Speech Synthesis from Electrocorticography during Imagined Speech Using a Transformer-Based Decoder and a 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`

June 2025

Abstract. This study presents a method for synthesizing speech from Electrocorticogram signals recorded during imagined speech. To address the limitations posed by the available training data, we employed a Transformer-based decoder to generate log-mel spectrograms, which were then converted into high-quality audio using a pre-trained neural vocoder, Parallel WaveGAN. In experiments involving electrocorticography (ECoG) recordings from 13 participants, the synthesized speech achieved Pearson correlation coefficients ranging from 0.85 to 0.95. These results demonstrate the effectiveness of the Transformer decoder in reconstructing accurate spectrograms from ECoG signals, even under data-constrained conditions.

1. Introduction

Brain–Computer Interfaces (BCIs) are technologies that establish a direct pathway between the brain and the external environment. The development of techniques to interpret meanings and messages from recorded brain activity is of paramount importance in BCI research. Generally, the methods for measuring, processing, and analyzing brain activity vary depending on the type of information to be extracted. 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 significantly enhance communication capabilities for individuals affected by such conditions. To achieve such a BCI, one promising approach is to extract speech by measuring invasive electrocorticography (ECoG) using implanted subdural grid electrodes. ECoG offers superior spatiotemporal resolution and signal-to-noise ratio compared to surface electroencephalography (EEG), making it particularly suitable for analyzing speech-related brain activity in the high gamma band [12, 13].

A review of previous studies reveals that estimated speech from ECoG or stereo-EEG (sEEG) signals is primarily categorized into two forms: textual output [14, 15, 16] and synthesized speech [17, 18, 19]. Most studies producing synthesized speech in audio format adopt a two-stage approach: the first stage involves estimating a speech representation, followed by the synthesis of audio signals from that representation. Herff et al. [17] directly estimated discrete audio units from ECoG signals corresponding to the participant's voice and concatenated them to generate continuous speech. Anumanchipalli et al. [18] employed a bidirectional long short-term memory (BLSTM) network [20] to decode articulatory kinematics as an intermediate speech representation from cortical activity, which was subsequently transformed into audio signals. Angrick et al. [19] estimated log-mel spectrograms by applying linear discriminant analysis to predict energy levels across spectral bins. Then, these spectrograms were converted into audio using the Griffin–Lim algorithm [21].

The quality of synthesized speech has significantly improved with the advent 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 first estimated log-mel spectrograms using DenseNets [26], which were then input to WaveNet. Kohler et al. [27] utilized WaveGlow for spectrogram-to-speech conversion, estimating log-mel spectrograms from cortical activity using a gated recurrent unit (GRU) [28], which captured temporal dependencies more effectively than DenseNets. Shigemi et al. [29] employed a Transformer-based model [30] to estimate log-mel spectrograms from ECoG signals recorded during overt speech tasks and used a pre-trained Parallel WaveGAN [24] to synthesize high-quality speech from the estimated spectrograms.

Previous studies have focused on synthesizing speech from overt speech, the neural activity during vocalized speech. Still, the ultimate goal of speech BCIs is to decode or synthesize covert (imagined) speech. However, covert speech data for training is difficult to collect because the state of covert speech cannot be observed externally. To address this challenge, we developed a task presentation paradigm termed "Karaoke-like text highlighting," and demonstrated the feasibility of decoding text from covert speech using a Transformer model trained on overt speech data collected with this method [31]. These findings suggest that covert speech may share common neural signal patterns with overt speech, an idea supported by several previous studies [32, 5, 33, 34]. Based on several lines of evidence showing shared neural mechanisms between overt and covert speech [32, 5, 33, 34], we hypothesize that decoder architectures effective for overt speech synthesis could be successfully applied to covert speech. Specifically,

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

we propose that Transformer-based decoders for log-mel spectrogram estimation–well-suited for capturing long-range dependencies in sequential data–when combined with pre-trained neural vocoders, can address the data limitations inherent in covert speech processing.

To validate the hypothesis, we utilized ECoG recordings from 13 participants, identical to those used in our previous study [31]. These signals were used to synthesize speech by combining either a BLSTM-based or a Transformer-based decoder with a pre-trained Parallel WaveGAN. The quality and accuracy of the synthesized speech-derived from both overt and covert speech were evaluated using several metrics: mean squared error (MSE) loss, Pearson correlation based on a dynamic time warping (DTW) algorithm, character error rate (CER) from automatic speech recognition (ASR) outputs, and token error rate (TER) from dictation test results. The results demonstrated that covert speech could be synthesized with high fidelity (approximately 0.9 Pearson correlation) and semantic accuracy, as evidenced by a TER of 42.2%, which is significantly below the 50.0% chance level ($p < 0.05$).

2. Methods

2.1. Participants

In this study, the 13 volunteer participants (labeled js4 to js16; see Table 1) were undergoing treatment for drug-resistant temporal lobe epilepsy at the Department of Neurosurgery, Juntendo University Hospital. The group comprised seven male and six female individuals, with a mean age of 26.1 (range: 8–41) years. ECoG electrode arrays were surgically implanted on the cortical surface of each participant to localize seizure

foci. Prior to implantation, the language-dominant hemisphere was identified using functional magnetic resonance imaging (fMRI). ECoG and audio data were collected from all 13 participants, each of whom provided informed consent. The experimental protocol for data acquisition was approved by the Medical Research Ethics Committee of Juntendo University Hospital, and the data analysis was approved by the Ethics Committee of the Tokyo University of Agriculture and Technology. All research procedures were conducted in accordance with relevant institutional guidelines and regulations.

2.2. Listeners for Dictation Test

For the dictation test to evaluate the synthesized speech (see Section 2.13.3 for details), we recruited 14 native Japanese-speaking listeners with normal hearing to the synthesized speech. This test was approved by the Ethics Committee of the Tokyo University of Agriculture and Technology.

2.3. Data Acquisition

ECoG signals were recorded using intracranial electrodes implanted on the cortical surface. An epidural electrode was used as the ground (GND). All signals, except for the one from the epidural electrode, were recorded as differential signals with respect to the GND and amplified using a biomedical amplifier (g.HIAMP, g.tec Medical Engineering GmbH, Austria). Electrooculograms were simultaneously recorded using four disposable electrodes to monitor eye movements. In addition to ECoG signals audio signals, such as participants' vocalizations were recorded using a microphone (ECM360, SONY, Japan) and digitized in synchrony with the ECoG recordings. Both ECoG and audio signals were digitized at a sampling rate of $9,600$ Hz.

Trigger signals indicating event status changes were recorded using a photodetector circuit, which combined a linear light sensor (LLS05-A, Nanyang Senba Optical and Electronic Co., Ltd., Nanyang, China) with an operational amplifier (NJM3404AD, New Japan Radio, Tokyo Japan). The photodetector was positioned in a corner of the monitor screen, where pixel color changes were used to indicate event transitions. All signals were recorded using the Simulink toolbox (MathWorks, Natick, MA, USA), which also controlled the biomedical amplifier.

2.4. Experimental Design

We recorded ECoG signals under three conditions: auditory speech perception, overt speech, and covert speech. To collect these data, we designed the task schedule shown in Fig. 1a. According to this schedule, participants performed the perception, overt, and covert tasks sequentially, which we collectively refer to as a "track." For each track, a single sentence was randomly selected and presented. Intervals of more than 6 seconds between tasks were included to prevent time-correlated artifacts. Each participant

repeated the track 80 times. If an error occurred during a trial, that trial was repeated, and thus the total number of trials could exceed 80. As a result, we obtained 80 segments of ECoG and audio signals for each task and each participant. During the recordings, participants were seated in a chair at a desk equipped with a monitor, loudspeaker, microphone, and keyboard. The monitor and loudspeaker were controlled in accordance with the task schedule.

The track consisted of the perception, overt, and covert tasks in that order. At the beginning of each task, instruction and countdown periods were included to prepare the participants for the upcoming task (yellow and blue boxes depicted in Fig. 1a). In the perception task, participants were instructed to listen to the sentence played from the loudspeaker while gazing at a cross symbol displayed on the monitor. In the covert and overt tasks, participants were asked to read silently or aloud a sentence presented using a "Karaoke-like text highlighting" method on the monitor. The text highlighting proceeded at a constant rate of 5 characters per second. To prevent participants from reading faster or slower than the highlight speed at the beginning of the sentence, a dummy word–one that does not exist in the Japanese language and carries no meaning but is pronounceable–was inserted at the beginning of the sentence (gray boxes in Fig. 1a). At the end of each track, participants used the keyboard to indicate whether there were any mistakes during trial (red box in Fig. 1a). If no mistakes were reported, the next track commenced; otherwise, the current 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 (gray and green boxes in Fig. 1a). The trigger signal was used to trim both the ECoG and audio signals.

2.5. Sentences

The sentences displayed to the participants comprised three Japanese phrases (Fig. 1b). The total number of sentences was eight. Therefore, each of the eight unique sentences appeared ten times across the 80 trials per task. Each phrase had two options to generate a 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.

2.6. Decoding Process

We synthesized audio signals from each segment of the ECoG signals by employing the decoding process illustrated in Fig. 1c. The process consists of neural network components, including temporal convolution, a log-mel spectrogram decoder, and a pre-trained neural vocoder.

First, as input to the temporal convolution, the segments of the ECoG signals were preprocessed by extracting envelopes of the high gamma bands and downsampled to

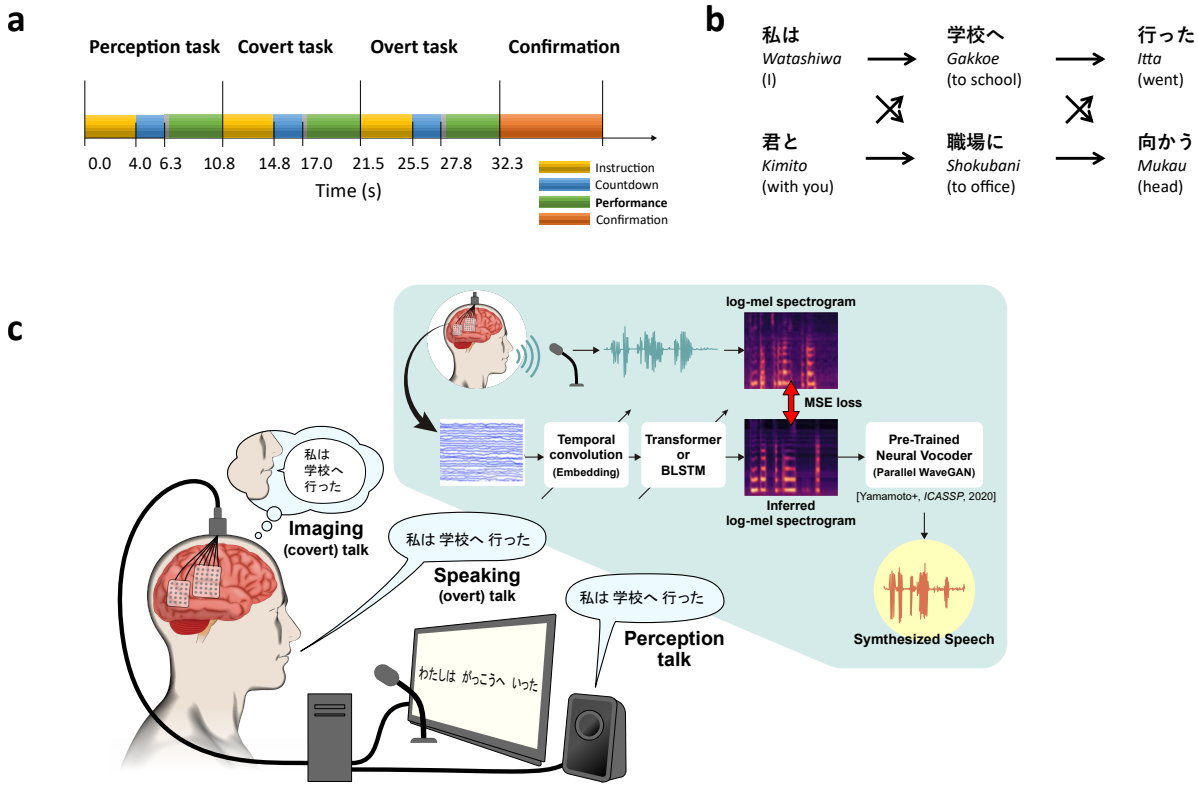

Figure 1. Study overview a Experimental design. A single track lasts over 33 seconds to record sentences for the three tasks. The yellow box indicates task instructions, the blue box indicates a countdown for preparation, and the green box indicates task performance: perception, covert, or overt. The red box indicates the prompt for error judgment. **b** Sentence generation. Sentences consist of three Japanese phrases, each with two options, resulting in a total of eight sentences. The figure depicts the structure of each sentence, showing the Hiragana representation at the top, Romaji (pronunciation) in the middle, and English translation in parentheses below. **c** Decoding process. The process involves a neural network, which composed of a temporal convolution, a log-mel spectrogram decoder, and a pre-trained neural vocoder.

a sampling rate of 200 Hz. A detailed explanation of the preprocessing is provided in Section 2.7. Next, the envelopes were fed into the temporal convolution module to effectively downsample the temporal dimension. Then, the output of the temporal convolution was passed to the log-mel spectrogram decoder. The predicted log-mel spectrograms were subsequently converted into speech waveforms using a pre-trained neural vocoder.

In this study, we implemented two types of log-mel spectrogram decoders: BLSTM-based and Transformer-based. The architectures of both decoders follow the designs described by Shigemi et al. [29]. A detailed description of the network architecture is provided in Section 2.9.

2.7. ECoG Preprocessing

Preprocessing to extract ECoG features from the raw signals for machine learning training and decoding involved the following steps [35]. First, electrodes exhibiting artifacts or excessive noise were excluded based on visual inspection by an epileptologist. The number of electrodes retained for analysis per participant were: 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 spatial coverage of valid electrodes across all participants is illustrated in Fig. S1. Second, we segmented both ECoG and audio signals using trigger timestamps indicating sentence start and end, discarding the initial dummy word periods. The length of each trimmed segment corresponded to the maximum sentence duration recorded during the experiment for each participant (approximately 3.5 s). This ensured uniform input length for the decoder and prevented bias from variable segment durations. Performing trimming prior to further signal processing, also mitigated time-correlated artifacts originating from adjacent segments. Third, the trimmed ECoG signals were anti-aliased using a low-pass filter at 200 Hz and downsampled to 400 Hz. Notch filters at 50 and 100 Hz (infinite impulse response) were applied to suppress power line noise and its harmonics. Finally, analytic signal amplitudes were extracted from eight adjacent frequency bands between 70 and 150 Hz. Each band was filtered by a finite impulse response bandpass filter with passbands: 68–78, 74–84, 82–92, 91–102, 101–112, 112–124, 124–136, and 137–150 Hz. These band-specific amplitudes were averaged and downsampled to 200 Hz. Then, the resulting analytic amplitudes were z-scored electrode-wise to produce the final ECoG features used for decoding.

2.8. Audio Preprocessing

The log-mel spectrograms used to train the decoder were computed from the trimmed audio signals using the `Python Speech Features` package [36] following this procedure. First, the audio data were upsampled to $24,000$ Hz. Next, log-mel spectrograms were calculated over the frequency range of 80–$7,600$ Hz, which effectively captures the acoustic features relevant for speech [37]. We used 80 mel frequency bins, with a window length of $1,200$ samples and a hop size of 300 samples (corresponding to 12.5 ms). These parameters were chosen to match those of the pre-trained neural vocoder model used for waveform synthesis (see Section 2.10).

2.9. Neural Network Architecture

We implemented both BLSTM- and Transformer-based architectures for the log-mel spectrogram decoder. BLSTM [38] is a type of recurrent neural network (RNN) designed to efficiently capture both long- and short-term dependencies, making it widely used in ASR and natural language processing (NLP). In contrast, the Transformer [30] employs self-attention mechanisms to model long-range dependencies and has become the state-

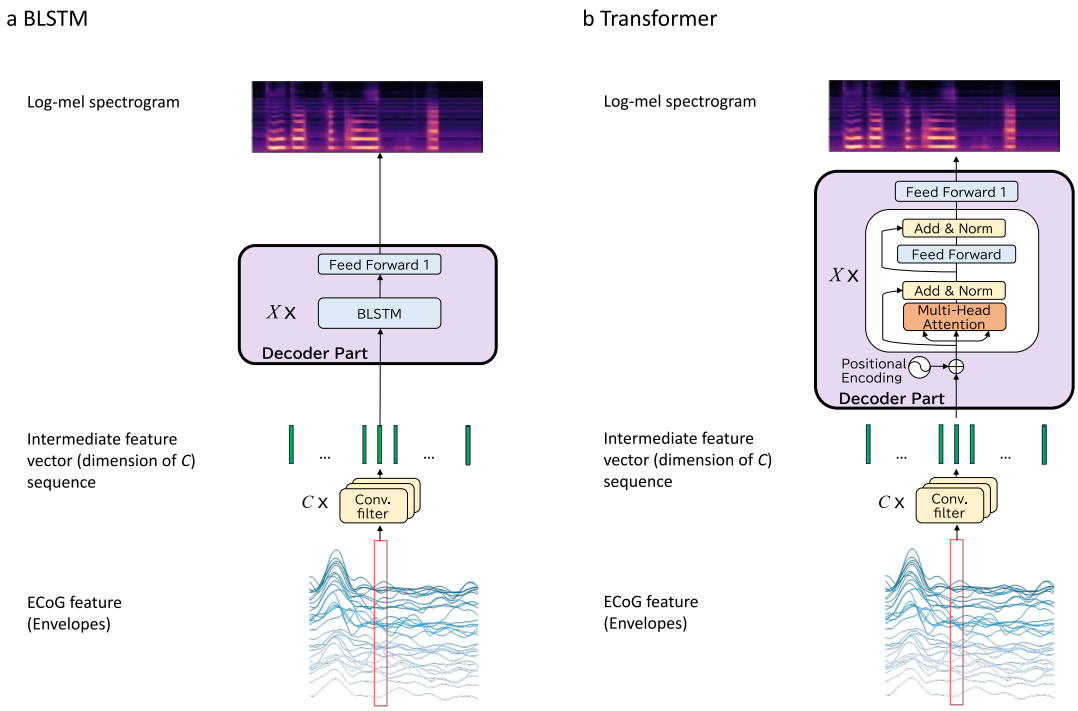

Figure 2. Detailed decoder architectures.

of-the-art architecture in ASR, NLP, and various other domains such as image processing since its introduction in 2017.

Figure 2 illustrates the detailed network architectures for both decoders. with implementation parameters summarized in Table 2. The decoding process occurs in two stages: a temporal convolutional network and a decoder network.

In the first stage, a temporal convolutional layer (CNN) is applied to downsample the ECoG feature sequences efficiently. The CNN input consists of K features, each of length L, where K is the number of electrodes (ranging from 24 to 72 across participants, and $L \approx 700$ corresponds to a 3.5-second sequence of sampled at 200 Hz. Convolutional filters with a kernel size of $W \times K$ are applied with stride W producing an output sequence of length $N = \lceil L/W \rceil$ with C-dimensional feature vectors.

In the second stage, the BLSTM or Transformer decoder takes the temporal convolution output as input. The decoder is composed of a stack of X identical layers. For the Transformer, each layer contains two sub-layers: a multi-head self-attention mechanism followed by a fully connected feed-forward network. Each sub-layer includes residual connections and layer normalization. The decoder outputs a sequence of N feature vectors of dimension C, matching the input sequence length. Finally, the output of the decoder stack is passed through a feed-forward network (labeled as "feed-forward 1" in Fig. 2) that transforms the feature vectors into an 80-dimensional log-mel spectrogram.

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. Neural Vocoder

In this study, we employed Parallel WaveGAN [24] (implemented as available at `https://github.com/kan-bayashi/ParallelWaveGAN`) to synthesize speech waveforms. Parallel WaveGAN is a waveform generator based on a non-autoregressive WaveNet architectures and is trained to generate speech from input log-mel spectrograms. We utilized a pre-trained vocoder model developed on the JSUT corpus dataset, which contains Japanese speech recordings. Specifically, we used the model weights from "jsut_parallel_wavegan.v1,"‡ which are publicly distributed online.

2.11. 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 depend on the data. In the decoder, we set the number of identical layers (X) to eight for the Transformer and four for the BLSTM. The number of model parameters of the Transformer and BLSTM decoders was 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.

‡ Available at `https://drive.google.com/open?id=1UDRL0JAovZ8XZhoH0wi9jj_zeCKb-AIA`

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. Network Training

Our neural networks are trained by minimizing the MSE between the predicted and target log-mel spectrograms output by the decoder. The training hyperparameters are listed in Table 3. We evaluated models trained for 800, $1,200$, $1,600$, and $2,400$ epochs for both the Transformer and BLSTM architectures based on the MSE loss on the training set, to determine the optimal number of training epochs. Based on this evaluation, $2,400$ epochs were selected for the Transformer and 800 epochs for the BLSTM.

For the overt task, we used audio signals that were temporally aligned with overt speech activity in the ECoG signals as the ground truth for training. Conversely, for the covert task, we used audio signals from perception tasks, which were not temporally aligned with covert speech in the ECoG signals.

2.13. Evaluation Method

To systematically evaluate our hypothesis, we employed multiple complementary assessment methods: (1) technical quality metrics, including MSE loss and Pearson correlation; (2) speech recognition performance using an automated system (Whisper ASR); (3) analysis with Gaussian noise input to investigate the decoder's intrinsic properties; and (4) human listener dictation tests to assess the real-world intelligibility of the synthesized speech.

The synthesized speech was evaluated using five-fold cross-validation. For all 13 participants who completed both overt and covert tasks, up to 80 trials were divided into five subsets: four subsets (up to 64 trials) were used to train the neural network model, and the remaining subset (up to 16 trials) was used for decoding the speech waveform. Separate models were trained for each participant and each fold. As a result, we obtained up to 80 synthesized speech samples per participant for both overt and covert tasks for evaluation.

2.13.1. Quality Evaluation To evaluate the quality of the decoded log-mel spectrogram and the synthesized speech waveform, we employed MSE loss and Pearson correlation based on a DTW algorithm [39] under two conditions: using the BLSTM and Transformer decoders. MSE loss was used as the training criterion for the decoders and serves to assess the accuracy of the predicted log-mel spectrograms. Pearson correlation

evaluates the synthesized speech, specifically the output produced by the neural vocoder. Especially, it measures the performance of the entire decoding pipeline, from ECoG signals to synthesized speech.

In addition to MSE loss and Pearson correlation, we incorporated an ASR evaluation using a general-purpose ASR model. For this evaluation, we applied Whisper [40] to transcribe the synthesized speech. We used the "Large-v3" model of Whisper and set the language option to "ja" for Japanese speech recognition. The transcribed output generated by Whisper was evaluated using the CER, as defined by 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 the reference sentence.

2.13.2. Gaussian Noise Input To investigate whether the log-mel spectrogram is generated in a regression-like or classification-like manner, we trained the model using Gaussian noise as input. If the model operates in a regression-like manner, it would fail to extract meaningful features from the noise, and consequently, a log-mel spectrogram with formant-like structures would not be generated. In contrast, if the model exhibits classification-like behavior, it may still produce structured outputs that reflect learned categorical information, even when the input lacks semantic content. First, the Transformer decoder was trained on 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, matching the median number of electrodes across the 13 participants. The log-mel spectrograms were computed from the audio playback presented to participants during the perception task. Next, the trained Transformer decoder was used to generate log-mel spectrograms from Gaussian noise inputs. For evaluating speech synthesis from Gaussian noise, we generated 80 Gaussian noise signals, corresponding to the number of trials for a single participant in one task. Following the same evaluation protocol as used for ECoG-based speech synthesis, we employed a five-fold cross-validation procedure and obtained 80 synthesized speech samples as decoding results.

2.13.3. Dictation Test To assess the real-world intelligibility of the synthesized speech, we recruited 14 listeners (see Section 2.2 for a detailed description). For each speech sample, listeners were instructed to select one sentence from a set of nine options, which included eight predefined sentences and a "not audible" option to be chosen if the sample was unrecognizable. As both the selected options (excluding "not audible") and the reference sentences consisted of three tokens, we employed the TER to evaluate the

accuracy of the selected sentences, defined as follows:

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

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 total number of tokens in the reference sentence.

For the dictation test in the overt task, listeners were presented with anonymized synthesized speech generated by the decoder. To safeguard the privacy of the 13 participants, the audio signals used for training were voice-converted. Voice conversion was conducted using "FreeVC" [41], with the speech from the perception task employed as the target speaker for the conversion.

2.14. Electrode Contributions

We introduced the saliency map method [42] to determine electrode contributions to the prediction of log-mel spectrograms, following its prior application in text decoding studies [15] [31]. The saliency map is computed as the derivative of the output log-mel spectrogram with respect to the input ECoG features. Its dimensions match those of the input with each element representing the contribution magnitude specific time point and electrode. Our aim was to examine the overall trend of electrode contributions across all trimmed ECoG signals used for text decoding in each participant. To capture this trend, we calculated the saliency values along the time axis. Rather than focusing on absolute contribution magnitudes, we compared their relative values among electrodes. For each trial, we normalized the electrode-wise variances by converting them to z-scores. Then, these z-scored values were averaged across all 80 decoding trials used for evaluation. Finally, the average z-score values representing electrode contributions were mapped to each participant's electrode locations, transformed into Talairach coordinates and visualized on the standard FreeSurfer pial surface image, "fsaverage" [43] [44] [45].

3. Results

We obtained decoded log-mel spectrograms and synthesized audio signals from both overt and covert speech of 13 participants using the decoding process depicted in Fig. 1c, which includes BLSTM-based and Transformer-based decoders. Examples of 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. In both overt and covert speech, the log-mel spectrograms generated by the BLSTM and Transformer exhibit formant-like stripe patterns, which are characteristic of natural human speech. The Transformer notably produced temporal patterns–particularly speech onset and offset points–that more closely resembled the ground truth compared to the BLSTM outputs.

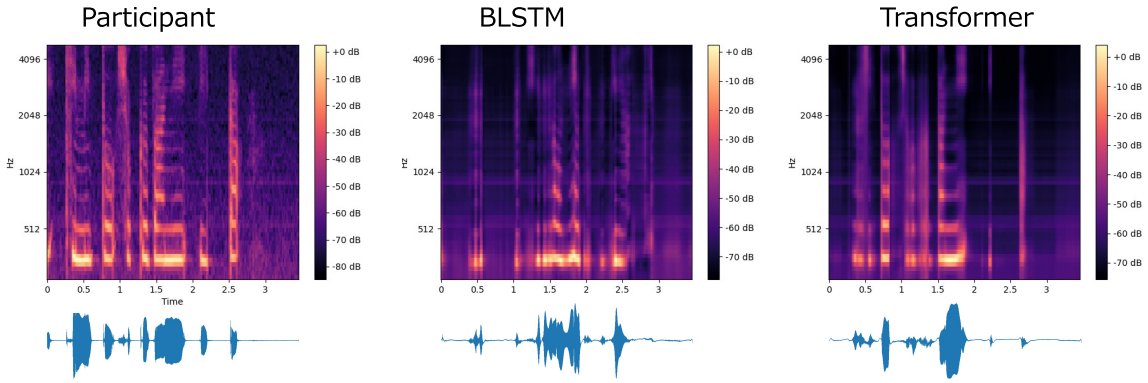

Figure 3. Decoded log-mel spectrograms and synthesized audio signals from overt speech: Comparison of participant (ground truth), BLSTM, and Transformer

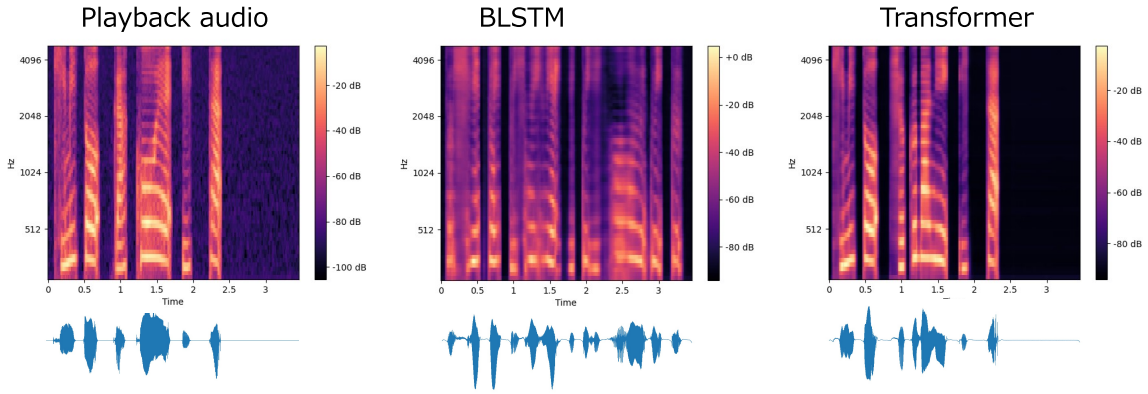

Figure 4. Decoded log-mel spectrograms and synthesized audio signals from covert speech: Comparison of playback audio (ground truth), BLSTM, and Transformer

3.1. Quality of Synthesized Speech

The quality of the synthesized speech was evaluated using multiple metrics: MSE loss of log-mel spectrograms, Pearson correlations of synthesized audio waveforms, and CERs from ASR.

3.1.1. MSE Loss Comparison The MSE losses of log-mel spectrograms are displayed in Fig. 5 for both overt speech (a) and covert speech (b). The green boxes represent the BLSTM decoder results, while the yellow boxes represent the Transformer decoder results. Each dot corresponds to the result of one participant, with lower values indicating better quality. According to Fig. 5, the Transformer exhibited lower MSE loss than the BLSTM in both overt and covert tasks, a difference confirmed to be statistically significant ($p < 0.001$) by a one-sided Wilcoxon signed-rank test. The corresponding Cohen's d values were as high as 0.7 and 2.8, respectively.

3.1.2. Pearson Correlation Analysis Fig. 6 presents the Pearson correlations, where higher values indicate better quality. As with the MSE loss, results are shown for

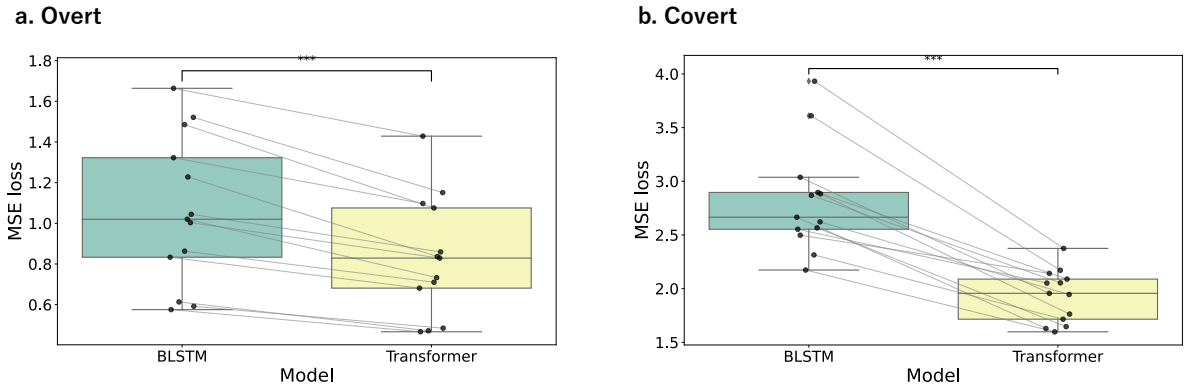

Figure 5. MSE loss comparison between BLSTM and Transformer-based decoders. Box plots compare the MSE losses between BLSTM and Transformer-based decoders for overt (a) and covert (b) speech synthesis tasks. Each dot represents the average MSE loss for an individual participant, with lines connecting corresponding values across the two models. Statistical significance was assessed using a one-sided Wilcoxon signed-rank test, with the resulting p-values reported accordingly ($* * * : p < 0.001$).

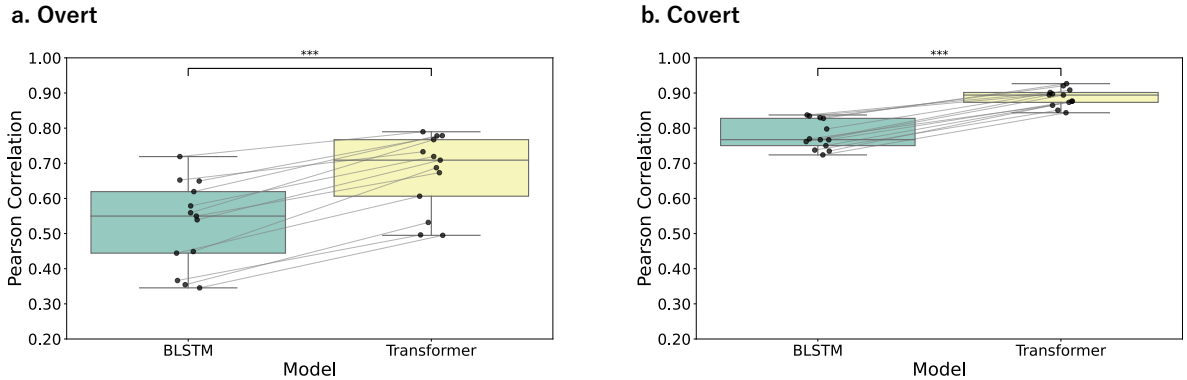

Figure 6. Pearson correlation comparison between BLSTM and Transformer-based decoders. Box plots compare the Pearson correlations between the BLSTM and Transformer-based decoders for overt (a) and covert (b) speech synthesis tasks. Each dot represents the average Pearson correlation for an individual participant, with lines connecting corresponding values across the two models. Statistical significance was assessed using a one-sided Wilcoxon signed-rank test, with the resulting p-values reported accordingly ($* * * : p < 0.001$).

both overt (a) and covert (b) speech, with green and yellow boxes representing BLSTM and Transformer decoders, respectively. The Transformer consistently yielded higher Pearson correlations than the BLSTM decoder, with p-values of less than 0.001 for both tasks (Fig. 6). The corresponding Cohen's d values were 1.3 and 3.2, respectively, demonstrating large effect sizes.

3.1.3. Character Error Rate Performance The results of the CERs obtained from the ASR outputs using Whisper for the BLSTM and Transformer decoders in both overt and covert tasks are presented as box plots in Fig. 7. Each dot represents the CER for an

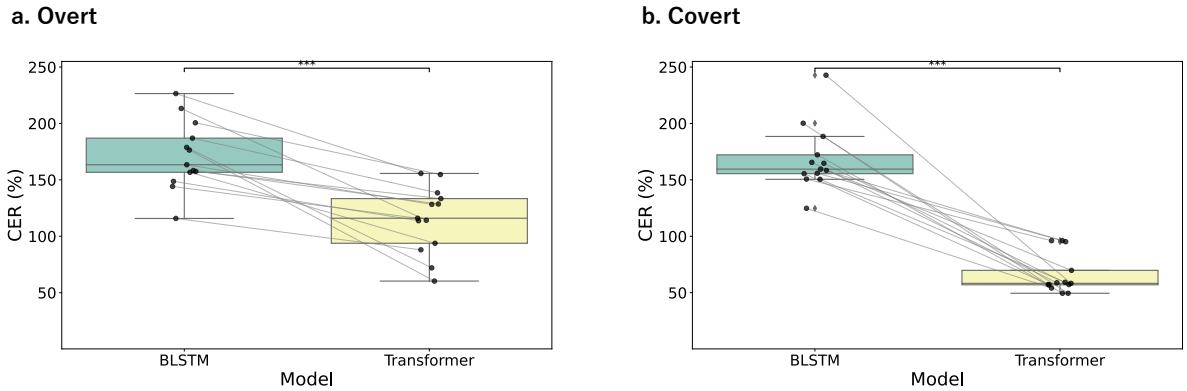

Figure 7. CERs by Whisper comparison between BLSTM and Transformer-based decoders. Box plots compare the CERs between BLSTM and Transformer-based decoders for overt (a) and covert (b) speech synthesis tasks. Each dot represents the average CER for each participant, and lines connect corresponding values for each participant across both models. The p-values were computed with a one-sided Wilcoxon signed test ($*** : p < 0.001$).

individual participant. The results indicate that the Transformer decoder consistently outperformed the BLSTM, with statistically significant differences ($p < 0.001$) for both overt and covert tasks. The corresponding Cohen's d values were 1.9 and 4.3 for the overt and covert tasks, respectively, indicating very large effect sizes. For the overt task, the Transformer achieved CERs ranging from 50% to 150%, whereas for the covert task, CERs ranged from 50% to 100%. Some CER values exceeded 100% due to a substantial number of insertion errors. These insertion errors were primarily attributed to Whisper's training bias towards the phrase "Thank you for watching" in Japanese, which is frequently encountered in YouTube videos. After excluding these insertion errors, the correct recognition rates ranged from 7.6% to 51.6% for the overt task and from 21.6% to 54.5% for the covert task. Notably, the covert speech synthesis task demonstrated superior recognition performance compared to the overt speech task.

3.2. Gaussian Noise Input

To investigate whether the Transformer decoder behaves more like a regressor or a classifier, we provided it with Gaussian noise as input. The underlying hypothesis was that only a classification-like model would produce structured spectrograms despite a lack of meaningful 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. The MSE loss was as low as 1.7, and the Pearson correlation was as high as 0.86 for 80 synthesized speech samples generated from Gaussian noise input. In contrast, when ECoG signals were input, the averages of the MSE loss and the Pearson correlation across 13 participants were 1.9 and 0.89, respectively.

To further analyze what enables the Transformer decoder to generate a high-quality

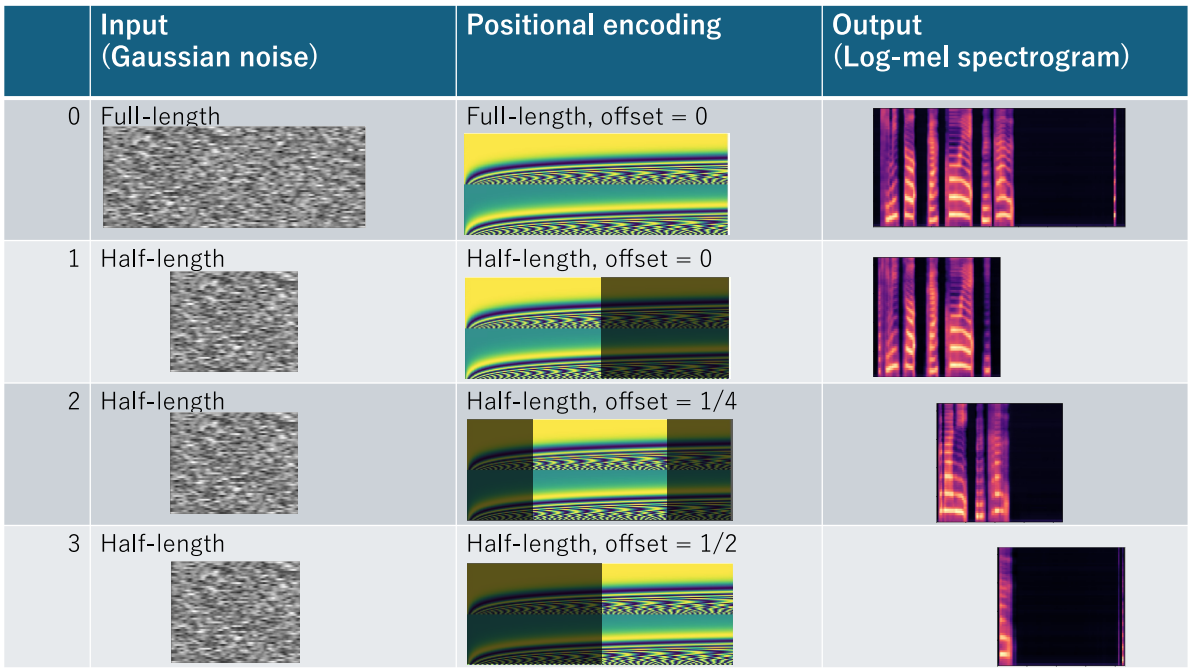

Figure 8. Offset control of positional encoding. Comparison of log-mel spectrogram outputs for Gaussian noise inputs with varying input lengths and positional encoding offsets applied to the Transformer-based decoder.

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 are shown in Fig. 8. As the figure demonstrates, the time position at which the patterns appeared in the output log-mel spectrogram shifted systematically with the offset of the positional encoding. This suggests that the decoder relies heavily on positional information when generating output, even without meaningful input content.

3.3. Dictation Test

In the initial results of the dictation test for synthesized speech from the Transformer decoder, the proportion of instances in which "not audible" was selected was 7.4% for the overt task and 0.8% for the covert task. The selection rate for the covert task was about approximately ten times lower than that for the overt task.

In the subsequent results of the dictation test, the TERs for synthesized speech from the Transformer decoder, calculated based on the selected options excluding "not audible," are presented in Fig. 9. The figure depicts the TERs for each participant in the overt and covert tasks. The green box corresponds to overt speech, while the yellow box corresponds to covert speech. Each dot represents the TER of an individual participant.

The median TERs for overt and covert speech were 37.6% and 47.7%, respectively. Conversely, the minimum TERs for overt and covert speech were 18.3% for js13 and 42.2% for js14, both of which were significantly lower than the chance level of 50%, as

determined by the binomial test ($p < 0.05$).

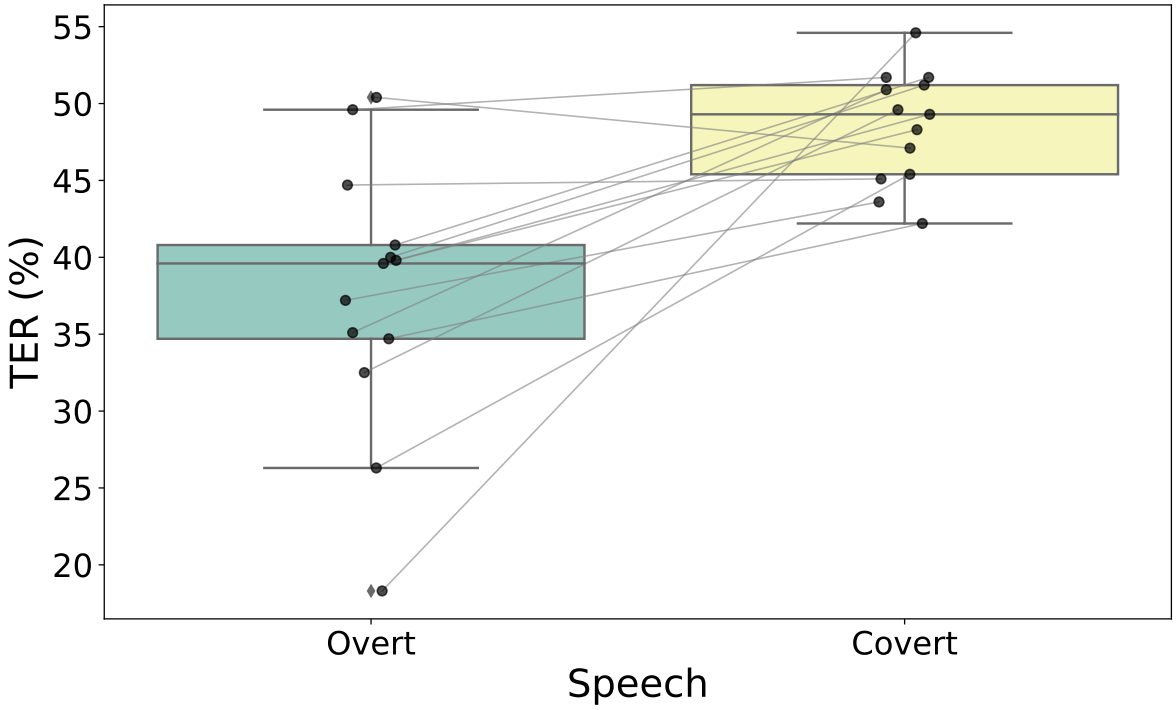

Figure 9. TERs by Dictation. Box plots comparing the TERs between overt and covert speech synthesis tasks. Each dot represents the average TER for one participant, and lines connect corresponding values for each participant across both models.

3.4. Electrode Contribution

The electrode contributions for decoding the mel-spectrogram were depicted using a three-dimensional brain object, as shown in Fig. 10. The figure displays example results for js13 in the overt task and js14 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 (SMC), inferior frontal gyrus, and temporal pole. In contrast, in the covert task for js14, the contributing area was the occipital lobe.

4. Discussion

In this study, we identified three key findings that warrant detailed discussion: (1) The Transformer-based decoder consistently outperformed BLSTM in both overt and covert speech synthesis across all quality metrics ($p < 0.001$ for MSE, Pearson correlation, and CER). (2) Despite the inherent challenges of covert speech, the synthesized output achieved remarkably high Pearson correlations (up to 0.95), with TERs significantly below the chance level for some participants. (3) Analysis using saliency maps revealed

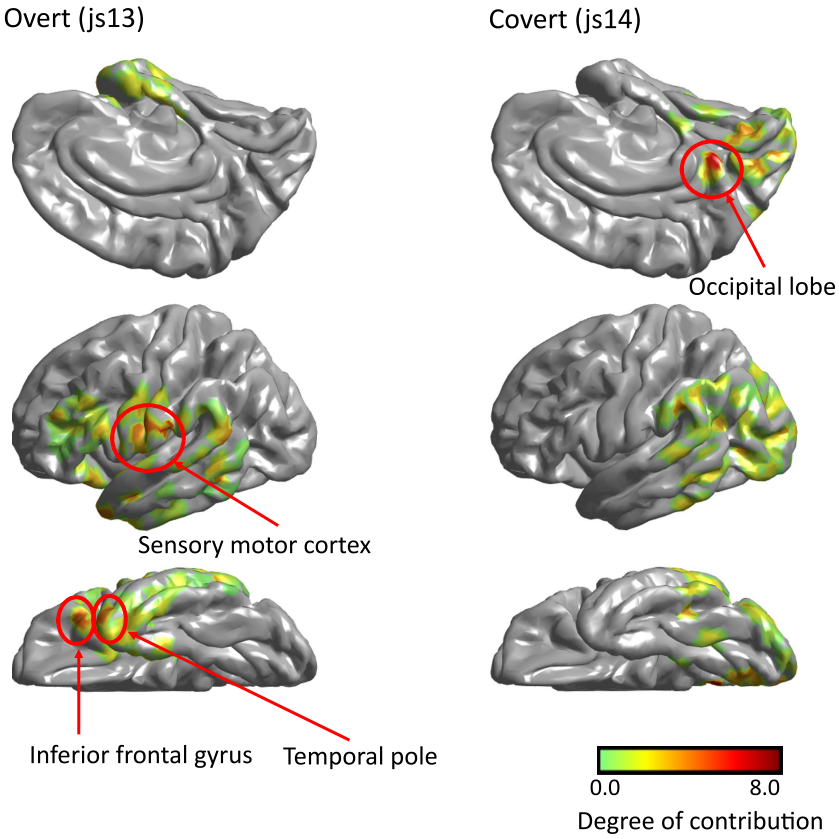

Figure 10. Electrode contributions for text decoding on averaged FreeSurfer pial surface image. The electrode contributions are illustrated using a color map, where green and red indicate the weak and strong contributions, respectively. The left shows the result of Overt (js13). Sensory motor cortex (SMC), inferior frontal gyrus, and temporal pole contribute. The right shows the result of Covert (js14). The occipital lobe contributes.

patterns of neural contributions that were consistent with those identified in our previous study on text decoding, suggesting a commonality in the brain regions involved in both speech and text decoding tasks.

These findings provide important insights into the neural decoding of imagined speech and highlight the potential of Transformer-based architectures for speech BCIs. In the following, we discuss the results from four aspects.

4.1. Comparison of Transformer and BLSTM

For the first aspect, the Transformer outperformed BLSTM in terms of MSE loss, Pearson correlation, and CER evaluated by Whisper. Previous studies by Komeiji et al. [16, 31] and Shigemi et al. [29] have demonstrated that the Transformer surpasses BLSTM for both text decoding and speech synthesis from ECoG signals. Consistent with these findings, long-range dependencies in the input ECoG signals may play crucial role in speech synthesis. Although Kohler et al. [27] achieved Pearson correlations of up to 0.8 using GRU, a type of RNN similar to BLSTM, for decoding Dutch sentences

from sEEG, replacing GRU with a Transformer may further improve result quality. Furthermore, Chen et al. [46] applied a Swin Transformer (a variant of Transformer) [47] to decode English speech from ECoG or sEEG and reported correlations exceeding 0.8. Therefore, it is anticipated that Transformer architectures will become the de facto standard for speech synthesis from ECoG or sEEG.

4.2. Ability to Generate High-quality Log-mel Spectrograms

For the second aspect, we analyzed the ability of the Transformer decoder to generate high-quality log-mel spectrograms. Interestingly, this capability was observed even when Gaussian noise was used as input. Further analysis revealed that positional encoding plays a crucial role in producing this high-quality output by providing temporal context to the input (Fig. 8). Even when Gaussian noise, which inherently lacks meaningful context, is used as input, the positional encoding enables the Transformer's output to maintain temporal structure. This suggests that the Transformer decoder may generate outputs that are somewhat independent of the specific characteristics of its input. Additionally, the quality of the synthesized speech appears to depend on the variability of the output signal patterns. Interestingly, synthesized speech from covert speech exhibited higher Pearson correlations than that from overt speech, despite the target audio signals for covert speech consisting of only eight variations, whereas overt speech involved up to 64. This implies that the model operates in a classification-like manner, effectively performing an eight-class classification task for covert speech. Nevertheless, we anticipate that output quality can be further improved by increasing the amount of training data, as demonstrated in domains such as image generation from noise vectors using techniques like stable diffusion. In the case of stable diffusion [48], the network model was trained on approximately 400 million images from the LAION-400M dataset [49]. Beyond speech generation from neural signals, stable diffusion technology has also been extended to estimate images from fMRI data [50]. These prior studies provide valuable insights into the impact of training data scale on the quality of generated outputs, which is particularly in relevant for the synthesis of speech from neural activity.

As aforementioned, the decoder behaved in a classification-like manner. This raises an important question: Does the decoder truly consider the ECoG signal patterns? To investigate this, we examine the results of the dictation test, which offer insight into whether the synthesized speech reflects meaningful neural patterns. For the dictation test using TERs, participant js13 in the overt task (TER: 18.3%) and participant js14 in the covert task (TER: 42.3%) demonstrated TERs significantly below the chance level of 50% (binomial test). These participants achieved high accuracy for text decoding in a previous study [31]. This suggests that the neural signals from js13 in the overt task and js14 in the covert task contained significant neural patterns relevant to speech content. The high-quality output with accurate content from covert speech synthesis supports our hypothesis that pre-trained neural vocoders can generate high-quality speech even

with limited ECoG training data. Moreover, it confirms that the synthesized speech was decoded based on information present in the ECoG signals.

In contrast, several participants exhibited TERs near the chance level of 50% during the dictation test. This indicates that while the synthesized speech quality was high, the accuracy of speech content was compromised. High-quality speech output may be attributed to the combined use of the Transformer decoder and pre-trained neural vocoder, along with the limited number of output variations. To improve content accuracy, additional training data will be necessary. In summary, while this study has largely succeeded in achieving high-quality speech output with limited training data, future work should emphasize synthesizing speech with precise and accurate content.

4.3. Electrode Contributions

Finally, in this study, we discuss anatomical analysis using saliency maps to better understand the regions that contribute to speech decoding. For participant js13 in the overt task, key regions included the sensorimotor cortex (SMC), inferior frontal gyrus, and temporal pole. In contrast, in the covert task for js14, the occipital lobe was identified as the main contributing area (see Fig. 10).

Interestingly, these regions correspond with those identified in our previous study on text decoding [31], where the SMC and temporal pole were significant for js13 in the overt task, and the occipital cortex was prominent for js14 in the covert task. This suggests that the same brain regions contribute to both speech and text decoding. In addition, js13 and js14 achieved the highest accuracy in both text decoding and the dictation tests for synthesized speech, further supporting the importance of these regions in decoding tasks. This insight may be crucial for improving speech synthesis accuracy, particularly in the selection of electrode placements to optimize performance.

5. Conclusion

This study demonstrated the successful synthesis of speech from neural activity during covert reading using ECoG signals. By employing a Transformer-based decoder in combination with a pre-trained neural vocoder, Parallel WaveGAN, we were able to generate high-quality audio despite the limited size of the training. Experiments involving ECoG signals from 13 participants yielded synthesized speech with a Pearson's correlation coefficients ranging from 0.85 to 0.95. These results highlight the effectiveness of the Transformer decoder in accurately reconstructing high-fidelity log-mel spectrograms from ECoG signals, even under data-constrained conditions. The ability to synthesize intelligible speech from covert reading underscores the potential of this approach for advanced speech BCI applications, particularly for enabling communication in individuals with speech impairments.

Acknowledgment

This work was partly supported 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] Gary M Oppenheim and Gary S Dell. Motor movement matters: The flexible abstractness of

inner speech. *Memory & cognition*, 38(8):1147–1160, 2010.

[33] Gary M Oppenheim and Gary S Dell. Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. *Cognition*, 106(1):528–537, 2008.

[34] Kai J Miller, Gerwin Schalk, Eberhard E Fetz, Marcel Den Nijs, Jeffrey G Ojemann, and Rajesh PN Rao. Cortical activity during motor execution, motor imagery, and imagery-based online feedback. *Proceedings of the National Academy of Sciences*, 107(9):4430–4435, 2010.

[35] 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.

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

[37] S. S. Stevens, J. Volkmann, and E. B. Newman. A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, 8(3):185–190, 1937.

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

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

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[49] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

[50] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models

from human brain activity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14453–14463, 2023.