Towards Voice Reconstruction from EEG during Imagined Speech

Young-Eun Lee[1*], **Seo-Hyun Lee**[1*], **Sang-Ho Kim**[2], **Seong-Whan Lee**[2†]

[1] Department of Brain and Cognitive Engineering, Korea University, Seoul, Republic of Korea
[2] Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea
{ye_lee, seohyunlee, sh_k, sw.lee}@korea.ac.kr

Abstract

Translating imagined speech from human brain activity into voice is a challenging and absorbing research issue that can provide new means of human communication via brain signals. Efforts to reconstruct speech from brain activity have shown their potential using invasive measures of spoken speech data, but have faced challenges in reconstructing imagined speech. In this paper, we propose NeuroTalk, which converts non-invasive brain signals of imagined speech into the user's own voice. Our model was trained with spoken speech EEG which was generalized to adapt to the domain of imagined speech, thus allowing natural correspondence between the imagined speech and the voice as a ground truth. In our framework, an automatic speech recognition decoder contributed to decomposing the phonemes of the generated speech, demonstrating the potential of voice reconstruction from unseen words. Our results imply the potential of speech synthesis from human EEG signals, not only from spoken speech but also from the brain signals of imagined speech.

Introduction

Brain signals contain various information related to human action or imagery, making them valuable materials for understanding human intentions. Brain-computer interface (BCI) is a technology that analyzes users' brain activity to derive external commands to control the environment through brain signals, therefore, can benefit paralyzed or locked-in patients (Chaudhary, Birbaumer, and Ramos-Murguialday 2016). Brain-to-speech (BTS) is a novel research stream in the field of BCI, which aims to directly synthesize audible speech from brain signals (Lee, Lee, and Lee 2022, 2020). While current studies on decoding human brain signals related to speech mainly focus on using spoken speech brain signals measured with invasive methods (Anumanchipalli, Chartier, and Chang 2019; Angrick et al. 2019; Herff et al. 2019), reconstructing imagined speech using non-invasive modalities is an intriguing topic that enables practical and silent communication via brain signals. However, due to the fundamental constraint of the imagined speech lacking the ground truth (GT) voice, it is challenging to synthesize the user's own voice from imagined speech brain signals.

As the potential of reconstructing speech from brain signals of spoken speech has been demonstrated, we expect that there must be relevant brain activation patterns that encode significant features of the speech. Imagined speech is known to resemble the neural patterns of spoken speech, which is mainly located on the ventral sensorimotor cortex (Watanabe et al. 2020; Si et al. 2021; Cooney, Folli, and Coyle 2018; Lee, Lee, and Lee 2019). If imagined speech has similar features to spoken speech, it may be feasible to establish a correlation between the brain signals of spoken speech, the audio of spoken speech, and the brain signals of imagined speech. Furthermore, if we could train and infer phonemes from imagined speech, several unseen words composed of already trained phonemes could also be reconstructed from the trained word sets.

In this study, we proposed a NeuroTalk framework that can correlate imagined speech electroencephalography (EEG) with spoken speech EEG and its corresponding audio, to reconstruct voice from imagined speech. The imagined utterances were decoded from the EEG signals to reconstruct the voice at the word level. Moreover, we estimated the possibility of reconstructing unseen words using the pre-trained model trained with only a few words, to potentially increase the degree of freedom using the model trained with a minimum number of words, including various phonemes. Based on our results, we aim to find the potential of speech reconstruction from imagined speech brain signals to the user's own voice. The main contributions are as follows:

Main Contribution

- We propose a generative model based on multi-receptive residual modules with recurrent neural networks that can extract frequency characteristics and sequential information from neural signals, to generate speech from non-invasive brain signals.
- The fundamental constraint of the imagined speech-based BTS system, which lacks the GT voice, has been addressed by the domain adaptation approach to link the imagined speech EEG, the spoken speech EEG, and the spoken speech audio.
- Unseen words have shown the potential to be recon-

[*] These authors contributed equally.
[†] Corresponding author

structed from the trained model for both spoken and imagined speech EEG. This implies that the model could learn the phoneme-level information from the brain signal, demonstrating the potential for robust speech generation by training only a few words or phrases.

Background

Speech-Related Paradigms

Speech-related paradigms mainly used in the BTS studies can be largely divided into three categories: spoken speech, mimed speech, and imagined speech. While spoken speech indicates the natural speech that accompanies vocal output and movement of the articulators, mimed speech does not produce vocal output but accompanies the movement of the mouth and tongue as if speaking out loud (Schultz et al. 2017). Imagined speech is the mode of internally imagining speech, accompanying both the imagery of the mouth movement and the vocal sound, without producing actual movement or voice (Cooney, Folli, and Coyle 2018).

Invasive Approach

Invasive measurements involve a surgical process of implementation inside the skull to capture brain activation directly from the cortex. Therefore, medical risks and difficulties exist to be applied for healthy users (Wang and Ji 2021). However, due to the high signal-to-noise ratio (SNR), many previous studies focused primarily on synthesizing speech from invasive brain signals. Studies using electrocorticography (Akbari et al. 2019; Anumanchipalli, Chartier, and Chang 2019; Angrick et al. 2019; Herff et al. 2019) and attempts to decode speech from deeper brain structures using stereotactic electroencephalography depth electrodes (Angrick et al. 2021a, 2022, 2021b; Herff, Krusienski, and Kubben 2020; Meng et al. 2022) have reported the possibility of speech reconstruction using spoken speech data.

Non-invasive Approach

Electroencephalography EEG is the most widely used non-invasive modality for practical use since it does not involve any surgical process and is relatively easy to access (Krishna et al. 2020). However, non-invasive measures have relatively low SNR and artifact problems compared to the invasive modalities, which makes it hard to extract the user's intention from brain signals (Graimann, Allison, and Pfurtscheller 2009).

Spoken speech-based BTS Speech reconstruction from spoken speech or mimed speech brain signals, kinematic or EMG data has shown potential (Gaddy and Klein 2020, 2021; Gonzalez et al. 2017). Nonetheless, relying solely on a spoken speech-based BTS system is not a definitive solution for the essential goal of BCI, as it is redundant when users can speak aloud. Additionally, it is not applicable to patients who are unable to speak or move.

Decoding imagined speech Current EEG-based imagined speech decoding technology has shown promising results in terms of classification problems (Saha, Abdul-Mageed, and Fels 2019; Wang et al. 2013; Lee et al. 2019b). However, these approaches are constrained to the basic classification of the predefined set of classes (Makin, Moses, and Chang 2020), therefore, reconstructing natural speech from imagined speech brain signals is crucial for intuitive and silent BCI communication.

Imagined speech-based BTS The fundamental constraint of speech reconstruction from EEG of imagined speech is the inferior SNR and the absence of vocal GT corresponding to the brain signals. Therefore, speech synthesis from imagined speech using non-invasive measures has not yielded convincing results (Proix et al. 2022). Attempts to reconstruct speech from invasive data during whispered and imagined speech have been made, but have reported relatively inferior performance even when invasive measures were utilized (Angrick et al. 2021b). Speech synthesis from imagined speech could be the key to a new era of human communication, moving from current voice or text-based communication to brain-based communication. It may also be a technology that can help patients who are unable to speak or those who may lose their voice in the future.

Method

In this section, we describe the model frameworks used in this paper, including generator, discriminator, vocoder, and automatic speech recognition (ASR), as well as losses including reconstruction loss, generative adversarial network (GAN) loss, and connectionist temporal classification (CTC), as shown in Figure 1. The collected brain signals of spoken speech and imagined speech are represented as feature embeddings to extract the optimal features from brain signals. The generator applying GAN (Goodfellow et al. 2014) reconstructs a mel-spectrogram to match the target voice during spoken speech. The reconstruction loss for the generator is determined as the difference between the reconstructed mel-spectrogram from the EEG signals and the GT mel-spectrogram during spoken speech. The discriminator classifies the validity of whether the input samples of the mel-spectrogram are real or fake, and calculates an adversarial loss for the generator and discriminator. The pre-trained vocoder synthesizes the mel-spectrogram into a reconstructed voice. The pre-trained ASR model transforms the reconstructed voice into text and calculates the CTC loss for the generator.

Since imagined speech has no reference voice, voice samples during the spoken speech, which is recorded in the same sequence as imagined speech, were used as the target audio. To match the EEG to the voice of spoken speech, dynamic time warping (DTW) was applied between the reconstructed mel-spectrogram from EEG and the mel-spectrogram of voice during spoken speech. Furthermore, domain adaptation (DA) was conducted to transfer the architecture of spoken speech to that of imagined speech.

Architectures

Embedding vector It is known that spatial, temporal, and spectral information are all important for speech-related brain signals, and vector-based brain embedding features can represent the contextual meaning in brain signals (Gold-

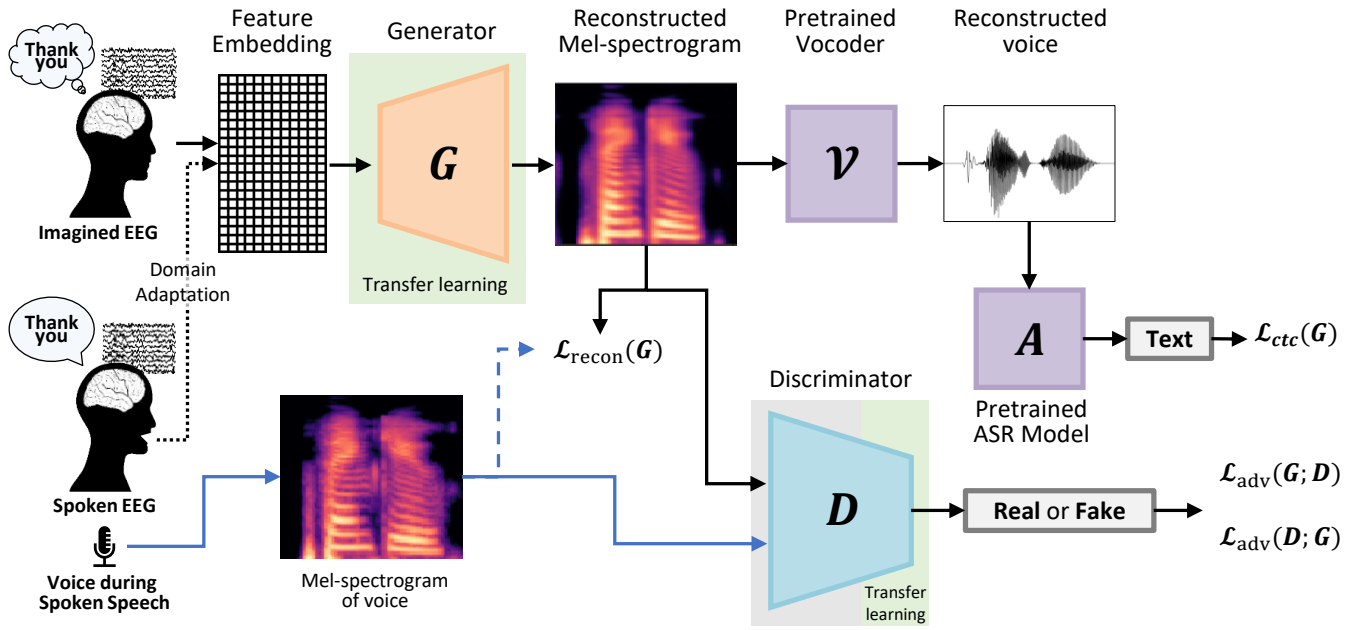

Figure 1: Overall framework of this study. Imagined speech EEG was given as the input to reconstruct the corresponding audio of the imagined word or phrase with the user's own voice. G refers to the generator, which generates the mel-spectrogram from the embedding vector. D is the discriminator, which distinguishes the validity of the input. At the bottom, the two models, a pre-trained vocoder V and a pre-trained ASR model A, generate text from the mel-spectrogram.

stein et al. 2022; Lee, Lee, and Lee 2020). The embedding vector was generated using a common spatial pattern (CSP) to maximize spatial patterns and log-variance to extract temporal oscillation patterns. CSP finds the optimal spatial filters using covariance matrices (Devlaminck et al. 2011), and helps to decode the brain signals related to speech (Nguyen, Karavas, and Artemiadis 2017).

To reduce the difference between the data distribution of spoken EEG and imagined EEG, CSP filters were shared with both EEG signals. CSP filters were trained with imagined EEG, which contains relatively pure brain signals, rather than spoken EEG which may contain some noise. By sharing the CSP filters, the spoken EEG domain could be adapted to the subspace of imagined EEG.

The CSP filters were trained with eight CSP features and sixteen segments without overlap using the training dataset. Each trial of EEG signals has a size of time point $\times$ channels (5,000 $\times$ 64). After applying CSP, the embedding vector, transformed from EEG signals, has 104 features $\times$ 16-time segments, where the features consist of 13 classes $\times$ 8 CSP features.

Generator The main architecture of the proposed generator consists of gated recurrent units (GRU) (Cho et al. 2014) to capture the sequence information and several residual blocks to capture the temporal and spatial information, preventing vanishing gradient problems. Figure 2a describe the generator in detail. The input of the generator is given as the embedding vector of EEG signals and the output is generated as a mel-spectrogram. The embedding vector goes through a pre-convolutional layer consisting of 1d convolution and concatenates the features from the bi-directional GRU to extract the sequence features. After that, the generator upsamples it N times using transposed convolution with a stride of two or three and the multi-receptive field fusion (MRF) module, which is the sum of the outputs of multiple residual blocks with different kernel sizes. Finally, a post-convolutional layer and activation function are applied.

Discriminator The discriminator is similarly composed in the opposite direction to the generator, as described in Figure 2b. The input of the discriminator is the mel-spectrogram and the output is the validity of real/fake voice. Moreover, the discriminator was also trained using class information from only voice data.

Vocoder and ASR Vocoder and ASR models are used to clarify the reconstructed voice from brain signals by translating it into text. Vocoder is a category of speech synthesis technology converting intermediate representation such as mel-spectrogram to waveform audio. To adjust our framework for a real-world BTS system, we applied a pre-trained HiFi-GAN (Kong, Kim, and Bae 2020) which is a high-quality vocoder with fast inference speed. The same architecture and hyperparameters were applied with the pre-trained model 'Universal ver.1'.

ASR converts human speech waveform into written texts, which can represent the speech as a contextual sequence of discrete units (Baevski et al. 2020). The ASR is composed of pre-trained HuBERT (Hsu et al. 2021) with a large configuration, which is a self-supervised learning model of speech representations trained with the Libri-Light dataset and fine-tuned with the LibriSpeech dataset.

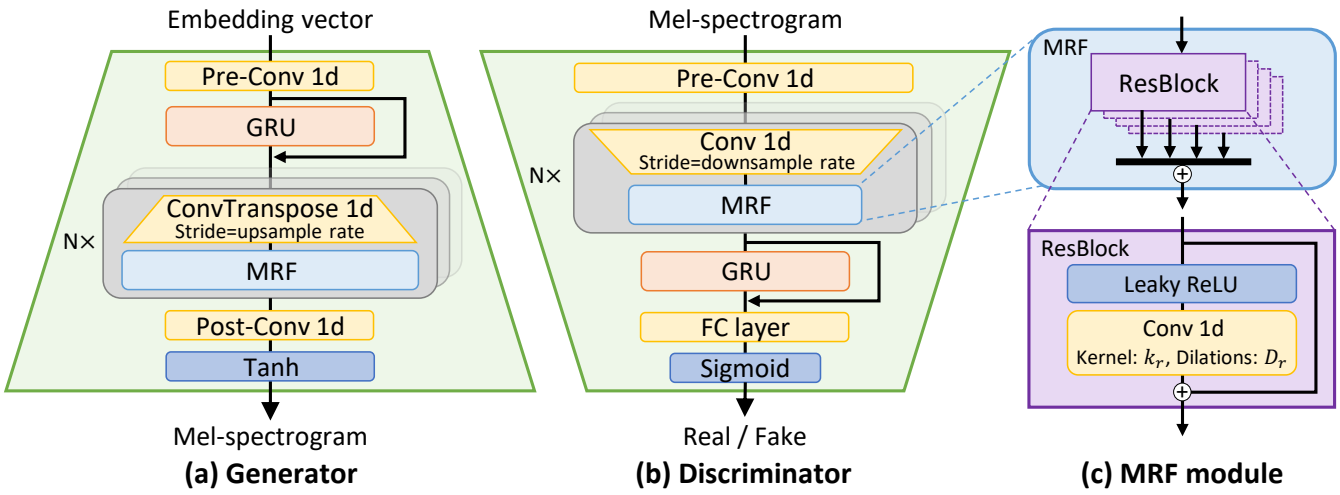

Figure 2: The architecture details in (a) generator, (b) discriminator, and (c) MRF module. The MRF modules in both the generator and the discriminator were repeated three times in our experiment. k_r indicates the kernel size of the residual block and D_r indicates the dilation rates of the residual block.

Training Loss Term

This section describes training losses: reconstruction loss L_{rec}, GAN loss L_{adv}, and CTC loss L_{ctc}. The generator is trained with reconstruction, adversarial, and CTC loss, while the discriminator uses an adversarial loss.

$$L(G) = \lambda_{g1} L_{rec}(G) + \lambda_{g2} L_{adv}(G; D) + \lambda_{g3} L_{ctc}(G) \quad (1)$$

$$L(D) = \lambda_d L_{adv}(D; G) \quad (2)$$

where loss coefficients are referred to λ_{g1-3} for the generator and λ_d for the discriminator.

Reconstruction loss Reconstruction loss has been verified in many studies (Kong, Kim, and Bae 2020; Isola et al. 2017), which can help improve the efficiency of the generator and the fidelity of reconstructed data.

$$L_{rec}(G) = E_s[(G(s) - x)^2] \quad (3)$$

where s refers to the input of the generator such as an embedding vector from EEG signals, and x refers to the input of the discriminator such as a mel-spectrogram.

GAN loss To reconstruct the mel-spectrogram to follow the real one, adversarial GAN loss L_{adv} was conducted on the generator G and discriminator D as follows.

$$L_{adv}(G; D) = E_s[log(1 - D(G(s)))] \quad (4)$$

$$L_{adv}(D; G) = E_{(x,s)}[log(1 - D(x)) + log(D(G(s)))] \quad (5)$$

where s refers to the input of generators such as embedding vector from EEG signals and x refers to the input of discriminator such as mel-spectrogram.

CTC loss CTC loss is a common performance metric for automatic speech recognition systems (Graves et al. 2006). CTC loss L_{ctc} allows to train the model using sequential data without the alignment information. CTC loss was primarily given to guide the prediction of characters and phonemes, to enhance the performance of unseen classes.

Domain Adaptation

The DA strategy was employed to resolve the fundamental constraint of speech reconstruction from imagined speech. Since imagined speech does not accompany the movement of the articulators, it is relatively reliable in terms of artifacts accompanied by the mouth movement and the vibration. However, since the GT audio for imagined speech does not exist, we designed a framework that adapts the domain of imagined speech from spoken speech, to exploit the natural correspondence of imagined EEG and the voice of spoken speech. The DA process was performed in two steps; 1) sharing the covariance matrix between imagined EEG and spoken EEG by applying the CSP filter of imagined speech and 2) applying transfer learning for the generator and discriminator from the trained model of spoken EEG.

Sharing subspace The CSP weights, trained with a training set (60%) of imagined EEG, were shared to generate embedding vectors. Sharing the CSP filters computed from imagined EEG allows the latent space of spoken EEG to be shifted into a comparable feature space of imagined EEG. Unlike most DA approaches that involve applying the weak domain to the well-trained classifier, we elected to the contrary, bringing the spoken speech feature space to that of the imagined speech. In that case, we could achieve a clear pattern more from the neural characteristics of speech rather than the movement artifacts or vibration artifacts.

Transfer learning The model was trained with a training set of spoken EEG, and then fine-tuned with a training set of imagined EEG at a smaller learning rate than the case of spoken EEG. This was to connect with the voice recordings of spoken speech, which acts as the GT of imagined speech. The trained model from spoken EEG can assist in training the model of imagined EEG that has relatively insufficient information, therefore, the spoken EEG could guide learning from the weak features of imagined EEG.

Experimental Setup

Dataset

Participants Six participants volunteered in the study. The study was conducted in accordance with the Declaration of Helsinki, approved by the Korea University Institutional Review Board [KUIRB-2019-0143-01]. Informed consent was obtained from all participants. Since the dataset contains human-derived biosignals, only a small sample dataset could be published to reproduce and execute code.

Paradigms For the spoken speech session, participants were instructed to naturally pronounce the randomly given thirteen phases, provided as an auditory cue of twelve words/phrases (ambulance, clock, hello, help me, light, pain, stop, thank you, toilet, TV, water, and yes) and a silent phase. Speech data were recorded in a rhythmic manner to avoid any visual or auditory disruptions. The imagined speech data was collected in the exactly same manner as the spoken speech, following the previous study (Lee, Lee, and Lee 2020). A hundred trials of both spoken speech and imagined speech per class were collected for each participant. Therefore, each participant had 1,300 trials for the spoken and imagined speech paradigm.

Recording The dataset used in this study consists of scalp EEG recordings of spoken/imagined speech and voice recordings of spoken speech. During the experiment, EEG signals were recorded at the sampling rate of 2,500Hz via Brain Vision/Recorder (BrainProduct GmbH, Germany), and the corresponding audio was simultaneously recorded at the sampling rate of 8,000Hz. Brain signals were recorded with a 64-channel EEG cap with active Ag/AgCl electrode placement following the international 10-10 system.

Pre-processing

EEG signals were extracted in 2-second intervals for each trial. The data was filtered with a 5th-order Butterworth bandpass filter in the high-frequency range of 30–120 Hz which is well-known to contain speech-related information (Lachaux et al. 2012; Lee, Lee, and Lee 2020). A notch filter was used to remove the line noise at 60 Hz with harmonics of 120 Hz. The electrooculography (EOG) and electromyography(EMG) of spoken speech were removed using blind source separation referencing (Gómez-Herrero et al. 2006). The baseline was corrected by subtracting the average value of 500 ms before each trial. Pre-processing procedures were performed in Python and Matlab using OpenBMI Toolbox (Lee et al. 2019a), BBCI Toolbox (Krepki et al. 2007), and EEGLAB (Delorme and Makeig 2004). For the voice data, we resampled the voice signals to 22,050 Hz and reduced the noise using the noise reduction library (Sainburg, Thielk, and Gentner 2020).

Dataset Composition and Training Procedure

Imagined speech lacks a reference voice to train a model. However, spoken speech provides both audio and EEG data in a perfectly time-aligned pair. Since the experimental design of imagined speech and spoken speech was completely identical, the voice of the identical sequence of spoken speech for each participant was used as the reference voice for the imagined speech. Moreover, the transfer learning approach was applied with the model trained on spoken speech EEG and spoken speech audio to imagined speech EEG.

The dataset was divided into 5-fold subsets in training, validation, and test dataset according to the random selection with a random seed. One unseen word, 'stop' was separated from the dataset, and was not included in the training set. It was chosen to test unseen cases since every phoneme composing the word 'stop' was covered with the remaining 11 words used for the training. That is, we trained the 11 words/phrases and a silent phase as a training dataset and validated 12 words/phrases and a silent phase in the validation and test dataset including the unseen word.

Model Implementation Details

The generator had three residual blocks with the kernel size of 3, 7, and 11, each dilation of 1, 3, and 5, and upsampling rate of 3, 2, and 2 with twice upsample kernel size. The number of the initial channel was 1,024, and the directional GRU dimension was half of the initial channel. The discriminator had the same residual block as the generator, but a downsampling rate of 3, 3, and 3 with twice the kernel size. The number of the final channel was 64, and the directional GRU dimension was half of the final channel. The mel-spectrogram was managed in a sampling rate of 22,050 Hz and the STFT and mel function was conducted with nFFT of 1,024, the window of 1,024, hop size of 256, and 80 bands of mel-spectrogram. Initial training was conducted with an initial learning rate of 10^{-4}, and the fine-tuning was conducted with a lower learning rate such as 10^{-5} in the maximum epoch of 500 and a batch size of 10. We trained the model on an NVIDIA GeForce RTX 3090 GPU. We used AdamW optimizer (Loshchilov and Hutter 2017) with searched parameters of β_1=0.8, β_2=0.99, and weight decay λ=0.01, which was scheduled by 0.999 factor in every epoch. We released the source code and sample data on GitHub at: https://github.com/youngeun1209/NeuroTalk.

Evaluation Metrics

For the evaluation metrics, we used root mean square error (RMSE), character error rate (CER), and a subjective mean opinion score (MOS) test. To evaluate the accurate reconstructing performance of the generator, we computed the RMSE between the target and reconstructed mel-spectrogram. To evaluate the clarity quantitatively, we computed CER after going through the ASR model. For the subjective evaluation, a MOS test was conducted to evaluate the quality of the reconstructed speech. We randomly selected 125 samples of voice from a test dataset. The samples were evaluated by more than 20 raters on a scale of 1-5 with 0.5-point increments. We compared the performance of generation with the voice of GT and the converted voice using mel-transform and ASR from the voice of GT. Moreover, to demonstrate the extension of NeuroTalk, we evaluated the generation performance of unseen word which is composed of phonemes that were contained in the trained word classes.

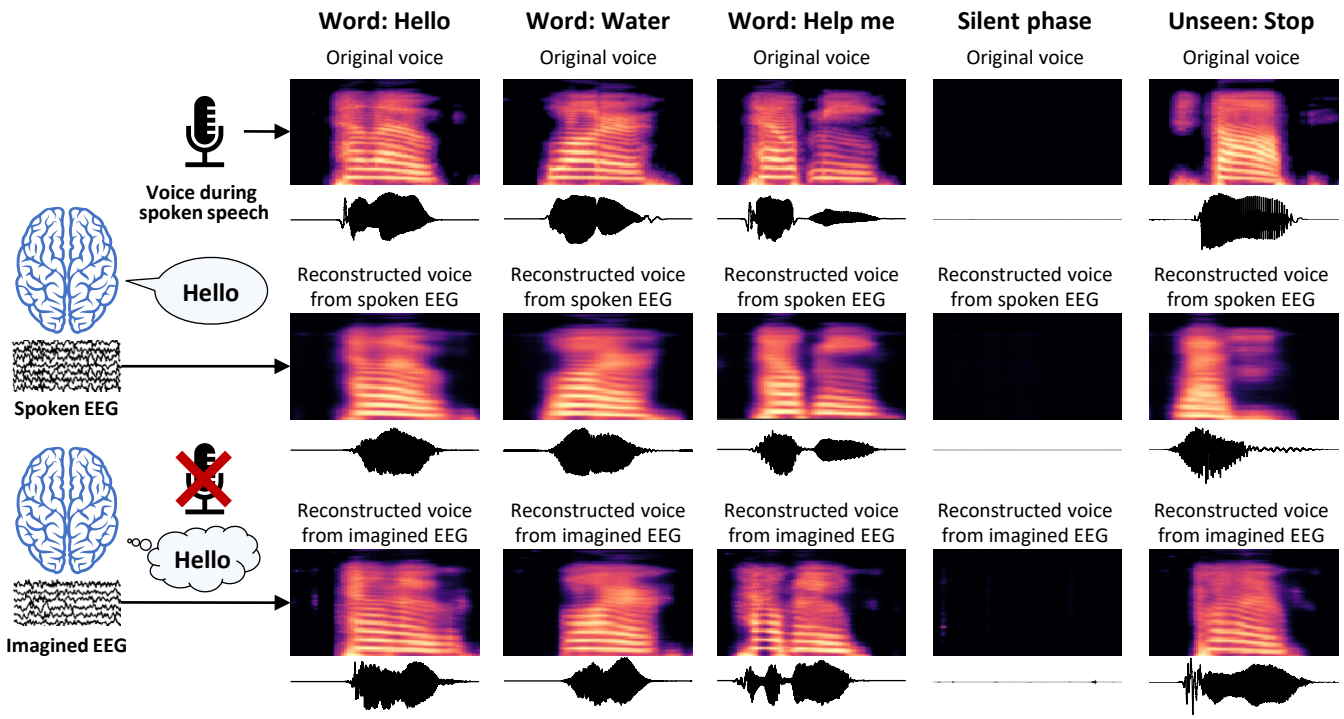

Figure 3: Mel-spectrogram and the audio wave of the original voice, voice reconstructed from EEG. Three examples of reconstruction include 'Hello', 'Water', and 'Help me'. Silent phases for both spoken and imagined EEG were successfully decoded. Unseen cases were also reconstructed despite their inferior performance.

Results and Discussion

Voice Reconstruction from EEG

The audio samples are included in the demo page at: https://neurotalk.github.io/demo. Figure 3 displays the mel-spectrogram and the audio wave samples of the original voice, reconstructed voice from spoken speech EEG, and reconstructed voice from imagined speech EEG. As shown in the figure, successfully reconstructed cases display similar patterns of the mel-spectrogram and the audio waveform with the original voice. Table 1 shows the evaluation results of the reconstructed voice from brain signals compared to the GT. MOS of spoken speech cases had no large difference from GT, which shows that the model can generate natural speech from spoken EEG. Objective measures of RMSE and CER have shown inferior performance in the case of imagined speech EEG compared to that of spoken speech EEG.

As depicted in Figure 3, the test samples of the silent phases were successfully reconstructed without any activation. Silent cases for both spoken and imagined EEG were successfully decoded except for only one case of imagined speech. According to this result, we can infer that our model accurately learned the silence interval and can detect the speech onset from both spoken and imagined speech EEG. Although imagined speech does not have GT voice, the results show that the proposed NeuroTalk framework effectively adapts the spoken speech-based model to the imagined speech EEG to decode the user's silent intention from brain signals and generate voice.

There were some instances of failure for the imagined speech case, as shown in Figure 4. In the case of synthesizing 'thank you', the main distinction between the success and failure cases was whether it detects the silence intervals within a phrase. As shown in Figure 4, an instance with a CER of 50% exhibited only a small silence interval between the words 'thank' and 'you'. Moreover, the instance of failure with a CER of 100% generated only few characters that cannot represent any of the syllables found in the GT.

Voice Reconstruction of Unseen Words

According to Figure 3 and Table 1, the unseen cases have shown CER of 78.9% and 83.1% for spoken and imagined speech, respectively. Although the word was not perfectly reconstructed, the model generated fairly high-quality audio with MOS over 2.5. The gap between the CER of spoken and imagined EEG was relatively small compared to the trained words. Although it still could be further improved, our result demonstrates that the NeuroTalk model has the potential to extend the degree of freedom of decodable words or sentences by training on the word-level dataset. We expect that CTC loss could learn the character or phoneme information of words even from brain signals, which contain human intention and phonetic information. Since we trained the model with a limited set of words/phrases, it may be simply classifying the EEG as one of the training classes. However, the model has shown the potential to generate the unseen word outside of the training set indicating the possibility of generalization and extension to the classes outside

Model	RMSE	CER (%)	MOS
GT	-	18.4 (±11.5)	3.67 (±1.0)
GT (trans.)	-	23.4 (±10.9)	3.68 (±0.9)
SpEEG	0.17 (±0.02)	40.2 (±13.5)	3.34 (±1.0)
ImEEG	0.18 (±0.03)	68.3 (±2.5)	2.78 (±1.1)
Unseen SpEEG	0.19 (±0.03)	78.9 (±7.4)	2.87 (±1.1)
Unseen ImEEG	0.19 (±0.03)	83.1 (±14.5)	2.57 (±1.2)

Table 1: Quantitative and qualitative evaluation of spoken and imagined EEG. GT (trans.) indicates the transformed voice from GT via mel-transform and vocoder. SpEEG and ImEEG refer to spoken and imagined EEG, respectively.

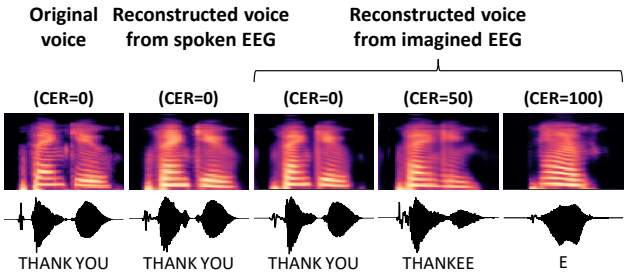

Figure 4: Success and failure cases. Mel-spectrogram and waveform were shown for original and reconstructed voices.

of the training set. Since our model generates speech rather than classifying into predefined classes, we expect that despite the inferior performance or somewhat mumbled voice, human listeners may be able to recognize the user's intention given the context in the real-world application. Our current results are in the preliminary stage, but we plan to increase the number of training vocabularies for future work.

Ablation Study

We performed an ablation study of GRU in the generator and discriminator, losses of GAN, reconstruction, and CTC to verify the effect of each module and loss on the model performance. As demonstrated in Table 2, CER of all cases were mostly inferior to the baseline, indicating that all approaches perform their roles. Specifically, the performance was notably compromised when reconstruction loss was omitted, followed by a slightly improved but still suboptimal performance without CTC loss. This suggests that the reconstruction loss followed by CTC loss has the most significant impact on the overall performance of the framework. In the subjective evaluation, the absence of GRU resulted in the poorest naturalness, emphasizing the importance of sequential features for natural speech synthesis.

Domain Adaptation

DA was performed by sharing the CSP subspaces and transferring the spoken speech-based trained model to imagined speech EEG. As shown in Table 2, the CER without DA has shown inferior performance compared to the baseline. This implies that the information from spoken speech EEG was

Input	RMSE	CER (%)	MOS
Baseline	0.18 (±0.03)	68.3 (±2.5)	2.78 (±1.1)
w/o GRU	0.19 (±0.03)	76.1 (±3.3)	2.18 (±1.2)
w/o GAN loss	0.18 (±0.02)	76.1 (±2.3)	2.86 (±1.2)
w/o recon. loss	0.62 (±0.12)	80.2 (±8.0)	2.50 (±1.3)
w/o CTC loss	0.39 (±0.07)	76.9 (±0.3)	2.52 (±1.2)
w/o DA	0.18 (±0.03)	72.3 (±1.7)	2.66 (±1.2)

Table 2: Ablation study of imagined EEG including performance without GRU module, GAN loss, reconstruction loss, CTC loss, and DA approach.

useful for training imagined speech EEG, which means the neural substrates of imagined and spoken speech have common features that can be represented in our embedding vector. Speech production and articulation are mainly known to be associated with the left sensory-motor and inferior frontal cortices (Proix et al. 2022). Angular gyrus functions to associate various language-related activation from the auditory, motor, and sensory regions, therefore, not only the left temporal lobe but the whole brain may function in the speech process (Watanabe et al. 2020). Our embedding vector, generated from the whole channel EEG, may contain both the articulatory information and the speech intention. Therefore, we demonstrate the potential of generating speech by extracting informative speech-related features, represented by the shared features of spoken and imagined speech EEG.

Leave-One-Out Scenario

For a more comprehensive discussion, we conducted an additional experiment of the leave-one-out (LOO) approach to further apply our NeuroTalk system to locked-in patients who can only use imagined speech. The model was trained with the spoken EEG of entire participants excluding one participant and was fine-tuned with the imagined EEG of the excluded participant. As a result, comparable performance inferior to the baseline but better than without DA was acquired. Based on this preliminary experiment, we have identified the possibility of extending our framework to an entirely new person, which could offer additional support to individuals who have lost their own voice.

Conclusion

We presented NeuroTalk, which reconstructs the user's own voice from the EEG during imagined speech. The DA approach was carried out through joint feature embedding and transfer-learning the models of imagined speech EEG, using the pre-trained models of spoken speech EEG. Our results demonstrate the feasibility of reconstructing voice from non-invasive brain signals of imagined speech at the word level. Furthermore, the generation of the unseen word with multiple characters, despite the inferior performance, indicates the potential to extend our study to larger datasets and to explore sentence-level speech synthesis in the future. We hope our study contributes to advancing the means of human communication and further provides increased freedom of communication for patients or people with disabilities.

Acknowledgements

This work was supported by Institute for Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) (No.2021-0-02068, Artificial Intelligence Innovation Hub; No. 2017-0-00451, Development of BCI based Brain and Cognitive Computing Technology for Recognizing User's Intentions using Deep Learning; No. 2019-0-00079, Artificial Intelligence Graduate School Program(Korea University)).

References

Akbari, H.; Khalighinejad, B.; Herrero, J. L.; Mehta, A. D.; and Mesgarani, N. 2019. Towards reconstructing intelligible speech from the human auditory cortex. *Scientific Reports*, 9(1): 1–12.

Angrick, M.; Herff, C.; Mugler, E.; Tate, M. C.; Slutzky, M. W.; Krusienski, D. J.; and Schultz, T. 2019. Speech synthesis from ECoG using densely connected 3D convolutional neural networks. *Journal of Neural Engineering*, 16(3): 036019.

Angrick, M.; Ottenhoff, M.; Diener, L.; Ivucic, D.; Ivucic, G.; Goulis, S.; Colon, A. J.; Wagner, L.; Krusienski, D. J.; Kubben, P. L.; et al. 2022. Towards closed-loop speech synthesis from stereotactic EEG: A unit delection approach. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1296–1300. IEEE.

Angrick, M.; Ottenhoff, M.; Goulis, S.; Colon, A. J.; Wagner, L.; Krusienski, D. J.; Kubben, P. L.; Schultz, T.; and Herff, C. 2021a. Speech synthesis from stereotactic EEG using an electrode shaft dependent multi-input convolutional neural network approach. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 6045–6048. IEEE.

Angrick, M.; Ottenhoff, M. C.; Diener, L.; Ivucic, D.; Ivucic, G.; Goulis, S.; Saal, J.; Colon, A. J.; Wagner, L.; Krusienski, D. J.; et al. 2021b. Real-time synthesis of imagined speech processes from minimally invasive recordings of neural activity. *Communications Biology*, 4(1): 1–10.

Anumanchipalli, G. K.; Chartier, J.; and Chang, E. F. 2019. Speech synthesis from neural decoding of spoken sentences. *Nature*, 568(7753): 493–498.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33: 12449–12460.

Chaudhary, U.; Birbaumer, N.; and Ramos-Murguialday, A. 2016. Brain–computer interfaces for communication and rehabilitation. *Nature Reviews Neurology*, 12(9): 513–525.

Cho, K.; Van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Cooney, C.; Folli, R.; and Coyle, D. 2018. Neurolinguistics research advancing development of a direct-speech brain-computer interface. *IScience*, 8: 103–125.

Delorme, A.; and Makeig, S. 2004. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1): 9–21.

Devlaminck, D.; Wyns, B.; Grosse-Wentrup, M.; Otte, G.; and Santens, P. 2011. Multisubject learning for common spatial patterns in motor-imagery BCI. *Computational Intelligence and Neuroscience*, 2011.

Gaddy, D.; and Klein, D. 2020. Digital voicing of silent speech. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5521–5530.

Gaddy, D.; and Klein, D. 2021. An improved model for voicing silent speech. *arXiv preprint arXiv:2106.01933*.

Goldstein, A.; Zada, Z.; Buchnik, E.; Schain, M.; Price, A.; Aubrey, B.; Nastase, S. A.; Feder, A.; Emanuel, D.; Cohen, A.; et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3): 369–380.

Gómez-Herrero, G.; De Clercq, W.; Anwar, H.; Kara, O.; Egiazarian, K.; Van Huffel, S.; and Van Paesschen, W. 2006. Automatic removal of ocular artifacts in the EEG without an EOG reference channel. In *Proceedings of the 7th Nordic Signal Processing Symposium*, 130–133. IEEE.

Gonzalez, J. A.; Cheah, L. A.; Gomez, A. M.; Green, P. D.; Gilbert, J. M.; Ell, S. R.; Moore, R. K.; and Holdsworth, E. 2017. Direct speech reconstruction from articulatory sensor data by machine learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12): 2362–2374.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Graimann, B.; Allison, B.; and Pfurtscheller, G. 2009. Brain–computer interfaces: A gentle introduction. In *Brain-Computer Interfaces*, 1–27. Springer.

Graves, A.; Fernández, S.; Gomez, F.; and Schmidhuber, J. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, 369–376.

Herff, C.; Diener, L.; Angrick, M.; Mugler, E.; Tate, M. C.; Goldrick, M. A.; Krusienski, D. J.; Slutzky, M. W.; and Schultz, T. 2019. Generating natural, intelligible speech from brain activity in motor, premotor, and inferior frontal cortices. *Frontiers in Neuroscience*, 13: 1267.

Herff, C.; Krusienski, D. J.; and Kubben, P. 2020. The potential of stereotactic-EEG for brain-computer interfaces: current progress and future directions. *Frontiers in Neuroscience*, 14: 123.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3451–3460.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1125–1134.

Kong, J.; Kim, J.; and Bae, J. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33: 17022–17033.

Krepki, R.; Blankertz, B.; Curio, G.; and Müller, K.-R. 2007. The Berlin Brain-Computer Interface (BBCI)–towards a new communication channel for online control in gaming applications. *Multimedia Tools and Applications*, 33(1): 73–90.

Krishna, G.; Tran, C.; Han, Y.; Carnahan, M.; and Tewfik, A. H. 2020. Speech synthesis using EEG. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1235–1238. IEEE.

Lachaux, J.-P.; Axmacher, N.; Mormann, F.; Halgren, E.; and Crone, N. E. 2012. High-frequency neural activity and human cognition: past, present and possible future of intracranial EEG research. *Progress in Neurobiology*, 98(3): 279–301.

Lee, M.-H.; Kwon, O.-Y.; Kim, Y.-J.; Kim, H.-K.; Lee, Y.-E.; Williamson, J.; Fazli, S.; and Lee, S.-W. 2019a. EEG dataset and OpenBMI toolbox for three BCI paradigms: an investigation into BCI illiteracy. *GigaScience*, 8(5): giz002.

Lee, S.-H.; Lee, M.; Jeong, J.-H.; and Lee, S.-W. 2019b. Towards an EEG-based intuitive BCI communication system using imagined speech and visual imagery. In *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 4409–4414. IEEE.

Lee, S.-H.; Lee, M.; and Lee, S.-W. 2019. EEG representations of spatial and temporal features in imagined speech and overt speech. In *Asian Conference on Pattern Recognition*, 387–400.

Lee, S.-H.; Lee, M.; and Lee, S.-W. 2020. Neural decoding of imagined speech and visual imagery as intuitive paradigms for BCI communication. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.

Lee, S.-H.; Lee, Y.-E.; and Lee, S.-W. 2022. Toward imagined speech based smart communication system: potential applications on metaverse conditions. In *2022 10th International Winter Conference on Brain-Computer Interface (BCI)*, 1–4. IEEE.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Makin, J. G.; Moses, D. A.; and Chang, E. F. 2020. Machine translation of cortical activity to text with an encoder–decoder framework. *Nature Neuroscience*, 23(4): 575–582.

Meng, K.; Lee, S.-H.; Goodarzy, F.; Vogrin, S.; Cook, M. J.; Lee, S.-W.; and Grayden, D. B. 2022. Evidence of Onset and Sustained Neural Responses to Isolated Phonemes from Intracranial Recordings in a Voice-based Cursor Control Task. In *Proceedings of the Annual Conference of the International Speech Communication Association (Interspeech)*, 4063–4067.

Nguyen, C. H.; Karavas, G. K.; and Artemiadis, P. 2017. Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features. *Journal of Neural Engineering*, 15(1): 016002.

Proix, T.; Delgado Saa, J.; Christen, A.; Martin, S.; Pasley, B. N.; Knight, R. T.; Tian, X.; Poeppel, D.; Doyle, W. K.; Devinsky, O.; et al. 2022. Imagined speech can be decoded from low-and cross-frequency intracranial EEG features. *Nature Communications*, 13(1): 1–14.

Saha, P.; Abdul-Mageed, M.; and Fels, S. 2019. SPEAK YOUR MIND! Towards imagined speech recognition with hierarchical deep learning. *Proc. Interspeech 2019*, 141–145.

Sainburg, T.; Thielk, M.; and Gentner, T. Q. 2020. Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires. *PLoS Computational Biology*, 16(10): e1008228.

Schultz, T.; Wand, M.; Hueber, T.; Krusienski, D. J.; Herff, C.; and Brumberg, J. S. 2017. Biosignal-based spoken communication: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12): 2257–2271.

Si, X.; Li, S.; Xiang, S.; Yu, J.; and Ming, D. 2021. Imagined speech increases the hemodynamic response and functional connectivity of the dorsal motor cortex. *Journal of Neural Engineering*, 18(5): 056048.

Wang, L.; Zhang, X.; Zhong, X.; and Zhang, Y. 2013. Analysis and classification of speech imagery EEG for BCI. *Biomedical Signal Processing and Control*, 8(6): 901–908.

Wang, Z.; and Ji, H. 2021. Open vocabulary Electroencephalography-to-text decoding and zero-shot sentiment classification. *arXiv preprint arXiv:2112.02690*.

Watanabe, H.; Tanaka, H.; Sakti, S.; and Nakamura, S. 2020. Synchronization between overt speech envelope and EEG oscillations during imagined speech. *Neuroscience Research*, 153: 48–55.