Tsinghua Science and Technology
Doi: 10.26599/TST.2026.9010004

Research Article

Reverse the auditory processing pathway: Coarse-to-fine audio reconstruction from human brain activity

Che Liu, Changde Du, Xiaoyu Chen, and Huiguang He (✉)

Abstract Drawing inspiration from the hierarchical processing of the human auditory system, which transforms sound from low-level acoustic features to high-level semantic understanding, we introduce a novel coarse-to-fine audio reconstruction method. Leveraging non-invasive functional Magnetic Resonance Imaging (fMRI) data, our approach first utilizes CLAP to decode fMRI signals coarsely into a semantic space, followed by a semantically-guided fine-grained decoding into the AudioMAE latent space. These fine-grained neural features then serve as conditions for high-fidelity audio reconstruction through a Latent Diffusion Model (LDM). Extensive validation on three public fMRI datasets demonstrates the superiority of our coarse-to-fine decoding method over conventional fine-grained approaches, achieving state-of-the-art performance across metrics including FD, FAD, and KL. Furthermore, reconstruction quality in challenging scenarios is enhanced through an innovative semantic prompting mechanism. This framework holds potential for advancing brain-computer interfaces and assistive technologies, such as improved hearing aids and neural communication systems for those with auditory or speech impairments. Reconstructed results are available at https://neurofusex.github.io/c2f-ldm/.

Keywords brain-to-audio reconstruction; coarse-to-fine; fMRI; auditory processing pathway

1 Introduction

Brain-to-audio reconstruction, which aims to reconstruct audio stimuli from brain signals, represents a significant challenge with promising applications. In brain-computer interfaces, it enables the decoding of attended sound streams [1–4] in cocktail party scenarios where multiple sources coexist. As an assistive technology, it can enhance hearing aids and cochlear implants [5, 6] to improve auditory experiences for people with hearing impairments. Furthermore, by potentially translating silent speech imagination into audible output [7, 8], this technology opens new possibilities for neural communication systems. These capabilities demonstrate the potential of brain-to-audio reconstruction in advancing multimodal human-computer interaction and accessibility [9–12].

The common brain-to-audio reconstruction tasks can be categorized into *brain-to-sound* task [3, 13] for reconstructing all natural sounds in the environment, *brain-to-music* task [14–16] for music, and *brain-to-speech* task [17–22] for human voice, based on the different stimulus audios.

Some researchers first attempt to map brain signals to the spectrograms of stimulus audios using linear regression [14, 17–19]. Others introduce non-linear units and use simple networks such as MLP [14, 18, 19], BiLSTM [15, 20], and Transformer [20]. This approach can restore the overall temporal and frequency information of the spectrogram, but the reconstructed audio lacks semantic and detailed information, especially for non-invasive brain signals.

Recent neuroscience research has revealed a crucial insight: Deep Neural Network (DNN) features show stronger correlations with neural responses in the human brain compared to traditional acoustic representations like spectrograms [23–26]. Building on this finding, researchers [3, 21, 22] have developed approaches that first decode neural signals into DNN features as an intermediate representation before reconstructing the spectrogram using generative models. The intermediate representation is typically chosen from the intermediate layers of DNN [27, 28], serving as fine-grained features that contain both semantic and acoustic information of sound. However, decoding the fine-grained features is often challenging due to the high dimensionality, yielding limited outcomes in reconstruction.

There are also works that decode neural signals coarsely into the low-dimensional semantic space. For example, Denk et al. [16] decode fMRI data into 128-dimensional MuLan [29] embeddings, which are aligned with simple music descriptions in natural language, and then generates music

• Che Liu and Huiguang He are with the State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China and the School of Future Technology, University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: liuche2022@ia.ac.cn; huiguang.he@ia.ac.cn.
• Changde Du and Xiaoyu Chen are with the State Key Laboratory of Brain Cognition and Brain-inspired Intelligence Technology, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China and the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China. E-mail: changde.du@ia.ac.cn; chenxiaoyu2022@ia.ac.cn.
Manuscript received: 2025-08-30; revised: 2025-11-18; accepted: 2025-12-22

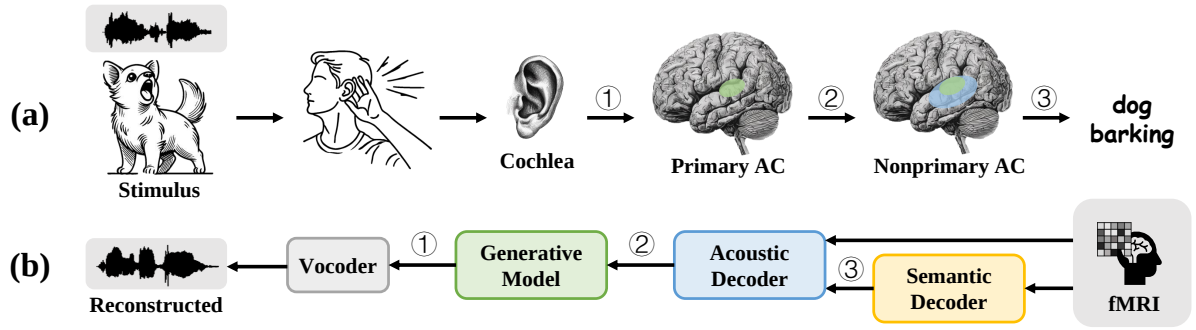

Fig. 1 (a) The hierarchical auditory processing pathway of humans. The stimulus audio is gradually decomposed into ① time-frequency representation, ② low-level acoustic features, and ③ high-level semantic characteristics. (b) The pipeline for our coarse-to-fine reconstruction from fMRI. Brain activity is decoded progressively into semantic, acoustic, and spectrogram levels, ultimately resulting in reconstructed audio.

using MusicLM [30]. Hence, Denk et al. primarily focus on decoding the semantic features within the music, while the acoustic details are largely inferred from the MusicLM's priors, resulting in limited reconstruction similarity. In addition, this model struggles to reconstruct audio beyond music and exhibits poor generalization capabilities.

To enhance the fine-grained decoding, we draw inspiration from neuroscientific findings. As shown in Fig. 1a, research has indicated that in the cochlea and subcortical structures of the human ear, the sound is decomposed into frequency-specific temporal patterns similar to spectrograms [31–34]. Further into the cerebral cortex, the human auditory system processes information along two hierarchical pathways, from low-level to high-level representations [35–39]. In recent years, an increasing amount of research has found that this cortical processing hierarchy aligns with the functional hierarchy of auditory DNN [23–26, 40–42]. The primary auditory cortex is more sensitive to shallow or intermediate DNN features, which represent low-level acoustic features, while the nonprimary auditory cortex is more sensitive to deep DNN features, which represent high-level semantic features.

Inspired by the acoustic-to-semantic stream, we propose a coarse-to-fine audio reconstruction method that reverses the auditory processing pathway, as shown in Fig. 1b. Using non-invasive fMRI as input, semantic features are first decoded, which we define as "coarse-grained" features, as they primarily capture high-level information. Then, guided by these semantic features, acoustic features are further decoded from fMRI that encompass both semantic content and acoustic details (e.g., rhythm, pauses), which we define as "fine-grained" features. After the coarse-to-fine decoding process, a conditional generative model is used to reconstruct the mel-spectrogram, followed by a vocoder to restore the original audio stimulus.

Unlike direct decoding methods that map fMRI directly to spectrograms and often produce overly smoothed outputs, our approach enables the generative model to reconstruct time–frequency details under the constraints of acoustic features. Compared with fine-grained decoding methods that directly regress high-dimensional latent representations and are highly sensitive to noise, our hierarchical coarse-to-fine design constrains the decoding space through low-dimensional semantic guidance, reducing the difficulty of acoustic decoding while preserving semantic and acoustic details.

To facilitate effective decoding and reconstruction across diverse stimuli, models pretrained on a wide range of audio samples are utilized, including natural sounds, music, and speech: CLAP [43], a contrastive audio-language model for semantic feature extraction; AudioMAE [44], a self-supervised audio model for comprehensive acoustic representation; and a Latent Diffusion Model[45] followed by a HiFi-GAN vocoder[46] for high-quality waveform synthesis.

The proposed approach is validated on three publicly available fMRI datasets with different kinds of audio stimuli: Brain2Sound [3], Brain2Music [47], and Brain2Speech [48]. Our model achieves state-of-the-art performance in metrics like Fréchet Distance (FD) and Fréchet Audio Distance (FAD). Experimental results show that our coarse-to-fine framework enhances the decoding of fine-grained audio embeddings and performs well across various datasets, showcasing its potential as a general framework.

Through extensive analysis of semantic information in decoded features, we reveal that while coarse-grained semantic guidance consistently improves overall reconstruction quality, its impact on semantic content varies with the quality of semantic features. It leads us to develop a conditional reconstruction method that further enhances performance by leveraging external semantic prompts.

Our contributions are as follows: (1) We propose a coarse-to-fine neural decoding model and reconstruct high-quality waveforms with both semantic and detailed information. We also confirm that coarse-to-fine decoding is superior to solely fine-grained decoding. (2) Our model achieves good results on datasets with three different kinds of stimuli, demonstrating its strong transferability. It can serve as a universal brain-to-audio framework. (3) We introduce an effective semantic prompting mechanism that improves reconstruction quality in challenging scenarios, providing a possible solution for handling imperfect neural signals. The code is released at https://github.com/cheee2000/C2F-LDM.

2 Method

Let $y \in \mathbb{R}^L$ represent an audio stimulus and $x \in \mathbb{R}^V$ represent the corresponding fMRI signal, where L is the length of the audio samples and V is the number of voxels in x. The brain-to-audio reconstruction process can be formulated as $\mathcal{R} : x \mapsto y$. Our approach is to first decode an intermediate representation c from x, and then generate y using a generative model $\mathcal{G}$ conditioned on c. To obtain the condition c, a coarse-to-fine process is adopted. First, a coarse-grained decoding is performed by a Semantic Decoder $\mathcal{D}^{Sem} : x \mapsto s$ to extract the semantic embedding s from fMRI. Then, a semantically-guided Acoustic Decoder $\mathcal{D}^{Aco} : (s, x) \mapsto c$ is employed to jointly decode the condition c with both semantics and acoustic details. After decoding, an LDM is utilized as the generative model $\mathcal{G} : c \mapsto y$ to reconstruct the stimulus audio conditioned on c. We will introduce the coarse-grained decoding process of $\mathcal{D}^{Sem}$ in Section 2.1.1, discuss the design of $\mathcal{D}^{Aco}$ and the fine-grained decoding process in Section 2.1.2, and describe the training of $\mathcal{G}$ in Section 2.2.

2.1 Coarse-to-fine brain decoding

2.1.1 *Coarse-grained semantic decoding*

CLAP feature is used as the coarse-grained semantic embedding of audio. CLAP, or contrastive language-audio pretraining [43], is a pretrained multi-modal model that aligns representations of audio with natural language descriptions. Pretrained on LAION-Audio-630K dataset [43] containing audios of human speech and song, natural sounds, and audio effects music, CLAP features are semantically aligned with various categories of audios, providing rich semantic information.

Semantic Decoder $\mathcal{D}^{Sem} : x \mapsto s$ is formulated as a ridge regression model. As shown in Fig. 2, the final-layer feature of CLAP's Audio Encoder is first used as the ground truth semantic feature of the stimulus audio y, denoted as $s_{gt} \in \mathbb{R}^{512}$. Then, L2-regularized linear regression is performed from x to s_{gt} using PyFastL2LiR[1] toolkit, which provides fast ridge regression and voxel selection functionalities. For each dimension of s_{gt}, we only select 500 voxels for regression based on the correlation coefficient. Thus, a sparse mapping matrix $W \in \mathbb{R}^{V \times 512}$ and a bias $b \in \mathbb{R}^{512}$ are obtained. The semantic embedding s of fMRI can be inferred by $s = xW + b$ and $s \in \mathbb{R}^{512}$.

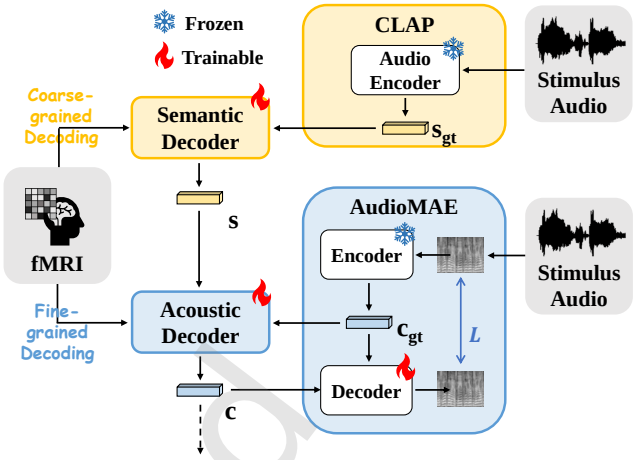

Fig. 2 (a) Coarse-to-fine brain decoding. In the **coarse-grained decoding**, fMRI is decoded into the semantic space of CLAP. In the **fine-grained decoding**, fMRI is decoded into the acoustic space of AudioMAE. (b) Detailed structure of Acoustic Decoder.

2.1.2 *Fine-grained acoustic decoding*

The AudioMAE latent feature serves as the fine-grained acoustic embedding of audio. AudioMAE, or audio mask autoencoder [44], is a self-supervised pretrained model, which consists of an encoder $\mathcal{E}^A$ and a decoder $\mathcal{D}^A$ and focuses on the reconstruction of the masked patches.

The reason we choose the AudioMAE latent embedding as the acoustic feature instead of other DNN features is threefold: (1) AudioMAE is trained on a generative task, which retains more low-level acoustic details compared to the discriminative models like VGGish-ish [27] used in Park et al. [3]. (2) Compared to the normal autoencoder used in Chen et al. [22], AudioMAE performs a masked patch prediction task, which models the whole patches of the spectrogram. The empirical evidence [49] shows that this makes the AudioMAE feature space more inclined to cluster audio of the same category together compared to VAE, indicating that AudioMAE better preserves high-level semantic information. (3) Pretrained on AudioSet-2M [50] which consists of natural sounds, human and animal sounds, and music, AudioMAE can work well in the general audio domain. In comparison, the MuLan [29] used in Denk et al. [16] and Wav2Vec 2.0 [28] used in Kim et

1 https://github.com/KamitaniLab/PyFastL2LiR

al . [21] can solely be utilized for music or speech. Considering all the points mentioned above, AudioMAE features are highly suitable for fine-grained features in our method, containing rich semantic and acoustic details.

As shown in Fig. 2, the stimulus waveform y is converted into mel-spectrogram m using 128 Kaldi [51] Mel-frequency bands following AudioMAE [44]. Then m is divided into 16×16 patches $m^p \in \mathbb{R}^{N_{patch}\times 256}$ and encoded into $c_{gt} = \mathcal{E}^A(m^p) \in \mathbb{R}^{N_{patch}\times 768}$ without masking, where N_{patch} represents the number of patches. c_{gt} is then decoded back to $m^p_{upp} = \mathcal{D}^A(c_{gt})$ and unpatchified to the mel-spectrogram m_{upp}. Here, we consider c_{gt} as the ground truth acoustic feature of the stimulus audio y and m_{upp} as an upper bound for the reconstructed mel-spectrogram.

Acoustic Decoder $\mathcal{D}^{Aco} : (s, x) \mapsto c$ is formulated as a Transformer-based model, which captures the dependencies between s and x, and decodes fMRI into the latent space of AudioMAE through a Seq2Seq generation. First, s and x are projected into the 768-dimensional representation space of the Transformer. s is projected to a semantic token s' through a linear layer. For x, 768 voxels with the highest responses are selected based on the mapping matrix W, forming the fMRI token x'. The tokens are then concatenated and encoded with a Transformer Encoder $\mathcal{E}^T$, yielding the neural embedding $n = \mathcal{E}^T([s', x'])$. A learnable embedding q is created as the query to a Transformer Decoder $\mathcal{D}^T$, with n serving as key and value, to obtain the decoded acoustic feature $c = \mathcal{D}^T(q, n)$.

Losses. $\mathcal{D}^{Aco}$ is trained from scratch using three loss functions: (1) $\mathcal{L}_{cond}$: L2 distance between c and c_{gt} in the latent space; (2) $\mathcal{L}_{perceptual}$: L2 distance between intermediate layer representations during the decoding process $m^p_{recon} = \mathcal{D}^A(c)$; and (3) $\mathcal{L}_{mel}$: L2 distance between the reconstructed mel-spectrogram m_{recon} and m_{upp}, where m_{recon} is obtained by unpatchifying m^p_{recon}. The overall loss is given by:

$$\mathcal{L} = \overbrace{\|c - c_{gt}\|_2^2}^{\mathcal{L}_{cond}} + \overbrace{\sum_{i\in layer} \|\mathcal{D}^A_i(c) - \mathcal{D}^A_i(c_{gt})\|_2^2}^{\mathcal{L}_{perceptual}} + \overbrace{\|m_{upp} - m\|_2^2 + \|m_{recon} - m\|_2^2}^{\mathcal{L}_{mel}}. \quad (1)$$

The pretrained AudioMAE is accustomed to handling masked patches, whereas our method leverages all patches to retain essential acoustic information for reconstruction. Therefore, $\mathcal{E}^A$ is frozen and $\mathcal{D}^A$ is fine-tuned to optimize the reconstruction performance.

Furthermore, we follow Liu et al. [49] by setting a $P_{gt} = 0.25$ during training, which means that $\mathcal{D}^{Aco}$ has a 0.25 probability of receiving the ground truth semantic feature s_{gt} as input and a 0.75 probability of receiving the decoded semantic feature s from $\mathcal{D}^{Sem}$. This trick helps reduce the impact of decoding noise and improve the stability of the reconstruction by bringing the decoded space closer to the original audio feature space. We will discuss it in Section 3.7.

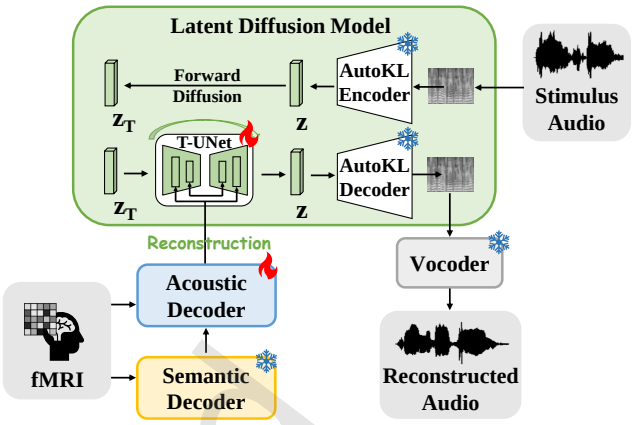

Fig. 3 Brain-to-audio reconstruction. The LDM generates mel-spectrograms under the condition of fine-grained acoustic features, followed by the Vocoder to generate reconstructed audios.

2.2 Brain-to-audio reconstruction

In this section, a generative model $\mathcal{G} : c \mapsto y$ is used to reconstruct the stimulus audio conditioned on c. When performing fine-grained decoding, although we use the AudioMAE Decoder to reconstruct the mel-spectrogram, it is not suitable to serve as the generative model for our method. We will discuss this in detail in Section 3.4. Instead, we model the process with a Latent Diffusion Model (LDM) [45]. LDM is a powerful generative model that can model complex data distributions in the latent space. It has been extensively used in the audio generation task, such as AudioLDM [52], AudioLDM2 [49] and DiffVoice [53].

The formulation in AudioLDM2 [49] is followed to implement the LDM. As shown in Fig. 3, we first use a Hanning window with 64 frequency bins, a window size of 1024, and a hop size of 160 to convert the stimulus audio into the mel-spectrogram. Then compress it to a latent representation z using a VAE. The forward diffusion process is a T steps Markov chain that gradually adds Gaussian noise as

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t}z_{t-1}, \beta_t\mathbf{I}) \quad (2)$$

where β_t is a variance schedule. Then the distribution of z_t given z_0 can be formulated as

$$q(z_t|z_0) = \prod_{s=1}^{t} q(z_s|z_{s-1}) = \mathcal{N}(z_t; \sqrt{\alpha_t}z_0, (1-\alpha_t)\mathbf{I}) \quad (3)$$

where $\alpha_t = \prod_{s=1}^{t}(1-\beta_s)$. The distribution of z_T at the final

step will be a standard Gaussian distribution [54]. The LDM learns a reverse denoising process from the prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$ to the data distribution z conditioned on c. The loss function [45, 54] in our method can be given as

$$\mathcal{L} = \mathbb{E}_{z_t, \epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I}), t \sim \{1,\dots,T\}}[\|\epsilon_\theta(z_t, t, c) - \epsilon\|_2^2] \quad (4)$$

where ϵ_θ is the denoising network, for which we utilize a Transformer-UNet (T-UNet) following AudioLDM2 [49]. After the LDM reconstructs the mel-spectrogram, it will be converted to the waveform using a pretrained HiFiGAN [46] vocoder. The LDM is initialized with the pretrained weights from AudioLDM2, and $\mathcal{D}^{Aco}$ and the T-UNet are fine-tuned during training, with other weights frozen.

2.3 Conditional reconstruction

Brain-to-audio reconstruction can sometimes leverage additional contextual conditions rather than being purely unconditional. First, the strong temporal correlations in audio signals enable semantically similar segments to serve as meaningful guidance when reconstructing specific target portions. Second, prior knowledge of simple sound categories (e.g., speech or animal sounds) can provide valuable semantic constraints for the reconstruction process. Based on these observations, we explore the potential of incorporating both audio and text prompts as semantic conditions to enhance the reconstruction quality.

The conditional reconstruction process is implemented as follows. CLAP's Text Encoder and Audio Encoder are used to extract the semantic embedding s_{prompt} of the text prompt and audio prompt. Then we replace s with s_{prompt} as input to $\mathcal{D}^{Aco} : (s_{prompt}, x) \mapsto c$, to obtain the fine-grained acoustic embedding c. Finally, we use $\mathcal{G} : c \mapsto y$ to reconstruct the stimulus audio conditioned on c.

3 Experiments

3.1 Experimental settings

3.1.1 Datasets

Three publicly available fMRI datasets are used to validate the method's performance across different kinds of stimuli: Brain2Sound [3], Brain2Music [47], and Brain2Speech [48] datasets.

Brain2Sound Dataset. The dataset[2] proposed by Park et al. [3] records the fMRI signals of five subjects (one female) while they are listening to natural sounds, including human speech, animal, musical instrument, and environmental sounds. fMRI data are acquired using a 3.0-Tesla Siemens MAGNETOM Verio scanner at the Kyoto University Institute for the Future of Human Society. Functional images that cover the entire brain are obtained with TR = 2,000 ms, TE = 44.8 ms, flip angle = 70 deg, FOV = 192 × 192 mm, voxel size=2 × 2 × 2 mm, number of slices = 76 and multi-band factor = 4. fMRI data preprocessed by Park et al. are utilized, primarily involving motion correction, slice time correction, co-registration, BOLD time-series resampling, etc.

Table 1 Number of voxels and samples in each dataset.

Dataset	ROI	Subject	#Voxel	#Train	#Test
Brain2Sound [3]	AC	S1	6,662	13,872	150
		S2	6,624	13,944	
		S3	6,713	13,944	
		S4	6,157	13,944	
		S5	7,143	13,944	
Brain2Music [47]	N/A	sub-001	60,784	4,800	600
		sub-002	53,927		
		sub-003	64,700		
		sub-004	61,899		
		sub-005	53,421		
Brain2Speech [48]	AC	UTS01	836	9,137	595
		UTS02	2,093		
		UTS03	1,303		
		UTS05	920		
		UTS06	980		
		UTS07	1,584		
		UTS08	1,109		

The stimuli consist of 1,250 8-s natural sound segments, with 1,200 for the training set and 50 for the test set, selected from the *VGGSound dataset* [55]. All the sounds are extracted from the videos uploaded to YouTube. To increase the sample number, audio segments are preprocessed in the same way as Park et al.: 4-s sliding windows are utilized with a 2-s stride to extract 3 4-s segments. All audio clips are resampled to 16kHz. During the collection of fMRI signals, each stimulus is repeated four times, resulting in 14,400 samples[3] for the training set (1,200 stimuli × 4 repetitions × 3 samples = 14,400 samples). For the test set, we average the multiple fMRI repetitions, resulting in 150 samples (50 stimuli × 3 samples = 150 samples).

Brain2Music Dataset. Following Denk et al. [16], we use the *music genre neuroimaging dataset*[4] from Nakai et al. [47], which records the fMRI signals of five subjects (two female) while they are listening to music clips. fMRI data are acquired using a 3.0T MRI scanner (TIM Trio; Siemens, Erlangen, Germany) at the Center for Information and Neural Networks (CiNet), National Institute of Information and Communications Technology (NICT), Osaka, Japan. Functional scanning

2 https://github.com/KamitaniLab/SoundReconstruction

3 When downloading, we discovered that some audios in the training set were no longer available on YouTube, hence, the amount of training samples is slightly less than 14,400.

4 https://openneuro.org/datasets/ds003720

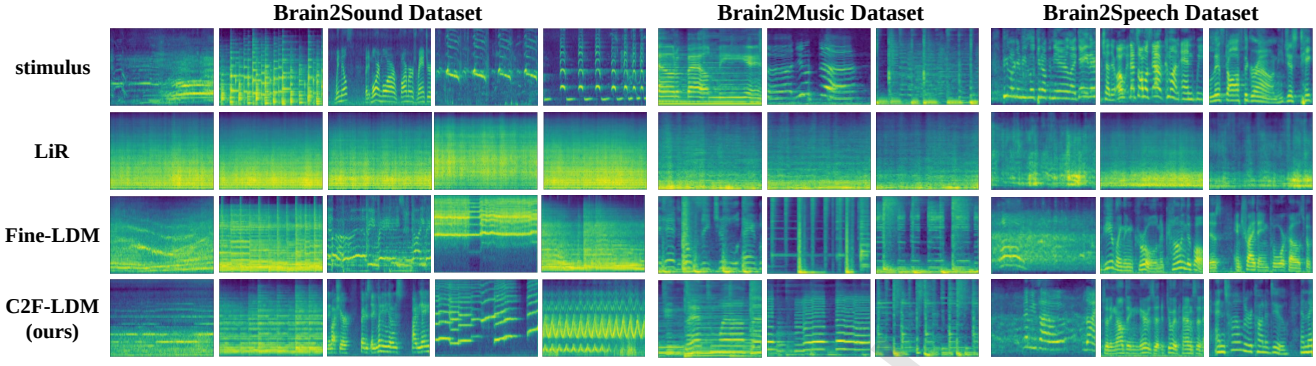

Fig. 4 Reconstruction results of S1, sub-001 and UTS01 on the three datasets.

is performed with TR = 1,500 ms, TE = 30 ms, flip angle = 62 deg, FOV = 192 × 192 mm, voxel size = 2 × 2 × 2 mm and multi-band factor = 4. fMRI data preprocessed by Denk et al. are utilized, encompassing essential steps such as motion correction, template alignment, low-frequency drift removal, response normalization, etc.

The dataset contains music stimuli from 10 genres (blues, classical, country, disco, hip-hop, jazz, metal, pop, reggae, and rock) which are sampled from the *GTZAN dataset* [56]. A total of 54 15-s music pieces are selected from each genre, with 48 for the training set and 6 for the test set. All music pieces are resampled to 16kHz and segmented into 10 clips of 1.5 seconds each to match the TR of functional scanning. As a result, the dataset consists of 4,800 samples (48 stimuli × 10 genres × 10 samples = 4,800 samples) for training and 600 (6 stimuli × 10 genres × 10 samples = 600 samples) for testing.

Brain2Speech Dataset. The dataset[5] proposed by LeBel et al. [48] contains fMRI responses recorded while 7 participants[6] (three female) are listening to 27 complete, natural, narrative stories. fMRI data are collected on a 3T Siemens Skyra scanner at the UT Austin Biomedical Imaging Center. Functional scans are collected with TR = 2.00 s, TE = 30.8 ms, flip angle = 71 deg, multi-band factor = 2, voxel size = 2.6 × 2.6 × 2.6 mm and FOV = 220 mm. fMRI data preprocessed by LeBel et al. are utilized, primarily incorporating motion correction, template creation and alignment, low-frequency drift removal, response normalization, etc.

The stimulus set consists of 27 10–15 minute stories from *The Moth* podcast. We select two stories (*Hang time* by a male speaker and *Where there's Smoke* by a female speaker) as the test set, and the remaining 25 stories are used as the training set. All stories are resampled to 16kHz and segmented into 2-s clips to match the TR of functional scanning. To account for the hemodynamic response, we form each sample pair by combining the fMRI signal at each TR with the corresponding stimulus audio clip from 4 seconds earlier. The last 10 stimulus audio clips from two stories in the test set are used as audio prompts. These prompts are not used for testing, ensuring that the participants could not have possibly heard the audio prompts in the preceding trials. As a result, the dataset consists of 9,137 samples for training and 595 samples for testing per subject.

5 https://openneuro.org/datasets/ds003020/versions/1.1.1
6 Subject UTS04 lacks a story, hence it will not be utilized.

3.1.2 *Metrics*

We use **PCC** (Pearson Correlation Coefficient) and **PSNR** (Peak Signal-to-Noise Ratio) to measure the similarity between the mel-spectrograms of reconstructed audio and stimulus audio, evaluating the low-level fidelity quality. In addition, we use FD, FAD, KL, and CLAP score, which are commonly employed in audio generation tasks, to evaluate the high-level perceptual quality of the reconstructed audio. **FD** (Fréchet Distance) calculates the distance in features between generated samples and target samples, extracted from an audio classifier PANNs [57]. **KL** (Kullback–Leibler divergence) calculates the KL divergence of classification logits based on PANNs. **FAD** (Fréchet Audio Distance) is similar to FD, but it uses VGGish [58]. **CLAP score** calculates the cosine similarity of CLAP [43] embeddings. In our experiments, each subject is trained and tested individually, and the metrics are averaged across subjects.

3.1.3 *Comparison models*

We compare the reconstruction results of three methods: (1) The direct decoding methods, which map fMRI signals to mel-spectrograms, including a linear regression model [14, 17–19] implemented through Ridge in sklearn, a three-layer MLP [14, 18, 19] implemented through MLPRegressor in sklearn following Bellier et al. [14], a Bidirectional LSTM [15, 20] and a Transformer Encoder [20], both with

the same configuration as our Transformer. (2) The fine-grained decoding methods, which map fMRI signals to high-dimensional intermediate features directly [3, 21, 22]. We remove the coarse-grained decoding process of our method and decode fMRI into the latent space of AudioMAE using the Acoustic Decoder $\mathcal{D}^{Aco}: x \mapsto c$. Then the LDM $\mathcal{G}: c \mapsto y$ is used to reconstruct the audio. This method is called *Fine-LDM*. In addition, for the Brain2Sound Dataset, codes and checkpoints open-sourced by Park et al. [3] are utilized to reproduce the experimental results. (3) The coarse-to-fine decoding methods proposed by us, including *C2F-Decoder*, which utilizes the AudioMAE Decoder as the generative model (see details in Section 3.4) and *C2F-LDM* using the LDM (ours).

3.1.4 *Experimental setup*

In the stage of coarse-grained decoding, only voxels from the auditory cortex (AC) area are utilized for the Brain2Sound and Brain2Speech datasets, whereas voxels from the entire brain are utilized for the Brain2Music Dataset. The specific brain regions and voxels can be found in Table 1.

In the stage of fine-grained decoding, we utilize a 4-layer Transformer Encoder and Decoder in $\mathcal{D}^{Aco}$ and use the default configuration of AudioMAE [44], initialized with the pretrained weights.[7] The AudioMAE Encoder is a vanilla 12-layer ViT-B, while the AudioMAE Decoder is a 16-layer Transformer with shifted local attention. Since AudioMAE requires 10-second audios (128×1024 mel-spectrograms) as inputs, the stimulus waveforms are duplicated to 10 seconds. After encoding with the AudioMAE Encoder $\mathcal{E}^{A}$, the embeddings of the first N_{patch} patches are selected as c_{gt}, corresponding to the length of the stimulus audio. N_{patch} is set to 208 for the Brain2Sound Dataset, 80 for the Brain2Music Dataset, and 112 for the Brain2Speech Dataset.

In the stage of brain-to-audio reconstruction, we follow the formulation in AudioLDM2 [49] to implement the LDM $\mathcal{G}$ and utilize two checkpoints[8] as the initialization weights: *audioldm2-full* for the Brain2Sound and Brain2Music datasets, and *audioldm2-speech-gigaspeech* for the Brain2Speech Dataset.

We use the AdamW [59] optimizer to train $\mathcal{D}^{Aco}$ and $\mathcal{D}^{A}$ with a learning rate of 1e-6, and train $\mathcal{G}$ with a learning rate of 1e-4. Models are trained on the Brain2Sound, Brain2Music, and Brain2Speech datasets with a batch size of 8 for 30, 40, and 30 epochs, respectively. All training is completed on a single NVIDIA A100 80GB GPU.

7 https://github.com/facebookresearch/AudioMAE
8 https://github.com/haoheliu/AudioLDM2

Table 2 Reconstruction results on the three datasets. **Bold** indicates the best, and underlined indicates that our method outperforms the fine-grained decoding methods.

Model	PCC↑	PSNR↑	FD↓	FAD↓	KL↓	CLAP↑
Brain2Sound Dataset [3]						
LiR	**0.607**	17.506	105.113	40.877	4.027	0.175
MLP	0.566	17.310	98.358	38.045	4.020	0.164
BiLSTM	0.580	17.381	112.031	39.895	3.948	0.180
Transformer	0.581	**17.676**	104.118	39.484	3.764	0.177
Park et al.	0.394	15.406	88.456	12.694	**2.251**	0.268
Fine-LDM	0.376	14.624	49.827	10.803	2.895	0.265
C2F-Decoder	0.595	17.385	95.565	35.775	3.748	0.179
C2F-LDM (**ours**)	0.418	15.103	**44.003**	**9.324**	2.697	**0.275**
Brain2Music Dataset [47]						
LiR	0.637	19.353	47.710	18.247	0.997	0.223
MLP	0.591	18.886	48.980	19.895	0.732	0.200
BiLSTM	0.628	19.078	57.030	22.673	1.008	0.209
Transformer	**0.646**	19.379	60.969	22.195	1.079	0.198
Fine-LDM	0.419	15.526	6.412	**1.273**	0.548	0.512
C2F-Decoder	0.643	**19.478**	63.039	26.053	1.191	0.195
C2F-LDM (**ours**)	0.454	15.883	**6.102**	1.504	**0.520**	**0.530**
Brain2Speech Dataset [48]						
LiR	0.511	17.500	68.146	24.988	3.483	0.112
MLP	0.409	16.389	75.174	27.983	4.153	0.094
BiLSTM	**0.526**	17.688	92.172	33.442	4.187	0.074
Transformer	0.526	**17.690**	74.048	27.526	3.817	0.041
Fine-LDM	0.357	14.385	12.706	4.820	0.885	0.420
C2F-Decoder	0.518	17.495	96.032	26.917	4.278	0.077
C2F-LDM (**ours**)	0.393	15.260	**9.726**	**4.623**	**0.616**	**0.471**

3.2 Reconstruction results

All reconstruction results are presented in Table 2, which are divided into three sections: direct decoding methods, fine-grained decoding methods, and our coarse-to-fine decoding methods. We select one representative from each section, *Linear Regression* (*LiR*), *Fine-LDM* and *C2F-LDM* to display the reconstructed mel-spectrograms[9] in Fig. 4.

It is found that direct decoding methods, which are optimized based on mean squared error, can achieve higher PCC and PSNR. However, as shown in Fig. 4, the reconstruction results are often overly smooth and lack high-frequency details, leading to poor perceptual quality.

In contrast, the fine-grained decoding methods exhibit a significant improvement in FD, FAD, KL, and CLAP score. The reconstructed spectrograms contain more time-frequency details, leading to more realistic reconstructed audios and a preliminary reconstruction of semantics. On the other hand, the fine-grained decoding methods fall short in terms of PCC and PSNR. In signal generation and reconstruction, a theoretical trade-off exists between perceptual quality and distortion metrics (such as PSNR) [60]. Improving perceptual quality often leads to lower PSNR values, and this trade-off is evident

9 The reconstructed audios can be accessed at https://neurofusex.github.io/c2f-ldm/.

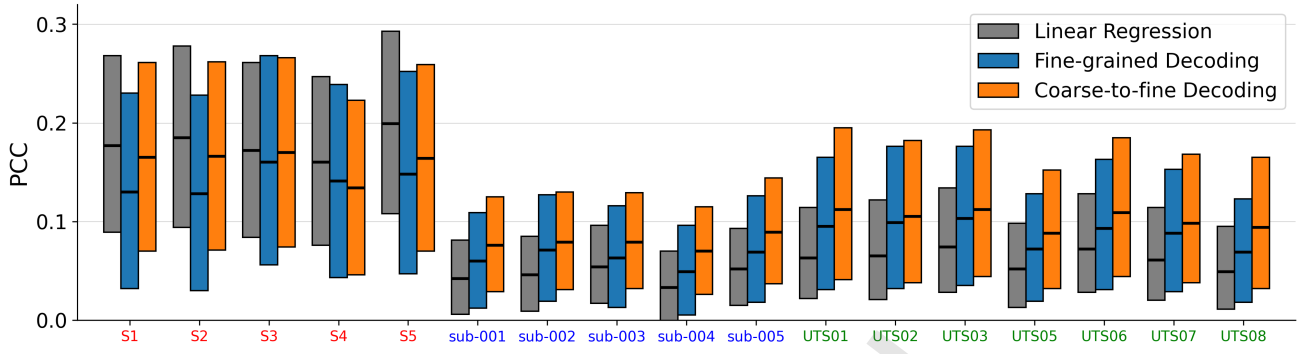

Fig. 5 PCC between the ground truth and decoded acoustic features for 17 subjects in the Brain2Sound, Brain2Music and Brain2Speech datasets. Our coarse-to-fine method consistently outperforms the directly fine-grained method.

in our reconstruction tasks. The use of generative models aims to achieve reasonable reconstruction accuracy while preserving high perceptual quality. Experimental results show that a PCC of approximately 0.4 is acceptable, considering the balance between perceptual quality and signal fidelity.

In comparison to fine-grained decoding methods, our coarse-to-fine approach excels in both low-level and high-level metrics, with spectrogram details closer to the stimulus audio. Our method achieves state-of-the-art performance in FD, FAD, KL, and CLAP score, while also enhancing PCC and PSNR, although it falls short of direct decoding methods. It demonstrates that coarse-to-fine decoding can effectively enhance the quality of reconstruction. Compared to Park et al. [3], our method significantly improves on PCC, FD, FAD, and CLAP score. The comparison of reconstructed samples can be found in Section 3.3. The comparison with *C2F-Decoder* can be found in Section 3.4.

We further analyze the impact of our coarse-to-fine method on decoding the fine-grained acoustic features. We compute the PCC between the ground truth and decoded acoustic features for 17 subjects across the three datasets. Then we compare the experimental results between the coarse-to-fine decoding and the directly fine-grained decoding, with a baseline established by directly mapping fMRI to the acoustic features using L2-regularized Linear Regression. As shown in Fig. 5, we visualize the interquartile range of the PCC across all feature dimensions using boxplots, which display the median and the first and third quartiles. Across almost all participants, our coarse-to-fine method consistently outperforms the fine-grained method. It suggests that coarse-to-fine decoding can effectively enhance the fine-grained acoustic features widely across the participants.

Notably, Brain2Sound achieves higher PCC than Brain2Music and Brain2Speech, which can be attributed to both neural and statistical factors. Natural environmental sounds evoke broader and more stable cortical activations and exhibit greater acoustic diversity, making them easier to decode from fMRI. In contrast, speech and music induce more localized and rapidly varying neural responses and show higher internal consistency with lower statistical variability, which together make them more difficult to predict.

3.3 Comparison with Park et al.

We reproduce the experimental results of Park et al. [3] using the features from the $conv5\_3$ layer of VGGish-ish [27] and voxels from the entire AC region. The results and the comparison with our method are shown in Fig. 6.

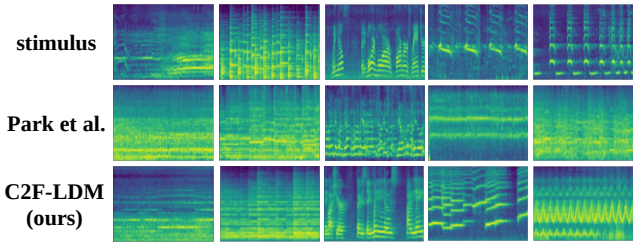

Fig. 6 Comparison with the reconstruction results of Park et al.

3.4 Comparison with *C2F-Decoder*

When performing fine-grained decoding, although we use the AudioMAE Decoder to reconstruct the mel-spectrogram, it is not suitable to serve as the generative model for our method. There are two main reasons for this: (1) The mel bins and window parameters of the mel-spectrograms in AudioMAE do not align with those of commonly used pretrained Vocoders. This mismatch prevents the generated mel-spectrograms from being directly converted into audio. Moreover, the cost of training a compatible Vocoder from scratch is prohibitively high. (2) The primary task of the AudioMAE Decoder is to predict masked patches, with a focus on low-level details of

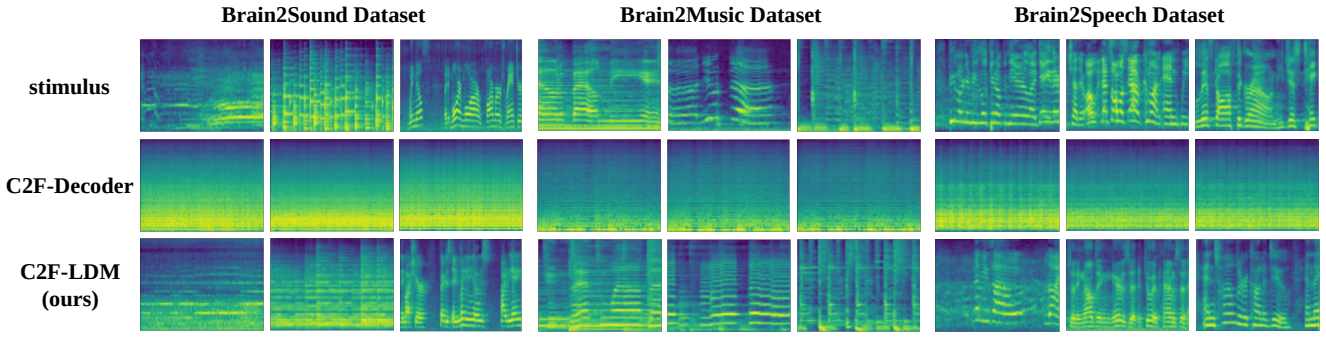

Fig. 7 Comparison of the reconstruction results between *C2F-Decoder* and *C2F-LDM*.

the spectrogram. This limitation leads to insufficient reconstruction quality in terms of semantic content. In contrast, the mel-spectrograms generated by LDM can be directly restored to audio using the pretrained HiFiGAN [46] vocoder, and the generated audio has richer semantic and acoustic details.

To investigate the reconstruction performance of *C2F-Decoder*, we need to transform the mel-spectrograms generated by the AudioMAE Decoder, denoted as m^A, into mel-spectrograms that the Vocoder can accept, denoted as m^V. We assume that m^A and m^V have a linear relationship, so we use an L2-regularized linear regression model trained on m^A and m^V of the stimulus audio in the training set. The results in the test set are as follows: PCC $= 0.938$ in the Brain2Sound Dataset, PCC $= 0.967$ in the Brain2Music and Brain2Speech datasets. Based on the results, we believe that this transformation is almost lossless. As shown in Fig. 7 and Table 2, *C2F-Decoder* is similar to the direct decoding methods in that they both focus on modeling the overall spectrograms but lack details and semantic information compared to *C2F-LDM*. It demonstrates the superiority of LDM over employing the AudioMAE Decoder directly.

3.5 Human rating

To complement the objective audio evaluation metrics, we design a subjective human rating experiment to assess the perceptual similarity between the reconstructed and the stimulus audio. Six human raters (R1-R6) participate in the evaluation. Each rater listens to the same 170 sets sampled from three datasets (17 subjects in total, 10 random samples per subject). Each set contains one stimulus audio together with multiple reconstructed audios generated by different methods: *C2F-LDM*, *Fine-LDM*, and for the Brain2Sound dataset, also the method of Park et al. [3]. The presentation order of the reconstructed audios in each set is fully randomized to prevent the raters from inferring the associated reconstruction method and to minimize potential bias.

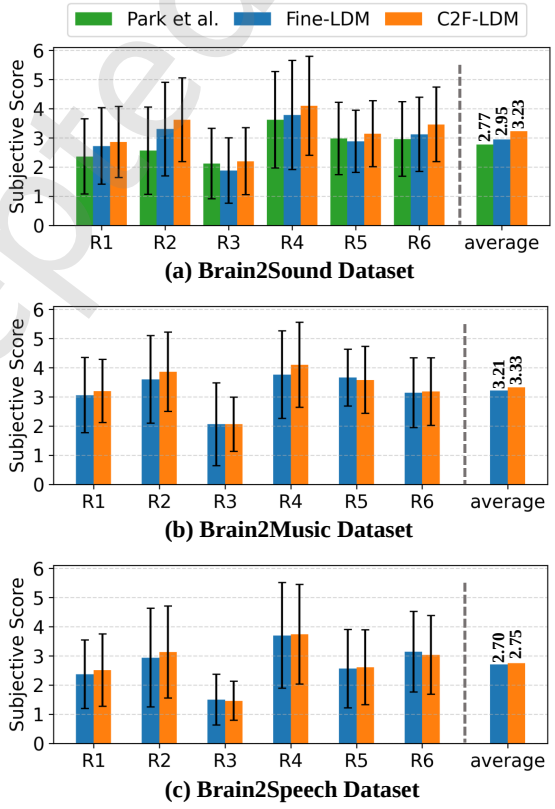

Fig. 8 Human rating results. Subjective scores from six human raters for each reconstruction method across the three datasets. Our proposed *C2F-LDM* achieves the highest subjective scores across most raters.

For each set, the raters independently assess the similarity between the stimulus audio and each reconstructed audio on a 7-point Likert scale [61–63] (1 = *completely dissimilar*, 3 = *slightly similar*, 5 = *moderately similar*, 7 = *highly similar*). Different reconstructed audios within the same set may receive identical scores when perceived equally similar.

We compute the mean subjective score for each method and each rater across the three datasets, as shown in Fig. 8. Overall, the scores center around 3, corresponding to a perception of *slightly similar*. The proposed *C2F-LDM* achieves the

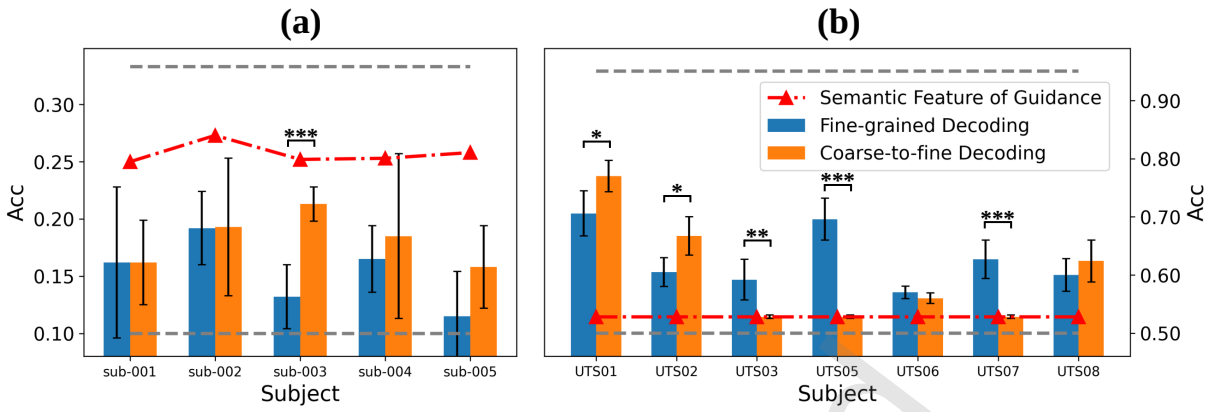

Fig. 9 Semantic decoding accuracy in the (a) Brain2Music and (b) Brain2Speech Dataset. The gray dashed lines represent the upper bound of accuracy and the chance level. Significance test is performed (paired t-test, $p < 0.001$(***), $p < 0.01$(**), $p < 0.05$(*)).

highest rating for most raters (6/6 on Brain2Sound, 5/6 on Brain2Music, and 4/6 on Brain2Speech). On the Brain2Sound and Brain2Music datasets, the average subjective scores of *C2F-LDM* across all raters reach 3.23 and 3.33 respectively, whereas the score on Brain2Speech is noticeably lower at 2.75. The perceptual advantage of *C2F-LDM* is most evident on Brain2Sound, where the method yields clearer semantic content and higher audio quality. In contrast, the two methods receive more comparable ratings on Brain2Speech, which is related to the overall lower perceived similarity observed for this dataset.

3.6 Semantic analysis of acoustic features

To better understand how coarse-grained semantic decoding improves the reconstruction, we evaluate its impact on the semantic content of decoded acoustic features. This analysis helps reveal whether the improvement comes from enhanced semantic information or other aspects of the acoustic features.

3.6.1 Experimental setup

It is generally believed that a representation space with strong semantic information typically clusters samples of the same category while separating those of different categories. Therefore, we conduct a classification experiment using two datasets with clear category labels, the Brain2Music Dataset (music genres, 10 classes) and the Brain2Speech Dataset (speaker genders, 2 classes).

Specifically for speech, we define audio semantics as high-level vocal characteristics (e.g., timbre, emotion) rather than transcribed content. Based on this definition, we consider speaker gender as a semantic label for two main reasons. From a neuroscientific perspective, gender perception engages higher-order auditory regions (STG, STS) that integrate timbre, emotion, and social cognition [64, 65]. From a modeling perspective, the LAION-Audio-630K dataset used to train CLAP explicitly encodes gender information (like "Woman saying yeah"), making gender a practical and well-supported choice.

We perform 5-fold cross-validation on the test set, using SVM to classify acoustic features obtained through coarse-to-fine decoding or directly fine-grained decoding. The average classification accuracy measures the semantic information in the acoustic features, with identical experimental conditions ensuring an unbiased comparison.

3.6.2 Results and analysis

Figure 9 summarizes the results. The chance levels are 0.1 and 0.5 for the two datasets, while the upper bounds are the classification accuracies obtained on the ground truth features, 0.33 and 0.95. After introducing the coarse-grained decoding, classification accuracy in the Brain2Music Dataset either improves or remains stable, whereas in the Brain2Speech Dataset some subjects show decreased accuracy.

This contrasting behavior prompts us to examine the quality of the coarse-grained semantic features used for guidance. When semantic features are of low quality, they might fail to enhance semantic information in the acoustic features while still improving other acoustic properties, resulting in better overall reconstruction despite decreased semantic content.

3.6.3 Interpretation

To further investigate this issue, we evaluate the quality of the coarse-grained semantic features by performing the same SVM classification task. As shown in Fig. 9, the classification accuracy of the decoded semantic features in the Brain2Music Dataset is relatively high (0.257 ± 0.008, compared with an upper bound of 0.518), whereas that in the Brain2Speech

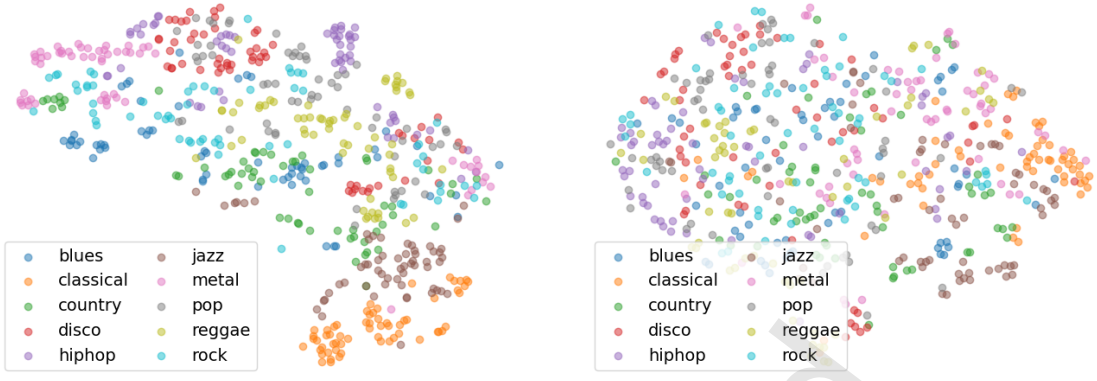

Fig. 10 Visualization on the space of ground truth (left) and decoded (right) semantic features on the Brain2Music Dataset.

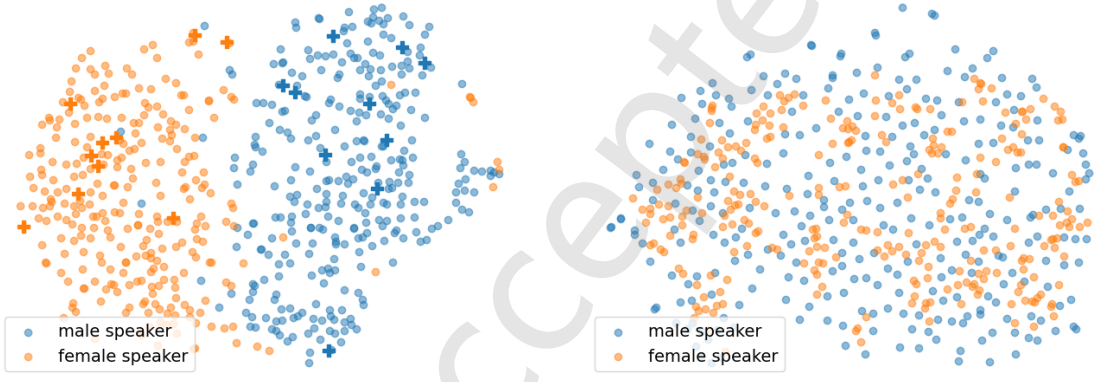

Fig. 11 Visualization on the space of ground truth (left) and decoded (right) semantic features on the Brain2Speech Dataset. The cross markers represent samples used as audio prompts, providing coarse-grained semantic features in the conditional reconstruction.

Dataset is poor (0.528 ± 0.000, upper bound 0.961), approaching the chance level. It indicates that there is little semantic content in the semantic features, making it nearly impossible to differentiate speakers' genders.

We further perform t-SNE visualization of the ground truth and decoded semantic feature spaces, using sub-003 and UTS05 as examples that show significant effects after incorporating the semantic features (Fig. 9). For the Brain2Music Dataset, some genres like *classical* and *jazz* exhibit good decoding performance, while others like *rock* show poor decoding performance, as shown in Fig. 10. For the Brain2Speech Dataset, the semantic decoding performance is poor, and it cannot differentiate between male and female speakers, as shown in Fig. 11. This finding explains why the semantic content in the acoustic features decreases for some subjects in the Brain2Speech Dataset: when the guiding semantic features lack clear category information, they cannot effectively enhance the semantic aspects of the acoustic features.

In summary, our analysis reveals that the effectiveness of coarse-to-fine decoding in enhancing semantic information depends critically on the quality of the decoded semantic features. When semantic features are well-structured (as in the Brain2Music Dataset), they can enhance both semantic and acoustic aspects of the reconstruction. However, when semantic features are poorly decoded (as in the Brain2Speech Dataset), the framework still improves overall reconstruction quality by enhancing low-level acoustic properties, even though it may not strengthen semantic information.

3.7 Conditional reconstruction results

As discussed in the previous section, the performance of semantic decoding can occasionally be suboptimal. To address this, in the conditional reconstruction task described in Section 2.3, we propose using prompt-based semantic features as coarse-grained representations, replacing the decoded features. We define two types of prompts: (1) Text prompts, which specify the stimulus audio category. For the Brain2Music Dataset, these include 10 music genres (e.g., *pop music*, *rock music*), while for the Brain2Speech Dataset, they indicate the speaker's gender (*man speaking* or *woman*

Table 3 Conditional reconstruction results with text prompts and audio prompts in the Brain2Music and Brain2Speech Dataset.

P_{gt}	Prompt	PCC↑	PSNR↑	FD↓	FAD↓	KL↓	CLAP↑
Brain2Music Dataset [47]							
0.0	no prompt	0.442	15.827	6.121	1.606	0.520	0.527
	text prompt	0.421	15.600	8.283	2.169	0.641	0.460
0.25	no prompt	**0.454**	15.883	**6.102**	1.504	**0.520**	**0.530**
	text prompt	0.405	**16.059**	7.358	2.219	0.584	0.470
0.5	no prompt	0.422	15.466	6.587	**1.341**	0.536	0.513
	text prompt	0.393	15.827	10.304	2.624	0.559	0.465
Brain2Speech Dataset [48]							
0.0	no prompt	0.379	14.898	11.636	4.866	0.758	0.438
	text prompt	0.331	14.531	10.772	5.265	0.513	0.444
	audio prompt	0.315	14.543	7.160	4.148	0.430	0.481
0.25	no prompt	**0.393**	15.260	9.726	4.623	0.616	0.471
	text prompt	0.340	14.680	9.265	4.712	0.449	0.476
	audio prompt	0.300	14.434	7.722	4.383	0.416	0.491
0.5	no prompt	0.374	**15.299**	7.957	4.013	0.493	**0.502**
	text prompt	0.350	14.912	**6.698**	5.102	**0.348**	0.487
	audio prompt	0.303	14.401	7.129	**3.930**	0.398	0.486

speaking). (2) Audio prompts, where the last 10 stimulus audio clips from two test set stories in the Brain2Speech Dataset are used as prompts, with results averaged.

To evaluate the conditional brain-to-audio reconstruction, we test the hyperparameter P_{gt} mentioned in Section 2.1.2 under three conditions: $P_{gt} = 0.0$, 0.25 and 0.5. The results are shown in Table 3. We observe that incorporating ground truth semantic features during training improves brain-to-audio reconstruction, but higher P_{gt} values do not always yield better results. For example, across the two datasets, all reported metrics are higher at $P_{gt} = 0.25$ than at $P_{gt} = 0.0$, confirming the benefit of limited ground truth guidance. However, further increasing P_{gt} to 0.5 leads to a decline in several metrics, such as PCC and the perceptual metrics on the Brain2Music Dataset. Since fMRI data is used exclusively during testing, excessively high P_{gt} creates a mismatch between training and testing, reducing reconstruction accuracy. Considering both datasets and all metrics, $P_{gt} = 0.25$ achieves the best overall performance and stability, and is adopted as the optimal setting.

For the Brain2Music Dataset, we observe a decrease in high-level metrics after incorporating text prompts. This unexpected result can be explained by two factors: (1) The simple genre labels may not capture the rich semantic variations within each music category, and (2) Using identical semantic features for all samples within the same genre eliminates sample-specific semantic information that was previously well-preserved by the decoded semantic features, according to Section 3.6. This loss of semantic specificity particularly affects the reconstruction quality at the semantic level. In contrast, the semantic decoding performs poorly in the Brain2Speech Dataset, so the introduction of both text prompts and audio prompts can significantly enhance the semantics of the reconstructed audio.

In summary, our conditional reconstruction experiments demonstrate that external semantic guidance can be an effective alternative when decoded semantic features are unreliable. However, the choice of semantic guidance should be carefully considered. While simple category-level prompts may suffice for datasets with poor semantic decoding, more detailed semantic representations may be necessary for datasets where fine-grained semantic information is important for high-quality reconstruction.

4 Limitations and Future Works

Temporal resolution. Given the advantages of high spatial resolution and high signal-to-noise ratio in non-invasive neural signals, fMRI has been commonly employed in the field of neural encoding and decoding. Research [13] has confirmed that the reconstruction from the BOLD response (TR=2.6s) can exhibit a temporal specificity of about 200 ms, sufficient for capturing essential auditory semantics. Since then, numerous works [3, 13, 16] and datasets [3, 47, 48, 66, 67] have emerged to support research on fMRI-to-audio tasks. However, the limited temporal resolution of fMRI consistently hampers the temporal decoding of audio. The aim of this study is to further enhance reconstruction performance from a neuroscientific perspective. To make a breakthrough in temporal decoding, it is imperative to leverage other neural signals with high temporal resolution, such as EEG and MEG.

It is worth noting that our hierarchical coarse-to-fine architecture is well suited for modality transfer beyond fMRI. The semantic representations are low-dimensional and relatively robust to noise and sparse spatial sampling, which helps maintain stable latent estimation across different neural signals. In addition, modality-specific adaptations such as source localization, spatial filtering, and region-informed channel pooling can enhance the extraction of auditory-related spatial information and support semantic decoding. Moreover, the higher temporal precision of EEG or MEG can be exploited by adjusting the acoustic decoder to capture rapid temporal dynamics while maintaining sufficient spectral resolution. These properties indicate that the proposed framework provides a feasible foundation for extending brain-to-audio reconstruction to neural modalities with higher temporal resolution.

Model and voxel selection. The main purpose of this article is to illustrate the superiority of hierarchical decoding over direct decoding. Therefore, we build a generic brain-to-audio

framework, selecting the most suitable models, CLAP and AudioMAE, without comparing them to other representation models. In the future, we will explore both model-level improvements and data-centric enhancements, such as incorporating more balanced or attribute-enriched training data, to further improve reconstruction performance within this framework. Furthermore, we utilize all the voxels of the auditory cortex (AC) in our work. However, there are gradients in the voxels of different brain regions within the AC [23, 25, 26]. In the future, we plan to consider the gradients of voxels to further enhance the hierarchy of information processing.

5 Conclusion

In this paper, we propose a novel coarse-to-fine framework for audio reconstruction from fMRI that reverses the hierarchical processing pathway of the human auditory system. Comprehensive experiments demonstrate that our decoding strategy significantly outperforms traditional fine-grained approaches, achieving state-of-the-art results on multiple evaluation metrics. Our findings underscore the value of incorporating neuroscientific principles into technical solutions, suggesting that mimicking biological processing pathways can lead to more effective neural decoding methods.

Acknowledgment

This work was supported in part by the Scientific and Technological Innovation (STI) 2030–Major Projects under Grant 2021ZD0201503 and the Lingang Laboratory under Grant No. LGL-1987-07. The authors would like to thank J.-Y. Park, T. Nakai, and A. LeBel for sharing the fMRI data.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Reference

[1] T. De Taillez, B. Kollmeier, and B. T. Meyer, Machine learning for decoding listeners' attention from electroencephalography evoked by continuous speech, *European Journal of Neuroscience*, vol. 51, no. 5, pp. 1234–1241, 2020.

[2] Z. Xu, Y. Bai, R. Zhao, Q. Zheng, G. Ni, and D. Ming, Auditory attention decoding from eeg-based mandarin speech envelope reconstruction, *Hearing Research*, vol. 422, p. 108552, 2022.

[3] J.-Y. Park, M. Tsukamoto, M. Tanaka, and Y. Kamitani, Sound reconstruction from human brain activity via a generative model with brain-like auditory features, *PLOS Biology*, vol. 23, no. 7, p. e3003293, 2025.

[4] X. Chen, C. Du, Q. Zhou, and H. He, Auditory attention decoding with task-related multi-view contrastive learning, in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6025–6033.

[5] Z. Wang, N. Shi, Y. Zhang, N. Zheng, H. Li, Y. Jiao, J. Cheng, Y. Wang, X. Zhang, Y. Chen *et al.*, Conformal in-ear bioelectronics for visual and auditory brain-computer interfaces, *Nature Communications*, vol. 14, no. 1, p. 4213, 2023.

[6] M. A. Tanveer, M. A. Skoglund, B. Bernhardsson, and E. Alickovic, Deep learning-based auditory attention decoding in listeners with hearing impairment, *Journal of Neural Engineering*, vol. 21, no. 3, p. 036022, 2024.

[7] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, Speech synthesis from neural decoding of spoken sentences, *Nature*, vol. 568, no. 7753, pp. 493–498, 2019.

[8] J. Tang, A. LeBel, S. Jain, and A. G. Huth, Semantic reconstruction of continuous language from non-invasive brain recordings, *Nature Neuroscience*, vol. 26, no. 5, pp. 858–866, 2023.

[9] C. Du, K. Fu, J. Li, and H. He, Decoding visual neural representations by multimodal learning of brain-visual-linguistic features, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 10 760–10 777, 2023.

[10] C. Du, K. Fu, B. Wen, Y. Sun, J. Peng, W. Wei, Y. Gao, S. Wang, C. Zhang, J. Li *et al.*, Human-like object concept representations emerge naturally in multimodal large language models, *Nature Machine Intelligence*, pp. 1–16, 2025.

[11] S. Li, P. Wang, X. Yu, P. Xia, X. Chen, L. Du, Y. Chen, and Z. Fang, Ganet: A convolution neural network with parallel convolutions and graph-based attention mechanism for event-related potential classification in brain-computer interface task, *Tsinghua Science and Technology*, vol. 31, no. 2, pp. 920–931, 2026.

[12] C. Tang, D. Jiang, Y. Guo, L. Chen, and B. Chen, Copula transfer entropy-based channel selection for meg motor imagery brain computer interfaces, *Tsinghua Science and Technology*, vol. 31, no. 3, pp. 1474–1486, 2026.

[13] R. Santoro, M. Moerel, F. De Martino, G. Valente, K. Ugurbil, E. Yacoub, and E. Formisano, Reconstructing the spectrotemporal modulations of real-life sounds from fmri response patterns, *Proceedings of the National Academy of Sciences*, vol. 114, no. 18, pp. 4799–4804, 2017.

[14] L. Bellier, A. Llorens, D. Marciano, A. Gunduz, G. Schalk, P. Brunner, and R. T. Knight, Music can be reconstructed from human auditory cortex activity using nonlinear decoding models, *PLOS Biology*, vol. 21, no. 8, p. e3002176, 2023.

[15] I. Daly, Neural decoding of music from the eeg, *Scientific Reports*, vol. 13, no. 1, p. 624, 2023.

[16] T. I. Denk, Y. Takagi, T. Matsuyama, A. Agostinelli, T. Nakai, C. Frank, and S. Nishimoto, Brain2music: Reconstructing music from human brain activity, *arXiv preprint arXiv:2307.11078*, 2023.

[17] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, Recon-

structing speech from human auditory cortex, *PLOS Biology*, vol. 10, no. 1, p. e1001251, 2012.

[18] M. Yang, S. A. Sheth, C. A. Schevon, G. M. McKhann II, and N. Mesgarani, Speech reconstruction from human auditory cortex with deep neural networks. in *Interspeech*, 2015, pp. 1121–1125.

[19] H. Akbari, B. Khalighinejad, J. L. Herrero, A. D. Mehta, and N. Mesgarani, Towards reconstructing intelligible speech from the human auditory cortex, *Scientific reports*, vol. 9, no. 1, p. 874, 2019.

[20] K. Shigemi, S. Komeiji, T. Mitsuhashi, Y. Iimura, H. Suzuki, H. Sugano, K. Shinoda, K. Yatabe, and T. Tanaka, Synthesizing speech from ecog with a combination of transformer-based encoder and neural vocoder, in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[21] M. Kim, Z. Piao, J. Lee, and H.-G. Kang, 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)*. IEEE, 2023, pp. 1–5.

[22] X. Chen, R. Wang, A. Khalilian-Gourtani, L. Yu, P. Dugan, D. Friedman, W. Doyle, O. Devinsky, Y. Wang, and A. Flinker, A neural speech decoding framework leveraging deep learning and speech synthesis, *Nature Machine Intelligence*, vol. 6, no. 4, pp. 467–480, 2024.

[23] A. J. Kell, D. L. Yamins, E. N. Shook, S. V. Norman-Haignere, and J. H. McDermott, A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy, *Neuron*, vol. 98, no. 3, pp. 630–644, 2018.

[24] Y. Li, G. K. Anumanchipalli, A. Mohamed, P. Chen, L. H. Carney, J. Lu, J. Wu, and E. F. Chang, Dissecting neural computations in the human auditory pathway using deep neural networks for speech, *Nature Neuroscience*, vol. 26, no. 12, pp. 2213–2225, 2023.

[25] G. Tuckute, J. Feather, D. Boebinger, and J. H. McDermott, Many but not all deep neural network audio models capture brain responses and exhibit correspondence between model stages and brain regions, *PLOS Biology*, vol. 21, no. 12, p. e3002366, 2023.

[26] B. L. Giordano, M. Esposito, G. Valente, and E. Formisano, Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds, *Nature Neuroscience*, vol. 26, no. 4, pp. 664–672, 2023.

[27] V. Iashin and E. Rahtu, Taming visually guided sound generation, in *British Machine Vision Conference (BMVC)*, 2021.

[28] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[29] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. Ellis, Mulan: A joint embedding of music audio and natural language, in *23rd International Society for Music Information Retrieval Conference (ISMIR)*, 2022, pp. 559–566.

[30] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, Musiclm: Generating music from text, *arXiv preprint arXiv:2301.11325*, 2023.

[31] J. Pickles, An introduction to the physiology of hearing, 1988.

[32] S. A. Shamma and C. Micheyl, Behind the scenes of auditory perception, *Current opinion in neurobiology*, vol. 20, no. 3, pp. 361–366, 2010.

[33] J. Schnupp, *Auditory Neuroscience: Making Sense of Sound*. MIT Press, 2011.

[34] B. C. Moore, *An introduction to the psychology of hearing*. Brill, 2012.

[35] J. P. Rauschecker and B. Tian, Mechanisms and streams for processing of "what" and "where" in auditory cortex, *Proceedings of the National Academy of Sciences*, vol. 97, no. 22, pp. 11 800–11 806, 2000.

[36] J. H. Kaas and T. A. Hackett, Subdivisions of auditory cortex and processing streams in primates, *Proceedings of the National Academy of Sciences*, vol. 97, no. 22, pp. 11 793–11 799, 2000.

[37] S. K. Scott and I. S. Johnsrude, The neuroanatomical and functional organization of speech perception, *Trends in neurosciences*, vol. 26, no. 2, pp. 100–107, 2003.

[38] G. Hickok and D. Poeppel, The cortical organization of speech processing, *Nature reviews neuroscience*, vol. 8, no. 5, pp. 393–402, 2007.

[39] J. P. Rauschecker and S. K. Scott, Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing, *Nature neuroscience*, vol. 12, no. 6, pp. 718–724, 2009.

[40] U. Güçlü, J. Thielen, M. Hanke, and M. Van Gerven, Brains on beats, *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[41] A. R. Vaidya, S. Jain, and A. Huth, Self-supervised models of audio effectively explain human cortical responses to speech, in *International Conference on Machine Learning*, 2022, pp. 21 927–21 944.

[42] J. Millet, C. Caucheteux, Y. Boubenec, A. Gramfort, E. Dunbar, C. Pallier, J.-R. King *et al.*, Toward a realistic model of speech processing in the brain with self-supervised learning, *Advances in Neural Information Processing Systems*, vol. 35, pp. 33 428–33 443, 2022.

[43] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, 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)*. IEEE, 2023, pp. 1–5.

[44] P.-Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, Masked autoencoders that

listen, *Advances in Neural Information Processing Systems*, vol. 35, pp. 28 708–28 720, 2022.

[45] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, High-resolution image synthesis with latent diffusion models, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[46] J. Kong, J. Kim, and J. Bae, Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis, *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.

[47] T. Nakai, N. Koide-Majima, and S. Nishimoto, Music genre neuroimaging dataset, *Data in Brief*, vol. 40, p. 107675, 2022.

[48] A. LeBel, L. Wagner, S. Jain, A. Adhikari-Desai, B. Gupta, A. Morgenthal, J. Tang, L. Xu, and A. G. Huth, A natural language fmri dataset for voxelwise encoding models, *Scientific Data*, vol. 10, no. 1, p. 555, 2023.

[49] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, Audioldm 2: Learning holistic audio generation with self-supervised pretraining, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2871–2883, 2024.

[50] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, Audio set: An ontology and human-labeled dataset for audio events, in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.

[51] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, The kaldi speech recognition toolkit, in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[52] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, Audioldm: Text-to-audio generation with latent diffusion models, in *International Conference on Machine Learning*, 2023, pp. 21 450–21 474.

[53] Z. Liu, Y. Guo, and K. Yu, Diffvoice: Text-to-speech with latent diffusion, in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[54] J. Ho, A. Jain, and P. Abbeel, Denoising diffusion probabilistic models, *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[55] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, Vggsound: A large-scale audio-visual dataset, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[56] G. Tzanetakis and P. Cook, Musical genre classification of audio signals, *IEEE Transactions on speech and audio processing*, vol. 10, no. 5, pp. 293–302, 2002.

[57] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, Panns: Large-scale pretrained audio neural networks for audio pattern recognition, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[58] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi, Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms, in *Proc. Interspeech*, 2019, pp. 2350–2354.

[59] I. Loshchilov and F. Hutter, Decoupled weight decay regularization, in *International Conference on Learning Representations*, 2017.

[60] Y. Blau and T. Michaeli, The perception-distortion tradeoff, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6228–6237.

[61] R. Likert, A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[62] S. Rimbert, N. Gayraud, L. Bougrain, M. Clerc, and S. Fleck, Can a subjective questionnaire be used as brain-computer interface performance predictor? *Frontiers in human neuroscience*, vol. 12, p. 529, 2019.

[63] H. Pan, P. Ding, F. Wang, T. Li, L. Zhao, W. Nan, Y. Fu, and A. Gong, Comprehensive evaluation methods for translating bci into practical applications: usability, user satisfaction and usage of online bci systems, *Frontiers in Human Neuroscience*, vol. 18, p. 1429130, 2024.

[64] P. Belin, R. J. Zatorre, P. Lafaille, P. Ahad, and B. Pike, Voice-selective areas in human auditory cortex, *Nature*, vol. 403, no. 6767, pp. 309–312, 2000.

[65] S. Lattner, M. E. Meyer, and A. D. Friederici, Voice perception: sex, pitch, and the right hemisphere, *Human brain mapping*, vol. 24, no. 1, pp. 11–20, 2005.

[66] S. A. Nastase, Y.-F. Liu, H. Hillman, A. Zadbood, L. Hasenfratz, N. Keshavarzian, J. Chen, C. J. Honey, Y. Yeshurun, M. Regev *et al.*, The "narratives" fmri dataset for evaluating models of naturalistic language comprehension, *Scientific data*, vol. 8, no. 1, p. 250, 2021.

[67] J. Li, S. Bhattasali, S. Zhang, B. Franzluebbers, W.-M. Luh, R. N. Spreng, J. R. Brennan, Y. Yang, C. Pallier, and J. Hale, Le petit prince multilingual naturalistic fmri corpus, *Scientific data*, vol. 9, no. 1, p. 530, 2022.

Author biography

Che Liu received the B.S. degree from Beihang University, China, in 2022. He is currently pursuing the Ph.D. degree with the Institute of Automation, Chinese Academy of Sciences, China. His research interests include neural encoding and decoding, as well as multimodal representation learning.

Changde Du received the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2019. He is currently an Associate Professor with CASIA. His current research interests include deep learning, computational neuroscience, brain-inspired intelligence, computer vision and brain-computer interfaces. He has published over 40 peer-reviewed research papers in prestigous conferences and journals. He won the following awards: National Scholarship for Doctoral Students (2018), President Prize of Chinese Academy of Sciences for Excellent Ph.D. Graduates (2019). His homepage: https://changdedu.github.io/.

Xiaoyu Chen received the B.S. degree in Communication Engineering from Xidian University, China, in 2016, and the M.S. degree in Electronic and Communication Engineering from Northwest University, China, in 2020. He is currently pursuing the Ph.D. degree with the Institute of Automation, Chinese Academy of Sciences, China. His research interests include neural encoding and decoding, as well as large language models (LLMs).

Huiguang He (Senior Member, IEEE) received the B.S. and M.S. degrees from Dalian Maritime University (DMU), Dalian, China, in 1994 and 1997, respectively, and the Ph.D. degree (Hons.) in pattern recognition and intelligent systems from the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China, in 2002. He was an Associate Lecturer with DMU from 1997 to 1999 and a Post-Doctoral Researcher with the University of Rochester, Rochester, NY, USA, from 2003 to 2004. He was a Visiting Professor with the University of North Carolina at Chapel Hill, Chapel Hill, NC, USA, from 2014 to 2015. He is currently a Full Professor with CASIA. His research has been supported by several research grants from the National Science Foundation of China. His current research interests include pattern recognition, brain–computer interfaces and medical image analysis. He has authored or coauthored more than 200 peer-reviewed papers including Nature Machine Intelligence, IEEE TPAMI, ICML, etc.