NeuralSpeak: Non-invasive Brain-to-Speech Synthesis with Language models

Anonymous ACL submission

Abstract

Speech Synthesis from non-invasive brain activities offers a promising avenue for restoring communication abilities in patients with neurological disorders. Significant progress has been made in reconstructing natural speech from invasive brain recordings; however, these methods face practical challenges such as the high risk associated with brain surgery and the difficulties encountered in maintaining such devices over time. In this work, we formulate the task of non-invasive brain-to-speech synthesis and propose *NeuralSpeak* tailored for this task, Specifically, we 1) leverage a multi-scale transformer model to address the challenges of handling excessively long sequences caused by the residual vector quantization-based neural codec in tokenization; 2) introduce a multi-window fMRI encoder, trained with contrastive learning to produce brain-derived embeddings that align closely with semantically rich text representations. *NeuralSpeak* achieves state-of-the-art results in both objective and subjective benchmark evaluation. Furthermore, we provide evidence that our model is biologically plausible and interpretable, mirroring established physiological processes.[1]

1 Introduction

Neurological disorders, such as stroke, brain tumors, and traumatic brain injury, often impair patients' communication abilities, making it crucial to find alternative ways for them to interact with their surroundings. Many patients rely on assistive communication devices that interpret nonverbal cues like residual head or eye movements, or utilize brain-computer interfaces (BCIs) to select letters and form words. While BCIs hold promise for restoring communicative functions (Owen et al., 2006; Claassen et al., 2019; King et al., 2013), their performance significantly lags behind the natural speech rate of about 150 words per minute. For instance, studies by Moses et al. (2021) have reported decoding rates of merely 15.2 words per minute with BCIs implanted in the sensorimotor cortex. Similarly, Metzger et al. (2022) have achieved a typing speed of 29.4 characters per minute using a similar BCI setup, presenting a potential alternative communication avenue for individuals with neurological impairments.

To approximate a more natural communication experience, researchers have turned to directly synthesizing speech from brain activity. Several investigations have utilized invasive techniques to decode verbal speech directly from neural activity. For instance, Anumanchipalli et al. (2019) proposed a system using a recurrent neural network to decode cortical signals into an articulatory representation, which was then translated into intelligible speech through electrocorticography (ECoG) signals. Kohler et al. (2021) explored a less invasive approach utilizing stereotactic EEG (sEEG) in conjunction with a recurrent encoder-decoder model to synthesize audible speech. Furthermore, Kim et al. (2023) have implemented transfer learning with a pre-trained self-supervised model to mitigate the limited availability of ECoG data. These advancements underscore the potential for developing more effective communication prosthetics for individuals afflicted by neurological disorders, aiming to bridge the gap between artificial and natural speech production.

However, the use of invasive recordings faces significant challenges; these include the high risk associated with brain surgery and the difficulties encountered in maintaining such devices over extended periods. Consequently, recent research by Défossez et al. (2023) has shifted focus to decoding speech from non-invasive brain activity recordings, such as magnetoencephalography (MEG) and electroencephalography (EEG). These modalities leverage self-supervised representations and con-

[1]Audio samples are available at https://NeuralSpeak.github.io

trastive learning to isolate the most probable word or speech segment from a predefined lexicon. Despite their non-invasive nature, both MEG and EEG are known to produce signals that are notoriously noisy (Gross et al., 2013; Muthukumaraswamy, 2013; Bai et al., 2023).

In contrast, functional Magnetic Resonance Imaging (fMRI) offers another non-invasive method for decoding brain activity into complex outputs, such as images (Ozcelik and VanRullen, 2023; Takagi and Nishimoto, 2023; Ozcelik et al., 2022), video (Chen et al., 2023), and languages (Tang et al., 2023). The superior spatial resolution of fMRI enables precise localization of brain activity to specific regions, which offers an advantage over MEG and EEG. However, fMRI also has its limitations, including: (1) The temporal resolution of fMRI is substantially inferior to that of the sampling rates employed for speech signals. For example, a speech signal sampled at a frequency of 16 kHz yields 16,000 discrete samples per second, while a single fMRI frame encompasses a 2-second interval. This discrepancy imposes a significant constraint on the capacity of fMRI to resolve the rapid temporal fluctuations that are characteristic of speech dynamics. (2) fMRI measures the Blood Oxygen Level-Dependent (BOLD) signal, which, while reliable, offers an indirect proxy for neural activity. This is referred to as the Hemodynamic Response (HR) (Buckner, 1998) and introduces a temporal delay between the occurrence of neuronal events and their manifestation in BOLD signals. Consequently, when a speech stimulus is presented, the associated BOLD signal will exhibit a delayed response in relation to the actual auditory event. (3) The nature and format of language representations in brain recordings remain largely unknown. Consequently, determining the most suitable representations for speech synthesis is an unresolved problem.

In this work, we propose *NeuralSpeak*, the first non-invasive brain-to-speech synthesis framework for synthesizing natural speech from fMRI recordings. NeuralSpeak first encoder fMRI signals with a multi-window fMRI encoder, which is trained through contrastive learning. The framework then transforms speech signals into discrete representations, which are refined through training with language models that have been enhanced specifically for fMRI-guided next-token prediction. Subsequently, our framework reconstructs high-fidelity waveforms using a unit-based vocoder. Additionally, we leverage a multi-scale Transformer model to address the challenges of handling excessively long sequences resulting from the residual vector quantization codec used in tokenization. Our contributions are summarized as follows:

- We formulate the task of non-invasive brain-to-speech synthesis and devise *NeuralSpeak*, a tailored framework that adopts multi-scale language models for managing the extended discrete representations of the speech signal.
- To alleviate hemodynamic response phenomenon and capture semantically rich representations, we introduce fMRI-language contrastive pretraining with a multi-window fMRI encoder.
- Experimental results demonstrate that NeuralSpeak achieves state-of-the-art performances. The attention analysis revealed mapping to the auditory cortex and higher cognitive networks suggesting our model is biologically plausible and interpretable.

2 Related Works

2.1 Speech Synthesis from Brain Activity

Speech synthesis from brain activity, also known as brain-to-speech, is an emerging field that aims to reconstruct or generate intelligible speech directly from neural signals. Early attempts at decoding brain activity into speech involved simple models that could predict a limited set of predefined words or phrases (Herff et al., 2015; Mugler et al., 2014; Brumberg et al., 2011). More recent works (Anumanchipalli et al., 2019; Angrick et al., 2019; Kohler et al., 2021) have focused on direct synthesis of speech from invasive brain recordings with advanced machine learning networks, typically using ECoG and sEEG. For instance, Anumanchipalli et al. (2019) have employed a two-stage decoding approach based on long short-term memory, wherein articulatory kinematic features are estimated from the ECoG signals. More recently, Défossez et al. (2023) have introduced a model trained with contrastive learning to decode self-supervised representations of perceived speech from non-invasive recordings. Concurrently, there is a growing body of research focusing on the reconstruction of music from brain activity (Denk et al., 2023; Ramirez-Aristizabal and Kello, 2022), utilizing modalities such as fMRI and

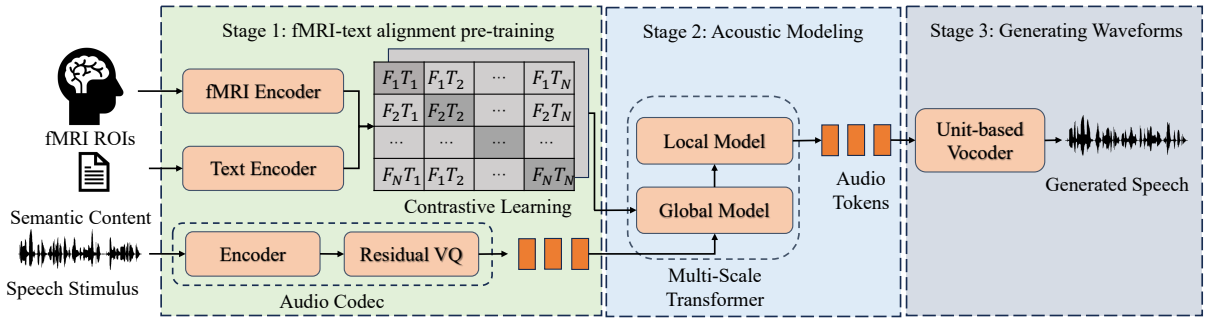


Figure 1: A high-level overview of NeuralSpeak. The framework consists of three core stages—(1) aligning fMRI representations with textual features, (2) autoregressively modeling audio tokens using multi-scale transformers, and (3) self-supervised waveform reconstruction. The framework employs the FLAN-T5 text encoder for linguistic feature extraction.

EEG. However, a notable research gap exists concerning non-invasive brain-to-speech synthesis. In this study, we present the pioneering framework for non-invasive brain-to-speech synthesis that leverages fMRI signals.

2.2 Speech Representation

Recent research has increasingly focused on efficiently encoding audio signals into compact discrete representations, aiming to optimize speech processing and high-fidelity audio coding. Pioneering techniques such as Wav2Vec (Baevski et al., 2020) and Hubert (Hsu et al., 2021) have employed k-means quantization to compress speech effectively. Additionally, SoundStream (Zeghidour et al., 2021) and Encodec (Défossez et al., 2022) have explored hierarchical vector quantization (VQ) methods to enhance the representation of acoustic information, showing promise in audio signal reconstruction with higher quality. A novel group-residual vector quantization (GRVQ) approach presented by (Yang et al., 2023) further advances audio coding. Our work builds upon SoundStream's progress to extract discrete representations for improved speech synthesis and processing, strengthening our proposed framework.

2.3 Language Models

Modeling audio within a compact discrete space has garnered significant attention, facilitating efficient audio representation through autoregressive transformers. Innovations like AudioLM (Borsos et al., 2022) and MusicLM (Agostinelli et al., 2023) treat audio synthesis as language modeling with a hierarchical coarse-to-fine structure, yielding high-quality audio synthesis with granular control. SpeechDLM (Nguyen et al., 2023), focusing on speech for dialogue, and MusicGen (Copet et al., 2023), which handles multiple streams of music representations, extend these concepts, offering realistic speech and complex musical compositions. In this study, we introduce a versatile and scalable framework for non-invasive brain-to-speech synthesis. This framework employs an autoregressive sequence-to-sequence (seq2seq) approach and leverages discrete representations.

3 Methods

3.1 Overview

NeuralSpeak is recognized as a scalable and adaptable framework that progressively improves the modeling of speech signals by integrating relevant fMRI information. This process is organized into three primary stages, as illustrated in Figure 1: 1) fMRI-text Alignment Pre-training: fMRI recordings are transformed into semantically rich representations using contrastive learning objectives by a multi-window fMRI encoder. 2) Acoustic Modeling Fine-tuning: Audio tokens are generated sequentially from the aligned fMRI features by multi-scale language models. 3) High-Fidelity Waveform Synthesis: A unit-based vocoder synthesizes high-fidelity waveforms from compressed acoustic representations. In the following sections, we describe these steps in detail.

3.2 Discrete Speech Representation

Recently, audio codec models such as SoundStream (Zeghidour et al., 2021) and Encodec (Défossez et al., 2022) have demonstrated the effectiveness of encoder-decoder architectures in learning acoustic information in a self-supervised manner. These architectures are capable of extracting rich representations from audio data, which can be leveraged for a variety of generative tasks.

The acoustic codec model typically comprises an audio encoder, a residual vector quantizer (RVQ),

and an audio decoder: 1) The audio encoder E is composed of multiple convolutional blocks with a total downsampling rate of 320, producing continuous representations at every 20-ms frame at 16 kHz. 2) The residual vector quantizer Q generates discrete representations $a_t \in \mathbb{R}^{T \times N_q}$, where T is the number of audio frames after downsampling and N_q is the number of vector quantization layers, utilizing a vector quantization technique (Vasuki and Vanathi, 2006). 3)The audio decoder G reconstructs the signal $\hat{y}$ from the compressed latent representation a_t.

3.3 Brain Representation

Multi-window fMRI Encoder While fMRI offers excellent spatial specificity, the BOLD signal it records is characterized by slow dynamics. An impulse of neural activity triggers the BOLD signal to rise and fall over a period of approximately 10 seconds (Logothetis, 2003). This implies that the fMRI data captured at a specific time may not fully capture the information about a corresponding auditory stimulus presented at the same time. Therefore, to adequately extract information for decoding each scanning window and to accommodate the hemodynamic response (HR), we propose a multi-window Transformer architecture. This architecture incorporates spatial-temporal attention mechanisms to effectively process sequential fMRI frames.

Consider a series of fMRI frames denoted as $x_t \in \mathbb{R}^{B \cdot W \times 1 \times V}$, where W, B, and V represent the window size, batch size, and the number of voxels, respectively. Inspired by the Vision Transformer (Dosovitskiy et al., 2020), the fMRI data undergoes an initial transformation through a patch embedding process to yield $x_p \in \mathbb{R}^{(B \cdot W) \times P \times D}$, where P denotes the patch size and D signifies the patch embedding dimension. Subsequently, spatial attention is computed as follows, with the query Q, key K, and value V all derived from the projected x_p.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \quad (1)$$

Here, d_k represents the hidden dimension of the key. The output $x_p^{spatial}$ is obtained by applying spatial attention. Subsequently, to compute temporal attention, we transpose the dimensions of p and w to obtain $x_p^{temp} \in \mathbb{R}^{B \cdot P \times W \times D}$. We then apply the same attention mechanism as in Equation 1, with the query, key, and value set to x_p^{temp}.

Contrastive fMRI-text Pretraining The encoding of acoustic, phonetic, lexical, and semantic information in brain recordings remains poorly understood, posing a significant challenge in identifying optimal representations for speech synthesis. To address this challenge, we enhance the multi-window fMRI encoder by incorporating fMRI-text pairs. Our objective is to align the fMRI-derived embeddings more closely with semantically rich text representations.

We process the fMRI data and corresponding text through distinct encoders: an fMRI encoder and a text encoder. This yields the fMRI representation $x_f \in \mathbb{R}^{B \times V}$ with dimensionality V, and the text representation $x_t \in \mathbb{R}^{B \times U}$ with dimensionality U, where B represents the batch size. Both representations are then projected into a joint multimodal space with dimension D, resulting in embeddings $E_f \in \mathbb{R}^{B \times D}$ and $E_t \in \mathbb{R}^{B \times D}$, achieved via a multilayer perceptron (MLP). Now that the fMRI and text embeddings are comparable, the contrastive loss $\mathcal{L}_{\text{CL}}$ is calculated as follows:

$$\begin{aligned}\mathcal{L}_{\text{CLIP}}(a, b) &= \text{Cross\_Entropy}(\epsilon \cdot (a \cdot b^\top), \text{range}(n)) \\ \mathcal{L}_{\text{CL}}(e_f, e_t) &= 0.5 \times (\mathcal{L}_{\text{CLIP}}(e_f, e_t) + \mathcal{L}_{\text{CLIP}}(e_t, e_f))\end{aligned} \quad (2)$$

Where ϵ is a scaling parameter. To create a contrastive learning scenario, following the common practice (Radford et al., 2021; Elizalde et al., 2023; Huang et al., 2022; Wu et al., 2023), we treat other elements of the batch as negative samples.

3.4 Multi-Scale Acoustic Modeling

Despite the effectiveness of our audio codec model in compressing raw waveforms into a condensed format with dimensions $T \times N_q$, the conventional Transformer architecture faces a significant limitation due to its intrinsic quadratic computational complexity, denoted by $\mathcal{O}(T^2 N_q^2)$. This complexity makes the model inefficient when processing even the compressed sequences, as they remain considerably lengthy. In response to this challenge, we draw inspiration from the work of Yu et al. (2023) and propose a multi-scale Transformer architecture tailored for discrete audio sequences. This hierarchical framework addresses correlations both within and between frames by incorporating distinct global and local Transformer modules. Specifically, the architecture segments every N_q consecutive token into global modeling units, subsequently managing the tokens within each segment at a local scale, as depicted in Figure 2.

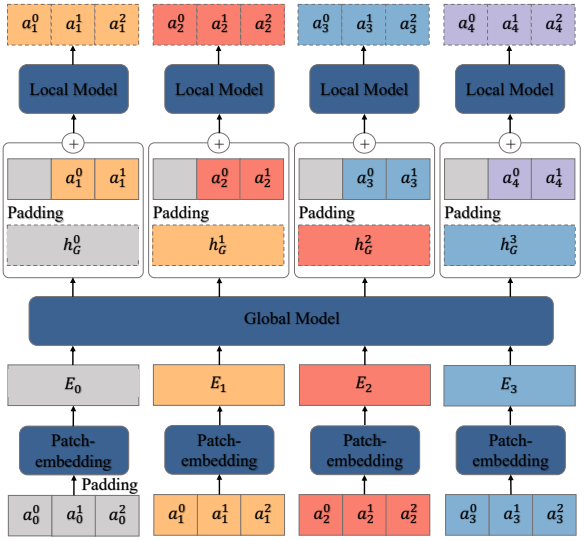


Figure 2: The architecture of multi-scale Transformer with patch size $P = N_q = 3$. Inputs to both the global and local models are padded by a single patch. The global model's output serves as the conditioning context for the local model, which then autoregressively predicts each patch in parallel. Note that the gray blocks denote the padding tokens.

Initially, to facilitate the patching of fMRI embeddings, we duplicate each embedding N_q times to populate a patch. This patch is then concatenated with the corresponding audio segment embedding a_p, incorporating special tokens such as '<fMRI_start>', '<fMRI_end>', '<audio_start>' and '<audio_end>' to identify boundaries. This process yields the patch embedding $E_t \in \mathbb{R}^{B\times K\times D_G\cdot N_q}$, where K represents the patch length, and D_G is the dimension of the global embedding. To enable autoregressive modeling, we subsequently augment the patched sequence with a trainable padding embedding at the beginning, while excluding the final patch from the input. This modified sequence is then processed by the global model to obtain the global hidden states $h_G \in \mathbb{R}^{B\times K\times D_G\cdot N_q}$. In the third step, we map the output of the global model to the dimension of the local model, D_L, and reshape the output sequence into $E_t^{Local} \in \mathbb{R}^{B\cdot K\times N_q\times D_L}$. For this local embedding, we introduce an offset by incorporating a trainable local padding embedding. Finally, we feed the local embedding into the local model and compute the probability distribution over the vocabulary, as described by the following equation:

$$p\left(\mathbf{a} \mid \mathbf{x_f}, \mathbf{a_p}; \theta_{ARs}\right) = \prod_{t=0}^{T} p\left(\mathbf{a_t} \mid \mathbf{a}< t, \mathbf{x_f}, \mathbf{a_p}; \theta_{ARs}\right), \tag{3}$$

Where θ_{ARs} represents the parameters of the autoregressive models (i.e., the global model and local model), and $\mathbf{a}_t$ denotes the audio token at time t.

3.5 High-Fidelity Waveform Synthesis

Upon completion of the training process, language models can be utilized to generate acoustic tokens based on the provided fMRI signals. Subsequently, a unit-based vocoder is employed to synthesize the corresponding speech waveforms. It is worth noting that the acoustic codec used, such as SoundStream, leverages multiple quantization levels, typically 12, to enhance the quality of speech reconstruction. Thus, reducing the number of codebooks during the inference stage might result in a noticeable drop in perceptual quality.

To maintain the quality of the generated audio waveforms, we employ a unit-based neural vocoder specially designed for waveform generation from acoustic units. This vocoder is trained from scratch and achieves high-quality audio reconstruction using only three quantization levels. Inspired by the architecture of the BigVGAN model (Lee et al., 2022), our synthesizer consists of a generator and a multi-resolution discriminator (MRD). The generator incorporates a set of look-up tables (LUT) for embedding the discrete representations, alongside a series of blocks. Each block consists of transposed convolutions and a residual block with dilated layers. The transposed convolutions are responsible for upsampling the encoded representation to match the input sample rate, while the dilated layers enhance the receptive field.

3.6 Training and Inference Procedures

Our model undergoes training through a bifurcated strategy comprising three distinct stages. Initially, the fMRI encoder is trained on our dataset utilizing contrastive learning, while concurrently, the language model undergoes training on audio samples with text conditioning. In the subsequent stage, both the fMRI encoder and the language models are fine-tuned jointly using paired fMRI-audio data. In the third stage, which focuses on synthesis, we train the advanced vocoder using a composite loss function that integrates the least-squares adversarial loss, feature matching loss, and spectral regression loss. During inference, we consistently employ top-k sampling to generate predictions, and then the audio output is synthesized from the tokens predicted by the language model.

Model	MOS (↑)	SMOS (↑)	WER (↓)	SIM (↑)
GT	4.25 ± 0.07	/	0.02	/
Model Performances				
Random	1.32 ± 0.15	1.14 ± 0.17	0.99	0.04
Regression	2.61 ± 0.10	2.43 ± 0.08	0.93	0.32
Cascaded	2.66 ± 0.09	3.12 ± 0.09	0.86	0.49
NeuralSpeak	**3.56 ±0.07**	**3.41 ±0.08**	**0.08**	**0.62**
Analysis Across Different Subjects				
Subject 2	3.49 ±0.09	3.36 ±0.10	0.14	0.58
Subject 3	3.52 ±0.08	3.40 ±0.09	0.10	0.59

Table 1: We summarize the results of comparison and analysis across different subjects in one table using objective and subjective metrics to evaluate the quality and style similarity of generated samples. By default, we use the data of Subject 1 for evaluation.

4 Experiments

4.1 Experimental Setup

4.1.1 The fMRI Dataset

We conduct preprocessing of the dataset from LeBel et al. (2023) following the methodology established by Jain et al. (2020). Detailed descriptions of the data collection and preprocessing steps are provided in the Appendix A. The stimulus set comprises 84 narratives, each lasting between 10 to 15 minutes, with a cumulative duration of approximately 15.8 hours.

4.1.2 Model Configurations

The sample rate of speech samples is 16,000 Hz. The dataset contains 3 subjects and each subject varies in the size of ROIs, where Subject 1, 2, and 3 have 1929, 4792, and 2747 voxels, respectively. For audio tokens, we train the SoundStream model with 12 quantization levels, each with a codebook of size 1024 and the same downsampling rate of 320. We take 3 quantization levels as the acoustic tokens. Language models are both 24-layer global transformers with an attention dimension of 1536 and 6-layer local transformers with the same dimension. As for the unit-based vocoder, we use the modified V1 version of BigVGAN. A comprehensive table of hyperparameters is available in Appendix B.1.

4.1.3 Training and Evaluation

During training, we train language models for 50K steps using 8/80 NVIDIA A100 GPUs with a batch size of 10000 tokens for each GPU on the publicly-available *fairseq* framework (Ott et al., 2019). Adam optimizer is used with $\beta_1 = 0.9, \beta_2 = 0.98, \epsilon = 10^{-9}$. The contrastive learning of fMRI encoder is optimized with an initial learning rate 10^{-3} using 8 NVIDIA A100 GPUs. Reconstructing audio model is optimized with a segment size of 8192 and a learning rate of 1×10^{-4} until 500K steps using 4 NVIDIA A100 GPUs. During inference, we use batch size 1 of autoregressive decoding in language modeling.

4.1.4 Evaluation Metrics

To evaluate the performance of NeuralSpeak on synthesized speech samples, we include both objective metrics and subjective metrics. For objective evaluation, Word Error Rate (WER) is used to evaluate the intelligibility of generated speech, Similarity Score (SIM) is for similarity in terms of speaker identity. For subjective evaluation, MOS is adopted to provide human-centric judgment for the quality of speech samples. Note all subjective results are obtained from Amazon Mechanical Turk for fair comparison. Appendix C shows details of the subjective evaluation process.

4.2 Model Performances

To comprehensively evaluate the superiority of NeuralSpeak and the effectiveness of our proposed methods, we compared it with other baselines using subjective and objective metrics. We compare the generated audio samples with other systems, including 1) GT, the ground-truth speech; 2) Random: a random baseline model that predicts the audio units using a randomly initialized version of NeuralSpeak; 3) Regression: we train a linear regression to predict the softmax probability of the true audio units generated by the audio codec model; 4) Cascaded: this baseline is composed of state-of-the-art fMRI-to-text model (Tang et al., 2023) in this dataset and Text-to-Speech MVoice

Model	MOS (↑)	SMOS (↑)	WER (↓)	SIM (↑)
NeuralSpeak	3.56 ±0.07	3.41 ±0.08	0.08	0.62
w/o Contrastive	3.49 ±0.08	3.32 ±0.08	0.15	0.60
w/o Multi-window fMRI Encoder	3.54 ±0.08	3.38 ±0.09	0.13	0.59
w/o Multi-scale Transformer	3.36 ±0.07	3.25 ±0.08	0.18	0.52
w/o Language Cortex	3.42 ±0.08	3.34 ±0.08	0.17	0.54
w/o Auditory Cortex	3.39 ±0.09	3.32 ±0.08	0.16	0.50

Table 2: The ablation studies to explore the effectiveness of our proposed contrastive learning, multi-window fMRI encoder, multi-scale transformer, and ROI regions. To replace the multi-window fMRI encoder and multi-scale transformer, we adopt vision transformers and the language models used by AudioGen (Kreuk et al.), which adopts parallel prediction.

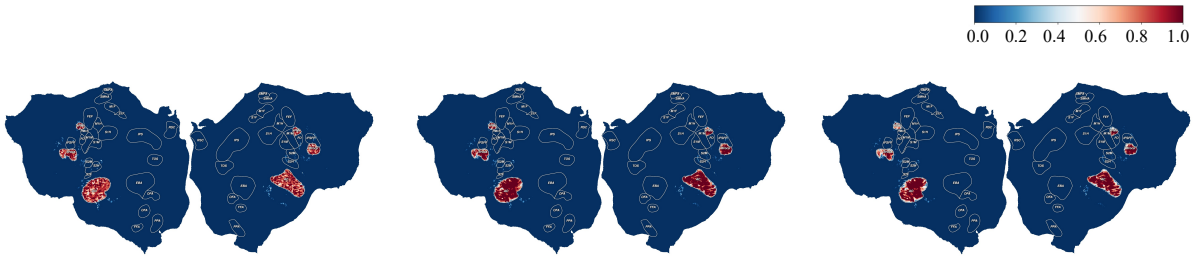
model (Huang et al., 2023). For easy comparison, the results are compiled and presented in Table 1, and we have the following observations: 1) For the intelligibility of the generated speech, NeuralSpeak has achieved a WER of 0.08, which is much lower than other systems. This indicates that NeuralSpeak could generate accessible speech of better quality compared to other baselines. 2) For audio quality, NeuralSpeak has achieved the highest MOS with scores of 3.56 ± 0.07 compared to the baseline models, demonstrating the effectiveness of our model in generating high-fidelity waveforms. 3) Regarding style similarity, NeuralSpeak scores the SMOS of 3.41 ± 0.08. The objective results of SIM further show that NeuralSpeak surpasses other baselines in generating identified voices.

Analysis Across Different Subjects To further analyze the performance across different subjects, we evaluate NeuralSpeak for Subject 2 and Subject 3. The number of voxels of ROIs varies in different subjects, where Subject 1, 2, and 3 have 1929, 4792, and 2747 voxels, respectively. The results are also included in Table 1, and the following observations are made: 1) The WER for Subject 2 and Subject 3 remains low, indicating the capability of NeuralSpeak to generate intelligible speech for different individuals. 2) For Subject 2 and Subject 3, NeuralSpeak consistently outperforms the baselines in terms of MOS, SMOS, and SIM. This indicates the robustness of NeuralSpeak in generating high-quality and style-consistent speech across different subjects. 3) The results of Subject 1 performs slightly better than Subject 2 and Subject 3, we contribute it to the smaller ROI size may produce better results with a larger batch size.

4.3 Ablation Studies

We conduct ablation studies to demonstrate the effectiveness of several key techniques on the test set in our model, including the contrastive learning, multi-window fMRI encoder, and multi-scale Transformer. we conduct ablation studies and discuss the key findings as follows. 1) Removing contrastive learning results in a significant degradation of generation quality. This indicates that NeuralSpeak has the ability to learn representations of language that are particularly valuable for brain-to-speech synthesis by aligning fMRI representations with text features. 2) Without the multi-window fMRI encoder designs, there is a distinct degradation in all metrics, which demonstrates that our model successfully alleviates the problem of hemodynamic response using a sliding window and spatial-temporal attention. 3) Multi-scale Transformer outperforms the parallel prediction approach used in AudioGen (Kreuk et al., 2023) in terms of generation quality. This is because the latter fails to preserve the property of autoregression when introducing concurrent prediction, while our multi-scale Transformer maintains this property. Moreover, our approach considerably reduces complexity from $T^2N_q^2$ to $\frac{T^2}{N_q^2} + TN_q$ by incorporating global and local modeling. 4) We further investigate the performance of using responses solely from language cortex and auditory cortex, and found that the use of responses from both the semantic regions and auditory cortex yielded better results. These findings highlight the importance of both semantic regions and the auditory cortex in representing valuable information for synthesizing high-fidelity waveforms.

Retrieval Results We further conduct the fMRI-to-text retrieval experiments to demonstrate the effectiveness of our proposed contrastive fMRI-text pre-training. The performance is evaluated based on the metrics of R@1 and R@10 for both text-to-fMRI (T-F) retrieval and fMRI-to-text (F-T) retrieval. The results are presented in Table 3, and



A. First Layer B. Middle Layer C. Last Layer

Figure 3: Visualization of Transformer Attention Maps. Attention maps of different transformer layers are shown in (A), (B), and (C). AC: Auditory Cortex; sPMv: superior premotor cortex.

we have the following observations:

Model	T-F retrival		F-T retrival	
	R@1	R@10	R@1	R@10
NeuralSpeak	17.2	55.6	24.1	58.2
w/o Contrastive	5.4	21.2	6.5	24.6
w CLAP	13.4	48.9	17.3	47.8
w RoBERTa	15.1	52.7	21.5	53.6

Table 3: The fMRI-to-text retrieval performance. We compare the performance of our text encoder T5-large with CLAP text encoder (Elizalde et al., 2023) and RoBERTa (Liu et al., 2019).

1) When the contrastive learning technique was removed, lower results were obtained for both T-F retrieval and F-T retrieval. This demonstrates the effectiveness of contrastive learning in improving retrieval performance. 2) Replacing the T5 encoder with alternative text encoders resulted in a degradation of retrieval performance. This highlights the importance of using advanced text encoders to enhance the retrieval results.

4.4 Interpretation Results

We calculate the average attention across the entire test set and visualize the voxel-wise self-attention value on a brain flat map. The resulting figure (Figure 3) shows a comprehensive distribution of attention throughout the entire brain region, from which we derive key insights:

1) The attention maps highlight the significant role played by the auditory cortex and language cortex (specifically the Broca and sPMv regions) in the natural speech synthesis process. These regions exhibit high attention values, indicating their crucial involvement in the processing and generation of speech. This finding aligns with our existing knowledge of the brain, where the auditory cortex is responsible for sound perception (King and Schnupp, 2007) and language-related regions are involved in language production and comprehension (Friederici, 2012).

2) The attention maps across different transformer layers demonstrate a hierarchical pattern of functionality within the fMRI encoder. In the initial layers (Fig. B), the self-attention layers are primarily focused on the structural characteristics of the input data, delineating brain regions based on their attention values in auditory processing. This observation echoes the brain's methodical approach to processing auditory information, where lower-level regions analyze basic acoustic features. Progressing to deeper layers (Fig. C and D), the attention becomes more dispersed, resulting in decreased differentiation between specific regions. This suggests a transition towards the acquisition of more holistic and abstract acoustic features in the deeper layers.

5 Conclusion

In this work, we proposed NeuralSpeak, a system specifically devised for the task of non-invasive brain-to-speech synthesis, offering a promising pathway to restore communicative functions in patients with neurological impairments. To tackle the obstacles associated with capturing optimal neural representations and addressing the hemodynamic response within brain recordings, we designed a multi-window fMRI encoder. This encoder, trained through contrastive learning, generates brain-derived embeddings that exhibit close semantic alignment with text representations. Additionally, to resolve the issue of excessively lengthy audio tokens, we have implemented a multi-scale transformer architecture. The efficacy of NeuralSpeak is affirmed by its state-of-the-art performance in both objective measurements and subjective assessments. Moreover, our model demonstrates biological plausibility and interpretability, reflecting well-established physiological processes. We envisage that our work will serve as a basis for future non-invasive brain-to-speech synthesis studies.

6 Limitation and Potential Risks

MindSpeak adopts auto-regressive models for high-quality synthesis, and thus it inherently requires iterative refinements for better results. Besides, a longer sequence length typically requires more computational resources, and degradation could be witnessed with decreased training data. One of our future directions is to develop lightweight and parallel models for accelerating sampling.

MindSpeak has the potential to revolutionize communication for individuals with speech impairments. However, as with any advanced technology, there are potential negative societal impacts that warrant consideration. This technology may potentially extract and vocalize thoughts without consent, leading to serious privacy violations and the possibility of unauthorized surveillance and misuse of personal information. In addition, there is the potential for leading to unequal access, with only those who can afford it benefiting from the technology, exacerbating social and economic inequalities.

References

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Miguel Angrick, Christian Herff, Emily Mugler, Matthew C Tate, Marc W Slutzky, Dean J Krusienski, and Tanja Schultz. 2019. Speech synthesis from ecog using densely connected 3d convolutional neural networks. *Journal of neural engineering*, 16(3):036019.

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

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. 2023. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934*.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2022. Audiolm: a language modeling approach to audio generation. *arXiv preprint arXiv:2209.03143*.

Jonathan S Brumberg, E Joe Wright, Dinal S Andreasen, Frank H Guenther, and Philip R Kennedy. 2011. Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech motor cortex. *Frontiers in neuroscience*, 5:7880.

Randy L Buckner. 1998. Event-related fmri and the hemodynamic response. *Human brain mapping*, 6(5-6):373–377.

Zijiao Chen, Jiaxin Qing, and Juan Helen Zhou. 2023. Cinematic mindscapes: High-quality video reconstruction from brain activity. *arXiv preprint arXiv:2305.11675*.

Jan Claassen, Kevin Doyle, Adu Matory, Caroline Couch, Kelly M Burger, Angela Velazquez, Joshua U Okonkwo, Jean-Rémi King, Soojin Park, Sachin Agarwal, et al. 2019. Detection of brain activation in unresponsive patients with acute brain injury. *New England Journal of Medicine*, 380(26):2497–2505.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. 2023. Simple and controllable music generation.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. 2023. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.

Timo I Denk, Yu Takagi, Takuya Matsuyama, Andrea Agostinelli, Tomoya Nakai, Christian Frank, and Shinji Nishimoto. 2023. Brain2music: Reconstructing music from human brain activity. *arXiv preprint arXiv:2307.11078*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. 2023. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Angela D Friederici. 2012. The cortical language circuit: from auditory perception to sentence comprehension. *Trends in cognitive sciences*, 16(5):262–268.

Joachim Gross, Sylvain Baillet, Gareth R Barnes, Richard N Henson, Arjan Hillebrand, Ole Jensen, Karim Jerbi, Vladimir Litvak, Burkhard Maess, Robert Oostenveld, et al. 2013. Good practice for conducting and reporting meg research. *Neuroimage*, 65:349–363.

Christian Herff, Dominic Heger, Adriana De Pesters, Dominic Telaar, Peter Brunner, Gerwin Schalk, and Tanja Schultz. 2015. Brain-to-text: decoding spoken phrases from phone representations in the brain. *Frontiers in neuroscience*, 9:217.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 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.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel PW Ellis. 2022. Mulan: A joint embedding of music audio and natural language. *arXiv preprint arXiv:2208.12415*.

Rongjie Huang, Chunlei Zhang, Yongqi Wang, Dongchao Yang, Luping Liu, Zhenhui Ye, Ziyue Jiang, Chao Weng, Zhou Zhao, and Dong Yu. 2023. Make-a-voice: Unified voice synthesis with discrete representation.

Shailee Jain, Vy Vo, Shivangi Mahto, Amanda LeBel, Javier S Turek, and Alexander Huth. 2020. Interpretable multi-timescale models for predicting fmri responses to continuous natural speech. *Advances in Neural Information Processing Systems*, 33:13738–13749.

Miseul Kim, Zhenyu Piao, Jihyun Lee, and Hong-Goo Kang. 2023. Braintalker: Low-resource brain-to-speech synthesis with transfer learning using wav2vec 2.0. In *2023 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–5. IEEE.

Andrew J King and Jan WH Schnupp. 2007. The auditory cortex. *Current Biology*, 17(7):R236–R239.

Jean-Rémi King, Frédéric Faugeras, Alexandre Gramfort, Aaron Schurger, Imen El Karoui, Jacobo D Sitt, Benjamin Rohaut, Catherine Wacongne, Etienne Labyt, Tristan Bekinschtein, et al. 2013. Single-trial decoding of auditory novelty responses facilitates the detection of residual consciousness. *Neuroimage*, 83:726–738.

Jonas Kohler, Maarten C Ottenhoff, Sophocles Goulis, Miguel Angrick, Albert J Colon, Louis Wagner, Simon Tousseyn, Pieter L Kubben, and Christian Herff. 2021. Synthesizing speech from intracranial depth electrodes using an encoder-decoder framework. *arXiv preprint arXiv:2111.01457*.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually Guided Audio Generation.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2023. Audiogen: Textually guided audio generation.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G Huth. 2023. A natural language fmri dataset for voxelwise encoding models. *Scientific Data*, 10(1):555.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. 2022. Bigvgan: A universal neural vocoder with large-scale training. *arXiv preprint arXiv:2206.04658*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nikos K Logothetis. 2003. The underpinnings of the bold functional magnetic resonance imaging signal. *Journal of Neuroscience*, 23(10):3963–3971.

Sean L Metzger, Jessie R Liu, David A Moses, Maximilian E Dougherty, Margaret P Seaton, Kaylo T Littlejohn, Josh Chartier, Gopala K Anumanchipalli, Adelyn Tu-Chan, Karunesh Ganguly, et al. 2022. Generalizable spelling using a speech neuroprosthesis in an individual with severe limb and vocal paralysis. *Nature Communications*, 13(1):6510.

David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. 2021. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227.

Emily M Mugler, James L Patton, Robert D Flint, Zachary A Wright, Stephan U Schuele, Joshua Rosenow, Jerry J Shih, Dean J Krusienski, and Marc W Slutzky. 2014. Direct classification of all american english phonemes using signals from functional speech motor cortex. *Journal of neural engineering*, 11(3):035015.

Suresh D Muthukumaraswamy. 2013. High-frequency brain activity and muscle artifacts in meg/eeg: a review and recommendations. *Frontiers in human neuroscience*, 7:138.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. *arXiv preprint arXiv:1904.01038*.

Adrian M Owen, Martin R Coleman, Melanie Boly, Matthew H Davis, Steven Laureys, and John D Pickard. 2006. Detecting awareness in the vegetative state. *science*, 313(5792):1402–1402.

Furkan Ozcelik, Bhavin Choksi, Milad Mozafari, Leila Reddy, and Rufin VanRullen. 2022. Reconstruction of perceived images from fmri patterns and semantic brain exploration using instance-conditioned gans. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.

Furkan Ozcelik and Rufin VanRullen. 2023. Brain-diffuser: Natural scene reconstruction from fmri signals using generative latent diffusion. *arXiv preprint arXiv:2303.05334*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision.

Adolfo G Ramirez-Aristizabal and Chris Kello. 2022. Eeg2mel: Reconstructing sound from brain responses to music. *arXiv preprint arXiv:2207.13845*.

Yu Takagi and Shinji Nishimoto. 2023. 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.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, pages 1–9.

A Vasuki and PT Vanathi. 2006. A review of vector quantization techniques. *IEEE Potentials*, 25(4):39–47.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. 2023. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*.

Lili Yu, Dániel Simig, Colin Flaherty, Armen Aghajanyan, Luke Zettlemoyer, and Mike Lewis. 2023. Megabyte: Predicting million-byte sequences with multiscale transformers.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.

A Dataset Collection and Preprocessing

A.1 fMRI data collection

The MRI data was acquired over six scanning sessions (15 sessions for the extended dataset) using a 3T Siemens Skyra scanner at the UT Austin Biomedical Imaging Center, employing a 64-channel Siemens volume coil. The initial session comprised an anatomical scan and functional localizers. Subsequent sessions involved passive listening to 4–5 stories, including the story designated for model testing. Each story was presented during a single EPI scan, incorporating a 10-second silent padding period at both the beginning and end of the narrative.The audio was transmitted via Sensimetrics S14 in-ear piezoelectric headphones. To reduce head movement, foam headcases (CaseForge, Inc., now defunct) were employed to snugly fit the gap between the participant's head and the head coil during data acquisition. Creating the headcases involved utilizing an RGB Structure.io sensor (Occipital Inc.) to capture a three-dimensional scan of each participant's head, while their hair was compressed using a swim cap. Subsequently, these scans were utilized to fabricate custom styrofoam headcases for individual participants.

A.2 Speech stimulus Collection

For three of these participants, the training set comprises a total of 82 stories, including two additional stories designated for use as a test dataset. During stimulus presentation, the audio for each story underwent filtering to rectify frequency response and phase errors caused by the headphones. This filtering process utilized calibration data supplied by Sensimetrics, augmented by custom Python code [2]. All stimuli were played at a sampling rate of 44.1 kHz using the pygame library in Python.

A.3 Data preprocessing

fMRI preprocessing was exclusively performed on the derivative data. This data underwent motion correction using the FMRIB Linear Image Registration Tool (FLIRT) from the FMRIB Software Library (FSL) version 5.028. Following motion correction, all volumes within each run were averaged to derive a single template volume. Cross-run alignment was subsequently conducted using FLIRT to align the template volume from each run with the template volume from the initial run in the first story session. These automated alignments underwent manual verification. The concatenated motion correction and cross-run transformations were then used to resample the original data into a motion-corrected and cross-run-aligned space, thereby minimizing unwanted blurring associated with multiple resampling steps. The motion correction and cross-run transformations were subsequently concatenated and applied to resample the original data into a motion-corrected and cross-run-aligned space. This approach mitigates the need for multiple resampling steps, thereby minimizing undesired blurring effects. Additionally, low-frequency voxel response drift was identified using a 2^{nd} order Savitzky-Golay filter with a 120-second window and subtracted from the signal. To minimize artifacts arising from onset transients and suboptimal detrending performance at the data boundaries, we trimmed the responses by discarding the initial and final 20 seconds (equivalent to 10 volumes) of each scan. This adjustment effectively eliminated the 10-second silent periods and the first and last 10 seconds of each story. In terms of audio preprocessing, we downsampled the audio samples to 16 kHz before temporally aligning them with the fMRI recordings.

B More Implementation Details

B.1 Model Configurations

We list the model hyper-parameters of NeuralSpeak in Table 4.

B.2 Unit-based Vocoder

The generator of the unit-based vocoder is constructed using a set of look-up tables (LUT) to embed the discrete representations, alongside a sequence of blocks comprising transposed convolutions and a residual block with dilated layers.

C Subjective Evaluation

For evaluating audio quality, we perform MOS (mean opinion score) tests, explicitly instructing the raters to "*(focus on examining the audio quality and naturalness, and ignore differences in style (such as timbre, emotion, and prosody))*". Testers present and rate the samples, and each tester is requested to evaluate subjective naturalness using a 1-5 Likert scale.

For style similarity evaluation, raters are explicitly instructed to "*(focus on the similarity of the*

[2] https://github.com/alexhuth/sensimetrics_filter

Hyperparameter		NeuralSpeak
fMRI Encoder	Patch Size	16
	Embed Dim	1024
	Encoder Layer	24
	Encoder Heads	16
Global Model	Transformer Layer	20
	Transformer Embed Dim	1152
	Transformer Heads	16
	Transformer FFN Dim	4608
	Dictionary Length	4163
	Patch Size	3
Local Model	Transformer Layer	6
	Transformer Embed Dim	1152
	Transformer Heads	8
	Transformer FFN Dim	4608
Vocoder	Upsample Rates	[5, 4, 2, 2, 2, 2]
	Hop Size	320
	Upsample Kernel Sizes	[9, 8, 4, 4, 4, 4]
	Number of Parameters	121.6M
Total Number of Parameters		1.4B

Table 4: Hyperparameters of NeuralSpeak.

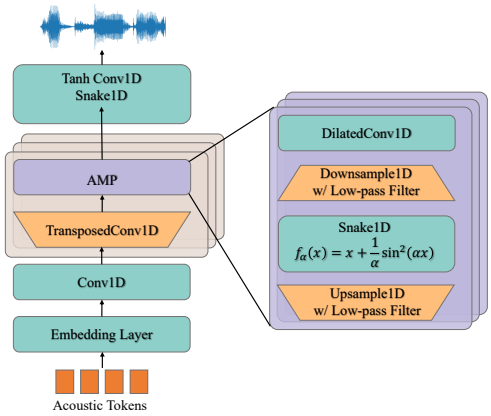


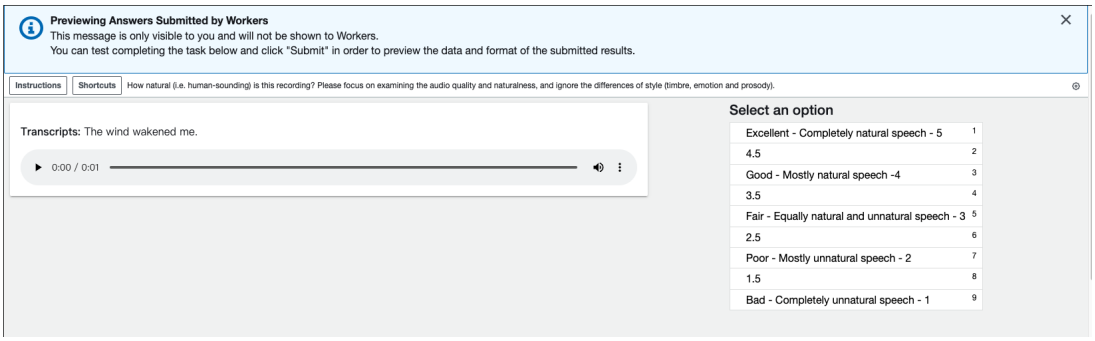
Figure 4: Overview of the unit-based vocoder.

style (timbre, emotion, and prosody) to the reference, and ignore the differences of content, grammar, or audio quality.)" during the SMOS (similarity mean opinion score) tests. In these tests, each synthesized utterance is paired with a true utterance to assess how closely the synthesized speech matches that of the target speaker. Each pair is rated by a single rater.

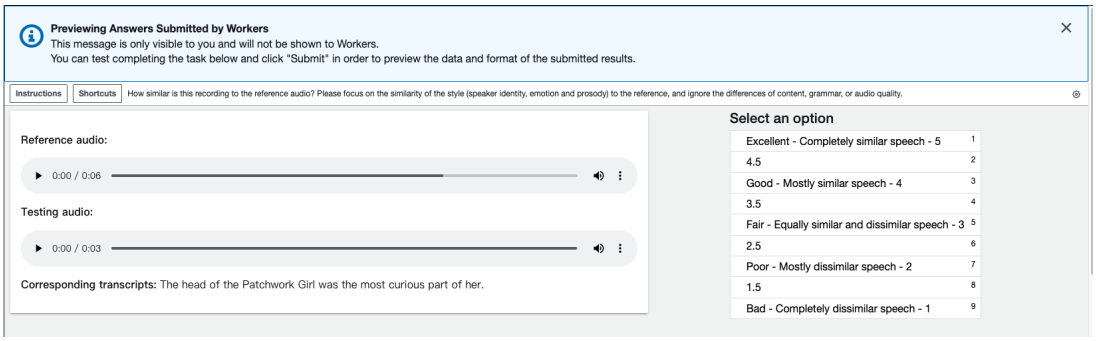
Our subjective evaluation tests were crowdsourced and carried out by 20 native speakers through Amazon Mechanical Turk. The screenshots of instructions for the testers are provided in Figure 5. Participants were compensated at a rate of $8 per hour, resulting in an expenditure of approximately $600 for participant compensation. A limited subset of speech samples utilized in the evaluation is accessible at https://NeuralSpeak.github.io/.

D Reproducibility Statement

We will release our code in the future. The NeuralSpeak model that we build upon is publicly available through the Fairseq code repository. To aid reproducibility, we have included an overview of the hyperparameters in Table 4.



(a) Screenshot of MOS testing.



(b) Screenshot of SMOS testing.

Figure 5: Screenshots of subjective evaluations.