing.

Here, we explain the detailed network architectures implemented by both the BLSTM and the Transformer-based decoder, as shown in Fig. 2. The network architectures are depicted in Fig. 2, and the implementation parameters are described in Table 2. The network processes the sequences in two stages: a temporal convolution network and a decoder network.

In the first stage, the temporal convolution (convolutional neural network; CNN) layer effectively downsampled the ECoG features. The input to the CNN consisted of K features, each of length L (where K represents the number of electrodes). The number of electrodes K ranged from 24 to 72 among all 13 participants, and the length L was approximately 700 (3.5 s sequence of a 200 Hz feature). Each convolutional filter with a kernel size of $W \times K$ was applied to the K features with a stride of W; consequently,

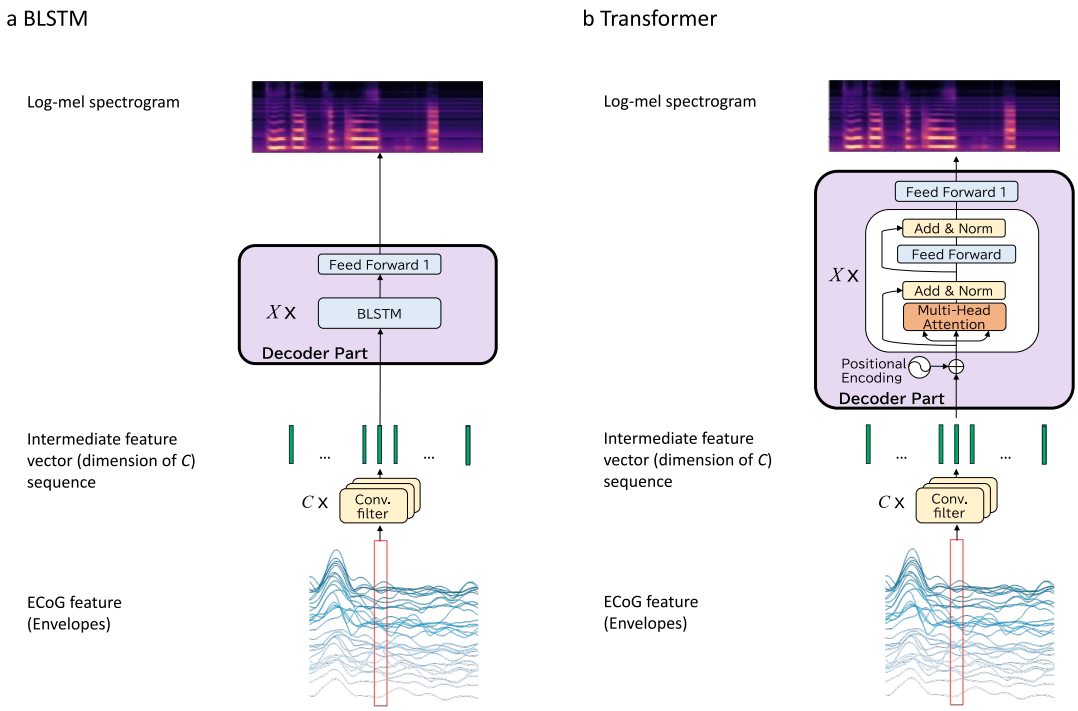

Figure 2. Detailed decoder architectures.

the CNN layer output a sequence of N ($= \mathrm{ceil}(L/W)$) vectors of C with a dimension of C.

In the second stage, the BLSTM or Transformer decoder received the outputs of the temporal convolution layer and comprised a stack of X identical layers. For the Transformer decoder, the identical layers have two sub-layers. The first sub-layer was the multi-head self-attention mechanism, and the second was a fully connected feed-forward network. Furthermore, these two layers had a residual connection, and layer normalization was applied to their outputs. The outputs of the BLSTM or Transformer decoder were the sequences of feature vectors of dimension C with a length of N, the same size as the inputs. The outputs of the stack of X identical layers were input to the feed-forward network called "feed-forward 1" to represent the log-mel spectrogram.

2.9. Neural Vocoder

In our study, Parallel WaveGAN [24] (as implemented here: `https://github.com/kan-bayashi/ParallelWaveGAN`) was used to synthesize the speech waveform. Parallel WaveGAN is a waveform generator based on a non-autoregressive WaveNet model and is trained to synthesize speech from a given log-mel spectrogram. We used an existing vocoder model pre-trained on the JSUT corpus dataset containing Japanese speech. The weights of the vocoder model we used are publicly available as "jsut_parallel_wavegan.v1."

Table 2. Network parameters

Layer name	Parameter	Output shape
Temporal convolution	filters = 100 kernel size = (12, K) strides = (2,1) dropout = 0.1	(B, N, 100)
Transformer decoder	$X = 8$ attention head = 10 hidden units = 100 dropout = 0.5	(B, N, 100)
BLSTM decoder	$X = 4$ forward units = 120 hidden units = 100 dropout = 0.5	(B, N, 100)
Feed-forward 1	units = 80	(B, N, 80)

2.10. Network Implementation and Parameters

The network hyperparameters are summarized in Table 2. In the temporal convolution layer, the number of convolutional filters (C) and kernel size (W) were set to 100 and 12, respectively. In contrast, the number of electrodes (K) remained variable depending on the participants. The filters were applied with a stride size of two. The output shape variables B and N represent the batch size for training and sequence length, which depends on the data. In the decoder, we set the numbers of identical layers (X) to eight for the Transformer and four for the BLSTM. The numbers of model parameters of the Transformer and BLSTM decoders were comparable: 1,129,305 for the Transformer and 1,182,905 for the BLSTM. In "feed-forward 1," the output dimension was set to 80, equivalent to the log-mel spectrogram dimension.

2.11. Network Training

Our neural networks are trained by reducing the MSE losses of log-mel spectrogram output from the decoder. The training hyperparameters are listed in Table 3. We tested 800, 1,200, 1,600, and 2,400 training epochs for the Transformer and BLSTM to identify the optimal performance. Based on these tests, 2,400 epochs were selected for the Transformer, and 800 epochs were selected for the BLSTM.

For the overt task, we utilized audio signals synchronized with overt speech in ECoG signals as the ground truth for training. Conversely, for the covert task, we employed audio signals from perception tasks, which were not synchronized with covert speech in ECoG signals.

Table 3. Training parameters

Parameter name	Value
Learning rate	0.0005
Batch size	16
Training epochs (Transformer decoder)	2,400
Training epochs (BLSTM decoder)	800

2.12. Evaluation Method

We evaluated the synthesized speech using five-fold cross-validation. For all 13 participants who completed both overt and covert tasks, up to 80 trials were divided into five sets: four sets, containing up to 64 trials, were used for training the neural network model, while the remaining set, consisting of up to 16 trials, was used for decoding the speech waveform. Consequently, we obtained up to 80 synthesized speech samples for all 13 participants across both tasks to be evaluated.

2.12.1. Quality evaluation To evaluate the quality of the decoded log-mel spectrogram and synthesized speech waveform, we introduced MSE loss and DTW Pearson correlation[35] under two cases: BLSTM and Transformer decoder. MSE losses were used as criteria for training the decoders and can assess the prediction accuracy of estimating the log-mel spectrograms by a decoder. DTW Pearson correlation evaluates the synthesized speech, specifically the output of the neural vocoder. In other words, it assesses the entire process of the decoding pipeline from ECoG signals to synthesized speech.

In addition to evaluating MSE loss and DTW Pearson correlation, we introduced speech recognition evaluation using a general ASR model. For the speech recognition evaluation, we applied Whisper [36] to obtain transcripts from the synthesized speech. We used the "Large-v3" model for Whisper and set the language option to "ja" for Japanese language recognition. The output texts generated by the Whisper were evaluated by CER as shown in the following equation:

$$\mathrm{CER} = \frac{S_c + D_c + I_c}{N_c} \times 100, \tag{1}$$

where S_c is the number of substituted characters, D_c is the number of deleted characters, I_c is the number of inserted characters, and N_c is the number of characters in reference sentence.

2.12.2. Gaussian noise input For further analysis, we replaced the ECoG signals with Gaussian noise as input to the Transformer decoder. First, the Transformer decoder was trained using pairs of Gaussian noise and log-mel spectrograms. We generated Gaussian noise with a duration of 3.5 seconds, a sampling rate of 9,600 Hz, and 54 channels, corresponding to the median number of electrodes across 13 participants. The log-mel spectrograms were computed from the playback presented to participants during

the perception task. Next, we obtained a log-mel spectrogram from Gaussian noise using the trained Transformer decoder. For the evaluation of speech synthesis from Gaussian noise input, we generated 80 Gaussian noise signals, comparable to the number of trials for one participant in one task. In the same way as the evaluation of speech synthesis from ECoG signals, we used five-fold cross-validation, and we obtained 80 speeches as decoding results.

2.12.3. Dictation test For the dictation test, we listened to the synthesized speech. For each speech sample, we authors chose one sentence out of nine options comprised of eight sentences displayed to the 13 participants and "not audible" if we could not recognize the speech sample. As the chosen options other than "not audible" for dictation and their reference sentences consist of three tokens, we introduced TER for evaluating the chosen sentence, as shown in the following equation:

$$\mathrm{TER} = \frac{S_t + D_t + I_t}{N_t} \times 100, \tag{2}$$

where S_t is the number of substituted tokens, D_t is the number of deleted tokens, I_t is the number of inserted tokens, and N_t is the number of tokens in the reference sentence.

For the dictation test in the overt task, we listened to the anonymized synthesized speech generated by the decoder, where the audio signals for training data were voice-converted to ensure the privacy of the 13 participants. To obtain voice-converted speech to train the decoder, we introduced "FreeVC" [37]. We used audio signals during the perception task as the target speaker of the voice conversion.

2.13. Electrode Contributions

We introduced the saliency map [38] to determine the electrode contribution for predicting log-mel spectrograms, as previously used in text decoding research [15] [31]. The saliency map represents the derivative of the output log-mel spectrogram for the input ECoG features. Its shape mirrors that of the input ECoG features, and each element signifies the magnitude of contribution for each time point and electrode. We were interested in determining the trend of electrode contributions across all trimmed ECoG signals for text decoding in each participant. To this end, we computed the variance along the time axis to observe the saliency trend over the time sequence. Our objective was not to obtain the absolute values for the magnitude of the contribution of each electrode but to compare the relative values among them. To normalize the contribution magnitudes across electrodes for a single trial, we calculated their z-score of variances. Finally, the z-scored values for each trial were averaged for the evaluation data sets of all 80 decoding times. The contribution magnitude, calculated as the average of the z-scores as described above, with the electrode positions for each participant, were transformed into Talairach coordinates and plotted on the averaged FreeSurfer pial surface image, known as "fsaverage" [39] [40] [41].

3. Results

We obtained decoded log-mel spectrograms and synthesized audio signals from both overt and covert speech of 13 participants using the decoding pipeline depicted in Fig. 1c, which includes BLSTM-based and Transformer-based decoders. Examples of the decoded log-mel spectrograms and synthesized audio signals, along with the ground truth, are shown in Fig. 3 and Fig. 4 for overt speech and covert tasks, respectively.

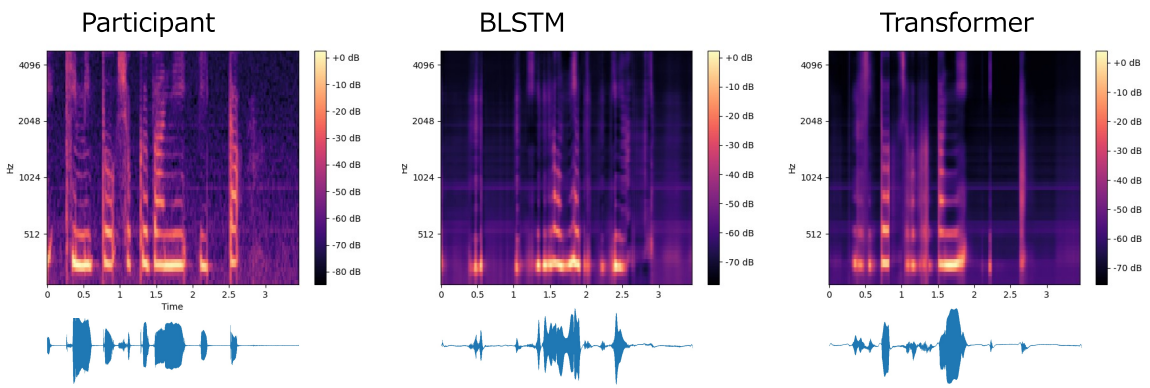

Figure 3. Decoded log-mel spectrograms and synthesized audio signals from the overt speech: Comparison of Participants(ground truth), BLSTM, and Transformer

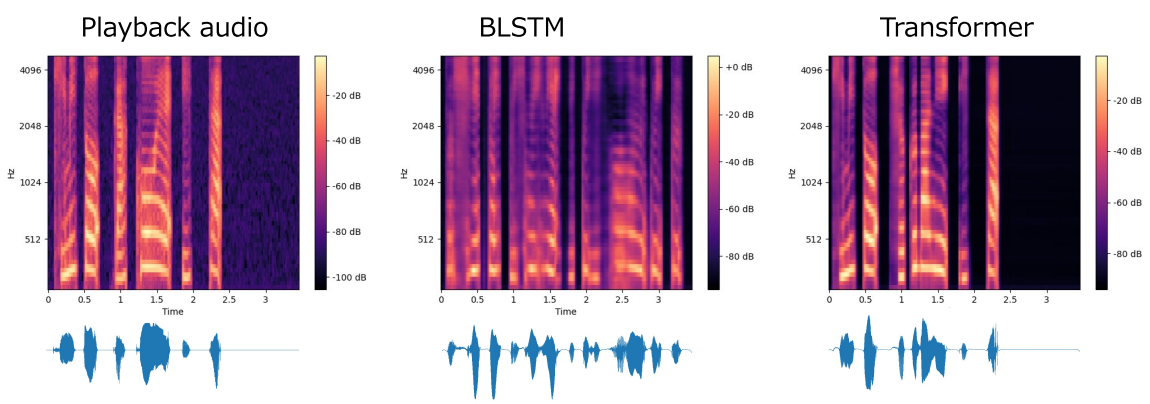

Figure 4. Decoded log-mel spectrograms and synthesized audio signals from the covert speech: Comparison of Playback audio (ground truth), BLSTM, and Transformer

3.1. Quality of Synthesized Speech

The results of the MSE loss of log-mel spectrograms and the DTW Pearson correlation of synthesized audio waveforms were displayed by box plots in Fig. 5 and Fig. 6 respectively. Each figure presents the results for overt speech (a) and covert speech (b). The green boxes represent the result of the BLSTM decoder, while the yellow boxes represent the result of the Transformer decoder. Each dot corresponds to the result of one participant. Note that for MSE loss, a smaller value indicates better quality, whereas for DTW Pearson correlation, a larger value indicates better quality.

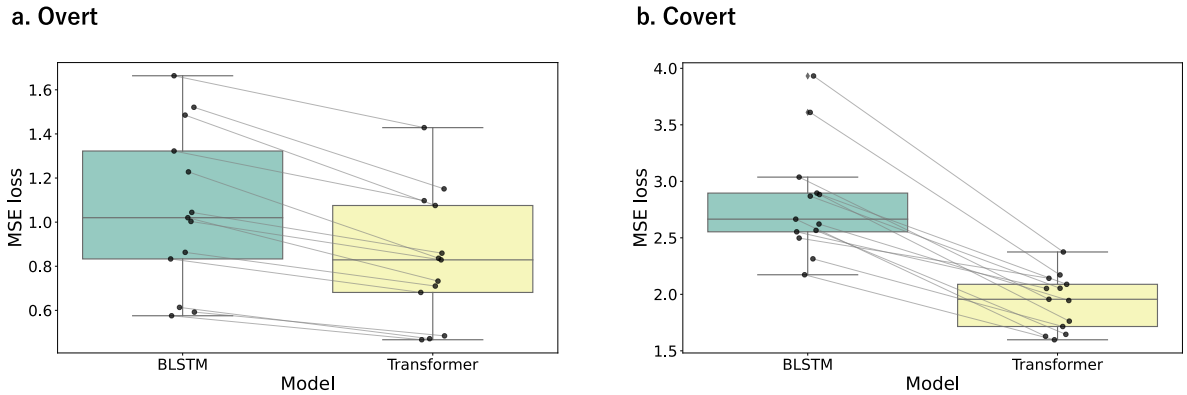

Figure 5. Decoded log-mel spectrograms and synthesized audio signals from the overt speech: Comparison of Participants(ground truth), BLSTM, and Transformer

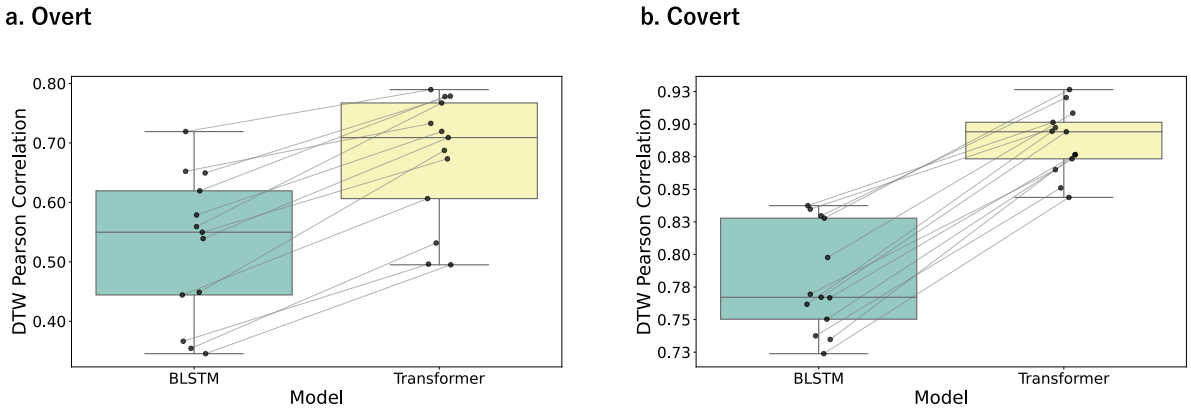

Figure 6. DTW Pearson correlation

According to the figures, the Transformer yielded lower MSE loss than the BLSTM with p-values of less than 0.05 for both the overt and covert tasks (Fig. 5). Cohen's d values were as high as 0.7 and 2.8, respectively. Moreover, the Transformer yielded a higher DTW Pearson correlation than the BLSTM decoder, with a p-value of less than 0.05 for both the overt and covert tasks (Fig. 6). Cohen's d values were as high as 1.3 and 3.2, respectively. The p-values were calculated using a one-sided Wilcoxon signed test.

The results of the CERs from the speech recognition output using Whisper by BLSTM and Transformer for overt and covert tasks are depicted in box plots (Fig. 7). Each dot corresponds to the result of one participant. According to the figure, the Transformer decoder outperformed the BLSTM with p-values of less than 0.05 for both overt and covert tasks. Cohen's d values were as high as 1.9 and 4.3, respectively. For the overt task, the Transformer achieved a CER ranging from 50% to 150%, while for the covert task, the Transformer achieved a CER ranging from 50% to 100%. Insertion errors accounted for a large portion of errors, resulting in CERs exceeding 100%. Most insertion errors were caused by the training bias of Whisper for the phrase "Thank you for watching" in Japanese YouTube videos. Correct rates, excluding insertion errors,

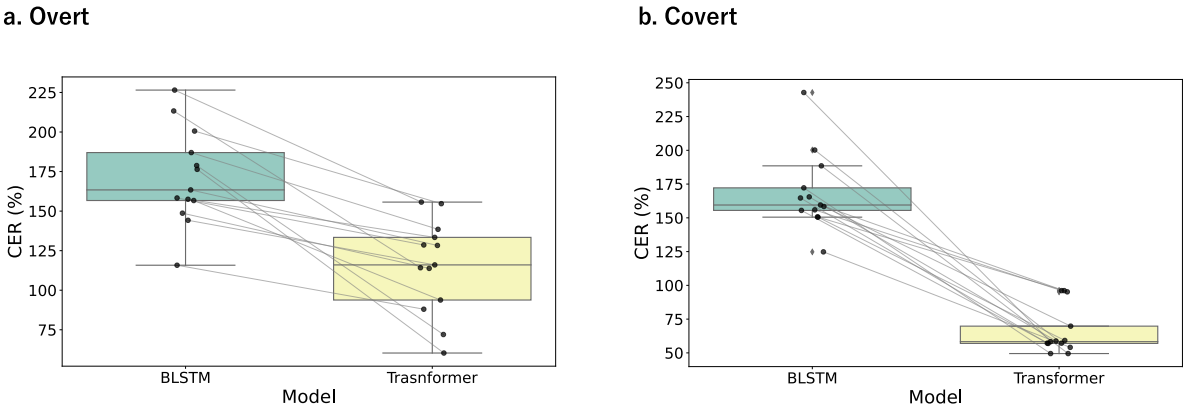

Figure 7. CERs by Whisper

ranged from 7.6% to 51.6% for the overt task and from 21.6% to 54.5% for the covert task. It is noteworthy that the covert task exhibited better performance than the overt task.

3.2. Gaussian noise input

The quality of the output log-mel spectrogram from the Transformer decoder, trained with pairs of Gaussian noise and log-mel spectrograms calculated from playback audio, was high when other Gaussian noise was used as input. To further analyze what enables the Transformer decoder to generate a high-quality log-mel spectrogram that maintains speech context, we input a half-length Gaussian noise sequence and adjusted the offset of the positional encoder. The results of this operation are depicted in Fig. **??**. According to the figure, the offset of the output half-length log-mel spectrogram was dependent on the offset of the positional encoder.

3.3. Dictation test

The results of the dictation test for synthesized speech by Transformer decoder are depicted in Fig. 9. The figure shows the TERs by each participant for overt and covert tasks. The green box represents the results of overt speech, while the yellow box represents the results of covert speech. Each dot corresponds to the result of one participant. The red dot corresponds to the result when the Gaussian noises were used for input.

Both the TERs of overt and covert speech were not significantly lower than when Gaussian noise was used as input, according to the permutation test. On the other hand, focusing on the minimum TERs for both overt and covert speech 18.3% for js13 and 42.2% for js16—these two TERs were significantly lower than the TER when Gaussian noise was used as input, according to the binomial test.

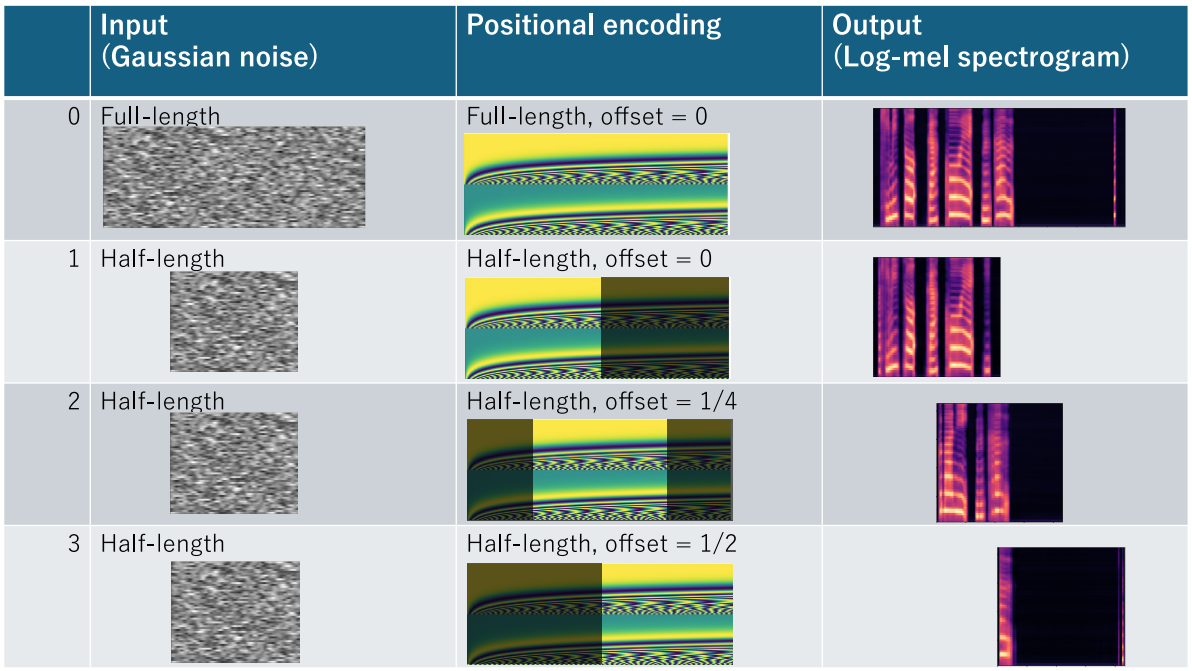

Figure 8. Offset control of positional encoding

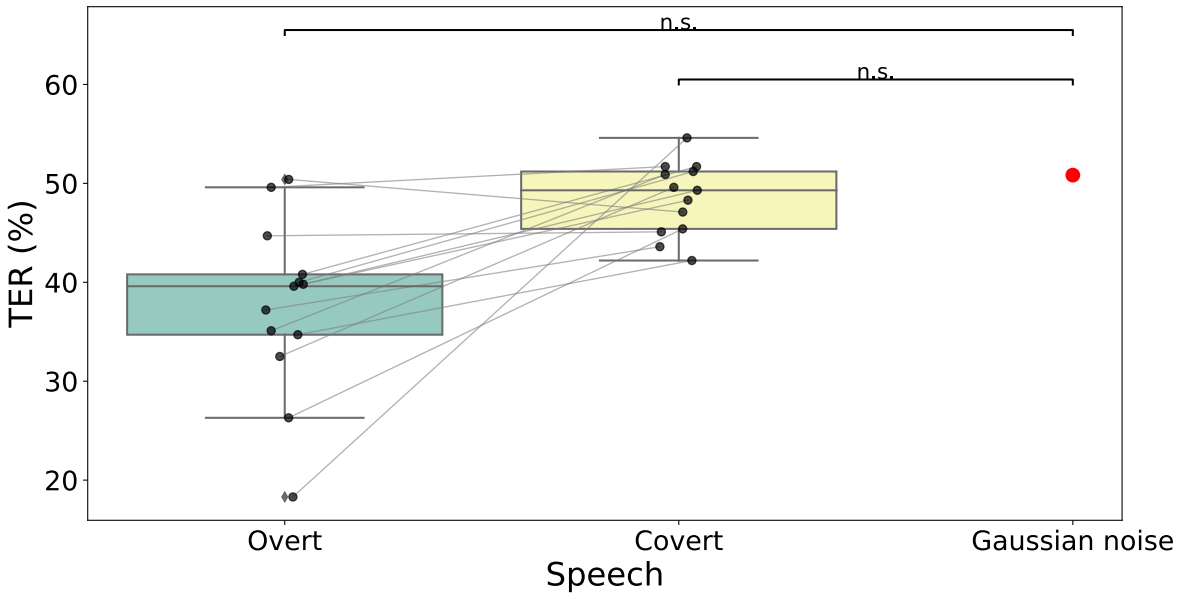

Figure 9. TERs by Dictation

3.4. Electrode Contribution

The electrode contributions for decoding the mel-spectrogram were depicted using a 3D brain object, as shown in Fig. 10. The figure displays example results for js13 in the overt task and js16 in the covert task. These participants achieved the minimum TER for both overt and covert tasks. In the overt task for js13, the contributing areas were the sensorimotor cortex, inferior frontal gyrus, and temporal pole. In contrast, in the

covert task for js16, the contributing area was the sensorimotor cortex.

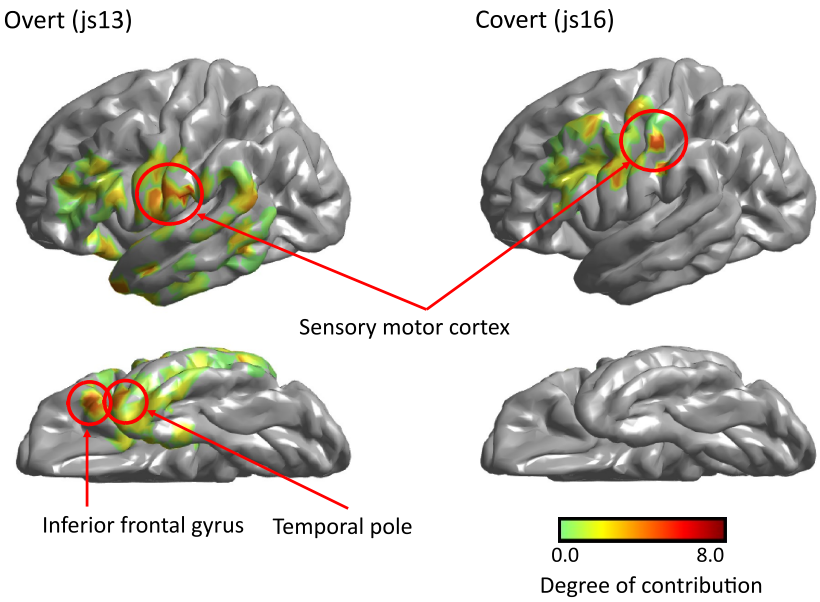

Figure 10. Electrode contribution

4. Discussion

In this section, we discuss the results from four aspects. Firstly, we discuss the results of the Transformer decoder outperforming BLSTM. Secondly, we analyze the ability of the Transformer decoder to output high-quality log-mel spectrograms even when Gaussian noise was used as input. Thirdly, we discuss the results of the dictation test and provide the conclusion of this study. Finally, the electrode contribution is discussed to provide a deeper understanding of this study.

For the first aspect, Transformer outperformed BLSTM on the MSE loss, DTW Pearson correlation, and CER by Whisper. Previous studies by Komeiji et al. [16, 31] and Shigemi et al. [29] have demonstrated that Transformer outperforms BLSTM for both text decoding and speech synthesis from ECoG signals. Consistent with these findings, the long-term dependencies of the input ECoG signals may play important roles in speech synthesis. Although Kohler et al. [27] achieved Pearson correlation coefficients of up to 0.8 using GRU, which is a type of RNN similar to BLSTM, for speech decoding of Dutch sentences from sEEG, adopting a Transformer instead of a GRU may further enhance the quality of results. Furthermore, Chen et al. [42] applied a Swin Transformer (a kind of Transformer) [43] to decode English speech from ECoG or sEEG and achieved correlation coefficients higher than 0.8. It is anticipated that Transformer will become the de facto standard for speech synthesis from ECoG or sEEG.

For the second aspect, we analyze the ability of the Transformer decoder to output high-quality log-mel spectrograms. Interestingly, this ability was evident even when Gaussian noise was used as input. Further analysis revealed that positional encoding

was crucial in producing this high-quality output. Positional encoding provides temporal context to the input. Therefore, even when Gaussian noise, which inherently lacks context, is used as input, the output of the Transformer possesses temporal context. This suggests that the Transformer decoder may generate output somewhat independent of its input. Consequently, we should evaluate not only the MSE loss or DTW Pearson correlation but also the accuracy of the content of the output speech and whether the input genuinely influenced the output.

The third aspect reveals that the dictation test provided insight into the Transformer decoder's ability to generate speech signals with accurate content based on the input ECoG signals. During the dictation test, the respondent (author) selected one sentence out of eight, not including "not audible" options, after listening to decoded audio speech. For participant js13 (TER: 18.3%) in the overt task and participant js16 (TER: 42.2%) in the covert task, the TERs were significantly lower than the TER when Gaussian noise was used as input (binomial test). The high-quality output and accurate content achieved in speech synthesis from the covert task validated our hypothesis that pre-trained neural vocoders can generate high-quality speech despite the limited ECoG signals available for training data. Additionally, this result clarified that the synthesized speech audio was decoded based on the information in the ECoG signal.

Conversely, for several participants, the TER during the dictation test was similar to when Gaussian noise was used as input. This indicates that the speech content was jumbled, even though the synthesis qualities were high. High-quality speech output can be attributed to the use of the Transformer decoder and pre-trained neural vocoder, as well as the limited number of output variations. More training data will be necessary to ensure the accuracy of the speech content. To summarize the third aspect, while this study has largely addressed achieving high-quality sound even with small amounts of training data, future work should focus on synthesizing speech with precise content.

Finally, we discuss the anatomical analysis using saliency maps for a deeper understanding of this study. For participants js13 in the overt task and js16 in the covert task (see 10), the sensory-motor cortex (SMC) commonly contributed. In the study of text decoding [31, 6], SMC was also identified as a contributing region. Whether the output target is speech or text, SMC could play a pivotal role in BCI.

Comparing it to text decoding, we observed two interesting points. Firstly, in the covert task, while the saliency map for text decoding exhibited the contribution distributions were broader than the overt task, speech decoding revealed narrower contribution distributions. According to Tian et al. [44], the cortical activity of covert speech involves a broader area in the cortex. Text is inherently more abstract than speech signals. We hypothesize that in the covert task, decoding a more abstract target may involve a broader range of ECoG signals, whereas decoding a more concrete target, such as speech mel spectrogram, may involve signals from a narrower area, specifically speech-specific regions. Secondly, in the overt task, the temporal lobe contributed more significantly than in text decoding. We posit that decoding speech involves ECoG signals associated with speech perception.

5. Conclusion

In this study, we demonstrated the successful synthesis of speech from neural activity during covert reading using ECoG signals. By leveraging a Transformer-based decoder and a pre-trained neural vocoder, Parallel WaveGAN, we achieved high-quality audio generation despite the constraints of limited training data. Our experiments with ECoG signals from 13 participants showed that the synthesized speech attained a DTW Pearson correlation ranging from 0.85 to 0.95. These results underscore the effectiveness of the Transformer decoder in accurately reconstructing high-fidelity log-mel spectrograms from ECoG signals, even with a limited-size dataset. The ability to generate high-quality speech from covert reading highlights the potential of this approach for advanced speech BCI applications, where decoding speech from neural activity can enable new forms of communication for individuals with speech impairments.

Acknowledgment

This work was supported in part by JSPS KAKENHI, 20H00235, and 23H00548.

Reference

[1] Wing-kin Tam, Tong Wu, Qi Zhao, Edward Keefer, and Zhi Yang. Human motor decoding from neural signals: a review. *BMC Biomedical Engineering*, 1:1–22, 2019.

[2] Sharlene N Flesher, John E Downey, Jeffrey M Weiss, Christopher L Hughes, Angelica J Herrera, Elizabeth C Tyler-Kabara, Michael L Boninger, Jennifer L Collinger, and Robert A Gaunt. A brain-computer interface that evokes tactile sensations improves robotic arm control. *Science*, 372(6544):831–836, 2021.

[3] Hamdi Altaheri, Ghulam Muhammad, Mansour Alsulaiman, Syed Umar Amin, Ghadir Ali Altuwaijri, Wadood Abdul, Mohamed A Bencherif, and Mohammed Faisal. Deep learning techniques for classification of electroencephalogram (eeg) motor imagery (mi) signals: A review. *Neural Computing and Applications*, 35(20):14681–14722, 2023.

[4] Stephanie Martin, Peter Brunner, Iñaki Iturrate, José del R Millán, Gerwin Schalk, Robert T Knight, and Brian N Pasley. Word pair classification during imagined speech using direct brain recordings. *Scientific reports*, 6(1):25803, 2016.

[5] Timothée Proix, Jaime Delgado Saa, Andy Christen, Stephanie Martin, Brian N Pasley, Robert T Knight, Xing Tian, David Poeppel, Werner K Doyle, Orrin Devinsky, et al. Imagined speech can be decoded from low-and cross-frequency intracranial eeg features. *Nature communications*, 13(1):48, 2022.

[6] Francis R Willett, Erin M Kunz, Chaofei Fan, Donald T Avansino, Guy H Wilson, Eun Young Choi, Foram Kamdar, Matthew F Glasser, Leigh R Hochberg, Shaul Druckmann, et al. A high-performance speech neuroprosthesis. *Nature*, 620(7976):1031–1036, 2023.

[7] Philip A Kragel and Kevin S LaBar. Decoding the nature of emotion in the brain. *Trends in cognitive sciences*, 20(6):444–455, 2016.

[8] Christof Koch, Marcello Massimini, Melanie Boly, and Giulio Tononi. Neural correlates of consciousness: progress and problems. *Nature reviews neuroscience*, 17(5):307–321, 2016.

[9] Lijuan Duan, Huifeng Duan, Yuanhua Qiao, Sha Sha, Shunai Qi, Xiaolong Zhang, Juan Huang, Xiaohan Huang, and Changming Wang. Machine learning approaches for mdd detection and emotion decoding using eeg signals. *Frontiers in Human Neuroscience*, 14:284, 2020.

[10] Nina Simmons-Mackie, Chris Code, Elizabeth Armstrong, Lillian Stiegler, and Roberta J Elman. What is aphasia? results of an international survey. *Aphasiology*, 16(8):837–848, 2002.

[11] Shiyu Luo, Qinwan Rabbani, and Nathan E Crone. Brain-computer interface: applications to speech decoding and synthesis to augment communication. *Neurotherapeutics*, 19(1):263–273, 2023.

[12] Nathan E Crone, Diana L Miglioretti, Barry Gordon, and Ronald P Lesser. Functional mapping of human sensorimotor cortex with electrocorticographic spectral analysis. II. Event-related synchronization in the gamma band. *Brain: A Journal of Neurology*, 121(12):2301–2315, 1998.

[13] Xiaomei Pei, Eric C Leuthardt, Charles M Gaona, Peter Brunner, Jonathan R Wolpaw, and Gerwin Schalk. Spatiotemporal dynamics of electrocorticographic high gamma activity during overt and covert word repetition. *Neuroimage*, 54(4):2960–2972, 2011.

[14] Christian Herff, Dominic Heger, Adriana De Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in neuroscience*, 8:141498, 2015.

[15] Joseph G Makin, David A Moses, and Edward F Chang. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature neuroscience*, 23(4):575–582, 2020.

[16] Shuji Komeiji, Kai Shigemi, Takumi Mitsuhashi, Yasushi Iimura, Hiroharu Suzuki, Hidenori Sugano, Koichi Shinoda, and Toshihisa Tanaka. Transformer-based estimation of spoken sentences using electrocorticography. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1311–1315. IEEE, 2022.

[17] Christian Herff, Lorenz Diener, Miguel Angrick, Emily Mugler, Matthew C Tate, Matthew A Goldrick, Dean J Krusienski, Marc W Slutzky, and Tanja Schultz. Generating Natural, Intelligible Speech From Brain Activity in Motor, Premotor, and Inferior Frontal Cortices. *Frontiers in Neuroscience*, 13:1267, 2019.

[18] Gopala K Anumanchipalli, Josh Chartier, and Edward F Chang. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753):493–498, 2019.

[19] Miguel Angrick, Maarten C Ottenhoff, Lorenz Diener, Darius Ivucic, Gabriel Ivucic, Sophocles Goulis, Jeremy Saal, Albert J Colon, Louis Wagner, Dean J Krusienski, et al. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Communications biology*, 4(1):1055, 2021.

[20] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610, 2005.

[21] Daniel Griffin and Jae Lim. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984.

[22] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A Generative Model for Raw Audio. *arXiv:1609.03499*, 2016.

[23] Ryan Prenger, Rafael Valle, and Bryan Catanzaro. Waveglow: A Flow-based Generative Network for Speech Synthesis. In *ICASSP 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3617–3621, 2019.

[24] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram. In *ICASSP 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203. IEEE, 2020.

[25] Miguel Angrick, Christian Herff, Emily Mugler, Matthew C Tate, Marc W Slutzky, Dean J Krusienski, and Tanja Schultz. Speech synthesis from ECoG using densely connected 3D convolutional neural networks. *Journal of Neural Engineering*, 16(3):036019, 2019.

[26] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[27] Jonas Kohler, Maarten C Ottenhoff, Sophocles Goulis, Miguel Angrick, Albert J Colon, Louis

Wagner, Simon Tousseyn, Pieter L Kubben, and Christian Herff. Synthesizing speech from intracranial depth electrodes using an encoder-decoder framework. *arXiv preprint arXiv:2111.01457*, 2021.

[28] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[29] Kai Shigemi, Shuji Komeiji, Takumi Mitsuhashi, Yasushi Iimura, Hiroharu Suzuki, Hidenori Sugano, Koichi Shinoda, Kohei Yatabe, and Toshihisa Tanaka. Synthesizing speech from ecog with a combination of transformer-based encoder and neural vocoder. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[31] Shuji Komeiji, Takumi Mitsuhashi, Yasushi Iimura, Hiroharu Suzuki, Hidenori Sugano, Koichi Shinoda, and Toshihisa Tanaka. Feasibility of decoding covert speech in ecog with a transformer trained on overt speech. *Scientific Reports*, 14(1):11491, 2024.

[32] Kristofer E Bouchard, Nima Mesgarani, Keith Johnson, and Edward F Chang. Functional organization of human sensorimotor cortex for speech articulation. *Nature*, 495(7441):327–332, 2013.

[33] James Lyons. Python speech features. `https://github.com/jameslyons/python_speech_features`, 2023-12-10.

[34] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[35] Stan Salvador and Philip Chan. Toward accurate dynamic time warping in linear time and space. *Intelligent Data Analysis*, 11(5):561–580, 2007.

[36] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[37] Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[38] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[39] Yasuo Nakai, Jeong-won Jeong, Erik C Brown, Robert Rothermel, Katsuaki Kojima, Toshimune Kambara, Aashit Shah, Sandeep Mittal, Sandeep Sood, and Eishi Asano. Three-and four-dimensional mapping of speech and language in patients with epilepsy. *Brain*, 140(5):1351–1370, 2017.

[40] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

[41] Alexander M Chan, Janet M Baker, Emad Eskandar, Donald Schomer, Istvan Ulbert, Ksenija Marinkovic, Sydney S Cash, and Eric Halgren. First-pass selectivity for semantic categories in human anteroventral temporal lobe. *Journal of Neuroscience*, 31(49):18119–18129, 2011.

[42] Junbo Chen, Xupeng Chen, Ran Wang, Chenqian Le, Amirhossein Khalilian-Gourtani, Erika Jensen, Patricia Dugan, Werner Doyle, Orrin Devinsky, Daniel Friedman, et al. Subject-agnostic transformer-based neural speech decoding from surface and depth electrode signals. *bioRxiv*, pages 2024–03, 2024.

[43] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin

transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[44] Xing Tian, Jean Mary Zarate, and David Poeppel. Mental imagery of speech implicates two mechanisms of perceptual reactivation. *Cortex*, 77:1–12, 2016.