MINDVOICE: TOWARDS INTELLIGIBLE SPEECH RECONSTRUCTION FROM NON-INVASIVE NEURAL SIGNALS VIA LINGUISTIC RECOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Reconstructing speech from neural activity offers a direct window into human auditory perception and holds promise for safe and scalable speech brain-computer interfaces. Yet intelligible reconstruction from non-invasive neural signals remains challenging. Existing methods are predominantly acoustic-centric, directly mapping noisy neural recordings to entangled speech representations, which can preserve acoustic characteristics while failing to reliably recover the underlying linguistic content. We introduce *MindVoice*, a neural speech reconstruction framework that addresses this *intelligibility gap* through explicit linguistic recovery. A linguistic reconstruction stream estimates what was perceived, while a complementary acoustic stream captures fine-grained characteristics of how it was spoken. Leveraging pretrained speech priors to compensate for incomplete neural evidence, *MindVoice* integrates both streams through text-to-speech generation and in-context voice cloning to synthesize natural, intelligible speech. Experiments on EEG and MEG datasets demonstrate substantial intelligibility gains over existing non-invasive methods. Further analyses reveal an empirical trade-off between spectral fidelity and linguistic intelligibility, with *MindVoice* achieving a favorable balance. Together, these findings support explicit linguistic recovery as an effective route toward intelligible non-invasive neural speech reconstruction.

1 INTRODUCTION

Auditory perception is central to human cognition, encoding rich acoustic information—including pitch, intonation, speaker characteristics, and emotional prosody—and linguistic content through multi-level neural processing. For decades, researchers have investigated how the brain represents and comprehends complex speech (Hickok & Poeppel, 2007; Rauschecker & Scott, 2009; Davis & Johnsrude, 2003). Directly reconstructing perceived speech from neural activity offers an intuitive way to probe these representations and lays the groundwork for next-generation speech brain-computer interfaces. Despite substantial progress in invasive speech decoding and reconstruction (Mesgarani et al., 2014; Wairagkar et al., 2025; Anumanchipalli et al., 2019; Kunz et al., 2025; Littlejohn et al., 2025; Metzger et al., 2023; Li et al., 2026), reconstructing intelligible speech from non-invasive recordings such as electroencephalography (EEG) and magnetoencephalography (MEG) remains more challenging. Their limited spatial resolution, low signal-to-noise ratio, and measurement noise obscure auditory-evoked activity, whereas natural speech contains highly structured linguistic information and fine-grained spectrotemporal dynamics. Reliably mapping these sparse, noisy neural observations to continuous, intelligible speech therefore remains fundamentally difficult.

Prior work on non-invasive speech decoding from auditory-evoked neural activity spans speech envelope reconstruction (Lalor & Foxe, 2010; Destoky et al., 2019), auditory-attention decoding (O'sullivan et al., 2015; Xu et al., 2022; Accou et al., 2023), voice activity detection (Sharon & Murthy, 2020; Özdogan et al., 2025; Landau et al., 2025; de Zuazo et al., 2025; Xu et al., 2026), acoustic-feature prediction or retrieval (e.g., mel-spectrograms) (Défossez et al., 2023; Li et al., 2024), phoneme classification (de Zuazo et al., 2025; Xu et al., 2026), and small closed-vocabulary word classification (d'Ascoli et al., 2025; McMurray et al., 2022). These studies provide important foundations for EEG/MEG speech decoding but largely target intermediate speech attributes rather than continuous speech waveforms. More recently, pioneering studies (Lee et al., 2024; 2025) have

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

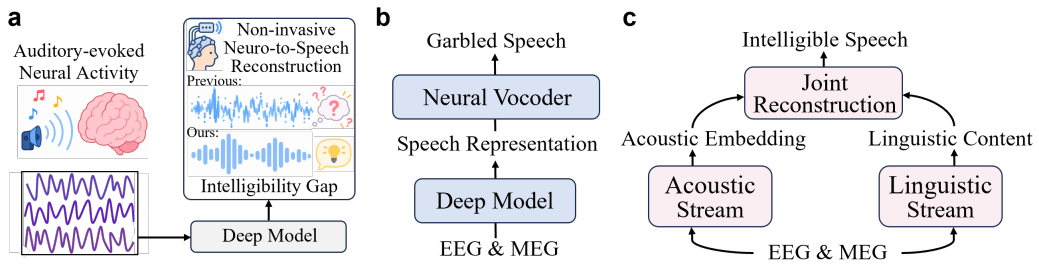


Figure 1: **Intelligibility gap for speech reconstruction from auditory-evoked non-invasive neural signals**. **a**, existing methods are predominantly acoustic-centric reconstruction, resulting in garbled speech waveforms, whereas our goal is to achieve continuous and intelligible speech reconstruction. **b**, previous methods directly map neural recording to entangled speech representation. **c**, our *MindVoice* explicitly recovers linguistic information from non-invasive neural signals and integrates it with complementary acoustic information.

explored direct speech reconstruction. As shown in Figure 1b, these approaches are predominantly acoustic-centric, mapping neural activity directly to deep acoustic or speech representations before synthesizing waveforms with pretrained neural vocoders.

Although acoustic-centric methods can recover some acoustic characteristics of perceived speech, they do not explicitly ensure accurate linguistic recovery. This limitation is particularly pronounced for non-invasive recordings: EEG and MEG provide noisy, spatially blurred, incomplete observations, whereas continuous acoustic representations entangle phonetic structure, lexical content, speaker characteristics, and prosody. Mapping these impoverished observations to such high-dimensional representations is therefore severely underdetermined. The predicted errors can affect speech attributes unevenly, preserving coarse spectral structure and speech-like acoustics despite substantial phonetic or lexical corruption. Reconstructions can thus resemble the original acoustically yet remain difficult to understand. This creates a critical ***intelligibility gap***: reconstructing speech-like acoustics does not necessarily imply recovering what was perceived.

We argue that explicit recovery of linguistic information, rather than its implicit emergence from acoustic reconstruction, can help bridge this intelligibility gap. To this end, we propose the ***MindVoice***, a neuro-to-speech framework centered on linguistic recovery from non-invasive neural signals. As illustrated in Figure 1c, its linguistic reconstruction stream directly estimates linguistic content—what was perceived, while a complementary acoustic reconstruction stream captures fine-grained acoustic characteristics of how it was spoken. This decoupled design reduces the burden on noisy neural recordings to jointly preserve linguistic content, speaker characteristics, and spectrotemporal details in a single deep representation. Given incomplete neural evidence, *MindVoice* further leverages prior knowledge from pretrained generative speech models to recover weak, ambiguous, or missing information. Finally, Text-to-speech generation and in-context voice cloning then integrate the recovered linguistic context and acoustic information to synthesize the natural, intelligible speech.

Extensive evaluations on widely used EEG (Brennan & Hale, 2019) and MEG (Gwilliams et al., 2023) datasets show that *MindVoice* substantially outperforms prior non-invasive neuro-to-speech methods across linguistic and acoustic metrics. More importantly, linguistic recovery substantially improves intelligibility, supporting our hypothesis that acoustic reconstruction alone cannot resolve this bottleneck. Meanwhile, the acoustic stream contributes complementary information to preserve speaker characteristics and fine-grained acoustic details. Results on both datasets support applicability to the evaluated EEG and MEG settings, with qualitative analyses and visualizations corroborating the quantitative gains. Together, these findings suggest that explicit linguistic recovery offers an effective route toward closing the intelligibility gap in non-invasive neural speech reconstruction.

Our main contributions are summarized as follows:

- We propose *MindVoice*, a non-invasive neural speech reconstruction framework centered on explicit linguistic recovery to address the speech **intelligibility gap**.
- We identify an empirical **trade-off** between spectral fidelity and linguistic intelligibility in reconstruction. Our *MindVoice* achieves a **favorable balance** between these objectives.
- Our *MindVoice* achieves **state-of-the-art speech reconstruction intelligibility** compared with existing baselines. We further investigate its interpretability and reconstruction preferences.

2 Related Works

Neural Speech Decoding. Neural speech decoding spans invasive recordings, such as ECoG and MEAs, and non-invasive recordings, including EEG, MEG, and fMRI, across perceived, produced, attempted, and imagined speech. Invasive studies have characterized speech-related neural representations (Mesgarani et al., 2014) and demonstrated speech decoding and synthesis, including real-time systems (Wairagkar et al., 2025; Anumanchipalli et al., 2019; Littlejohn et al., 2025; Metzger et al., 2023; Willett et al., 2023; Kunz et al., 2025; Li et al., 2026). Non-invasive studies encompass acoustic feature decoding (Lalor & Foxe, 2010; Destoky et al., 2019; Li et al., 2024), speech and semantic unit decoding (Lee et al., 2023b; Défossez et al., 2023; de Zuazo et al., 2025; Xu et al., 2026; d'Ascoli et al., 2025; McMurray et al., 2022; Abdulghani et al., 2023; Sato et al., 2024), and continuous speech or language reconstruction (Lee et al., 2024; 2025; Yang et al., 2024; Li et al., 2025; Rastogi et al., 2025; Liu et al., 2025; Yang et al., 2026). Despite this progress, recording modalities and task paradigms impose different constraints on speech reconstruction. Invasive recordings provide high-fidelity neural signals but require surgical implantation, while the limited temporal resolution of fMRI can hinder the tracking of rapid speech dynamics. Imagined speech lacks an observable acoustic reference for precise alignment, whereas overt speech may introduce articulatory and movement-related confounds. These considerations motivate our focus on perceived speech reconstruction from EEG and MEG during listening, which combines fine temporal resolution with an external speech reference for alignment and evaluation.

Pretrained Speech Models. Speech LLMs have gained increasing attention in research communities. For automatic speech recognition (Radford et al., 2023; Barrault et al., 2023a;b; Gao et al., 2023; Shi et al., 2026; Amodei et al., 2016; Gulati et al., 2020; Baevski et al., 2020), recent studies leverage large-scale pre-training to improve transcription robustness across diverse acoustic conditions. For text-to-speech synthesis (Du et al., 2024; Chen et al., 2025; Liao et al., 2024; Hu et al., 2026; Arık et al., 2017; Ren et al., 2019; 2020; Kim et al., 2020; 2021), LLM-based speech generation has advanced expressive audio modeling by predicting discrete or continuous speech representations. To achieve personalized speech, many works (Du et al., 2024; Chen et al., 2025; Liao et al., 2024; Hu et al., 2026) further explore speech cloning techniques that adapt synthetic speech to target speakers using limited reference audio, either by conditioning generation on speaker embeddings or by modeling codec-based speech tokens. These methods improve speaker similarity, prosody similarity, and expressive controllability. In our work, we leverage the priors of speech LLMs to supplement the incomplete information in neural recordings.

3 Methods

3.1 Overview

Speech carries heterogeneous information, including speaker timbre, prosody, and linguistic content. Directly mapping noisy neural recording $\mathcal{X}_i$ to speech representations that jointly encode these factors is challenging and can lead to poor intelligibility. Our *MindVoice* therefore introduces an explicit linguistic stream to enhance intelligibility. Specifically, *MindVoice* adopts a dual-reconstruction framework that predicts different aspects of speech-related information, with each stream augmented by pretrained priors. The **acoustic reconstruction stream** independently predicts discrete speech-codec embeddings $A_{\mathcal{Y}_i}$ to obtain specific pitch and timbre cues. The **linguistic reconstruction stream** reconstructs discrete representations $S_{\mathcal{Y}_i}$, decodes them into mel-spectrograms $\mathcal{M}_i$, and uses an adapted ASR model to recover linguistic content $\mathcal{T}_i$. Finally, the **joint speech reconstruction** module integrates the acoustic embeddings $A_{\mathcal{Y}_i}$ and text $\mathcal{T}_i$ through a frozen text-to-speech model. This design provides an explicit linguistic input alongside the complementary acoustic prediction.

3.2 Acoustic Reconstruction Stream

We begin with direct acoustic reconstruction from neural recordings. Given paired EEG/MEG and speech waveforms $\{(\mathcal{X}_i, \mathcal{Y}_i)\}_{i=1}^{N}$, this stream predicts acoustic representations $A_{\mathcal{Y}_i}$ from $\mathcal{X}_i$.

Neural Signal Embedder. We treat EEG/MEG signals with channel and temporal dimensions as single-channel images, i.e., $\mathcal{X}_i \in \mathbb{R}^{1\times \mathrm{T}\times \mathrm{C}}$. Based on this, cascaded temporal CNNs progressively compress the input $\mathcal{X}_i$ along the temporal dimension, while spatial convolutions integrate information across recording channels. The resulting features are rearranged into sequence latents in $\mathbb{R}^{\mathrm{T}'\times d}$ space,

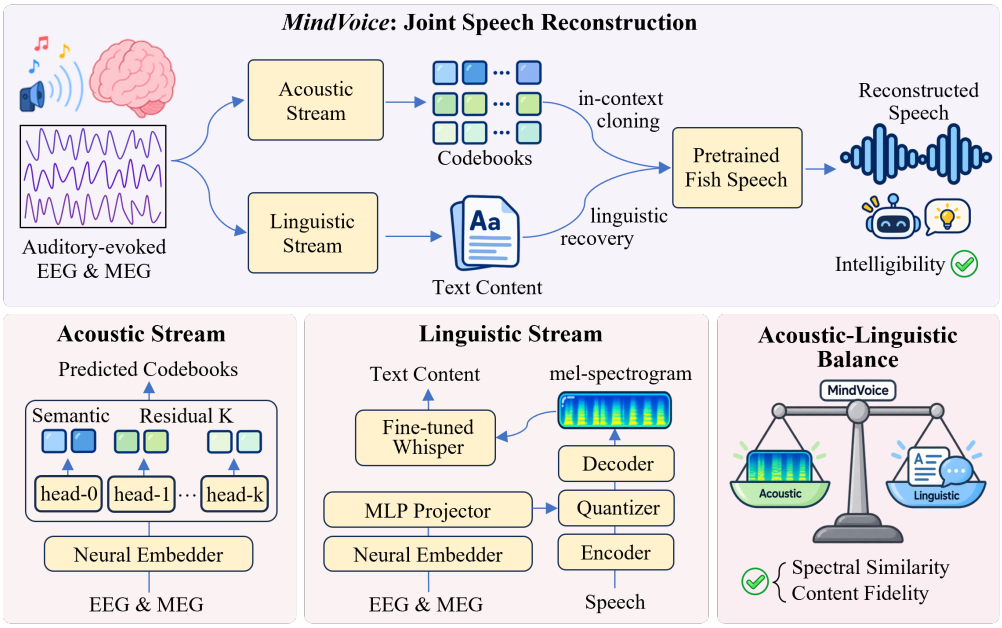


Figure 2: **Overview of *MindVoice* framework**. Acoustic and linguistic reconstruction streams recover complementary information. TTS model integrates them in a balanced manner to final results.

where each d-dimensional neural token encodes multichannel activity within a local temporal window. To account for inter-subject variability, we associate each subject with a learnable embedding. This embedding is added to every sequence latent of the corresponding subject, allowing the embedder to condition on subject identity. Then, we add learnable positional embeddings and apply a Transformer encoder to model inter-token dependencies, producing the final neural embeddings $E_{\mathcal{X}_i}$.

Discrete Acoustic Embeddings. We adopt the pretrained Fish-Speech codec (Liao et al., 2024) to provide structured acoustic embeddings. Given a speech waveform $\mathcal{Y}_i$, the codec encoder $E(\cdot)$ extracts continuous acoustic features $A_{\mathcal{Y}_i}$, comprising T' vectors aligned with T' temporal steps. The codec quantizer $Q(\cdot)$ maps these features into a discrete representation using a semantic-dominant codebook followed by K residual codebooks. The primary semantic-dominant codebook provides an initial quantized approximation, while subsequent codebooks progressively encode the remaining quantization error to preserve additional acoustic detail. The quantized embedding is expressed as:

$$A_{\mathcal{Y}_i} = E(\mathcal{Y}_i) = \{a_t\}_{t=1}^{T'}, \quad A^q_{\mathcal{Y}_i} = Q\left(A_{\mathcal{Y}_i}\right) = \{a^q_t\}_{t=1}^{T'}, \quad a^q_t = \sum_{k=0}^{K} a^q_{k,t}, \tag{1}$$

where $a^q_{k,t}$ denotes the quantized code from the k-th codebook, and $k = 0$ corresponds to the semantic-dominant codebook. For more details on Fish-Speech, please refer to Appendix A.1.

Codebook Classifiers. Given the neural embeddings $E_{\mathcal{X}_i} = \{e_t\}_{t=1}^{T'}$, we achieve acoustic embedding prediction using codebook classification heads $\{f_k\}_{k=0}^{K}$. We first predict the semantic-dominant code $\tilde{a}^q_{0,t}$ and then condition on it to predict the remaining residual codes $\tilde{a}^q_{k,t}$ in parallel. We freeze the pretrained codec and jointly optimize the neural signal embedder and codebook classifiers through the code-level smoothing cross-entropy (CE) supervision as follows.

$$\mathcal{L}_{\text{acoustic}} = \frac{1}{|\Omega|} \sum_{(i,t)\in\Omega} \left[\text{CE}\left(f_0(e_t), a^q_{0,t}\right) + \frac{1}{K} \sum_{k=1}^{K} \text{CE}\left(f_k(e_t, \tilde{a}^q_{0,t}), a^q_{k,t}\right) \right]. \tag{2}$$

Here, Ω denotes the set of valid sample-time pairs, f_0 and f_k denote the semantic-dominant and residual code classifiers, respectively. For simplicity, we reuse the code notation $a^q_{k,t}$ in CE loss.

3.3 Linguistic Reconstruction Stream

Although the acoustic reconstruction stream captures speech structure, it does not ensure the correct linguistic content. We therefore introduce the linguistic reconstruction stream to provide explicit

linguistic guidance. Given the transcript text $\mathcal{T}_i$ of $\mathcal{Y}_i$, this stream learns to reconstruct $\tilde{\mathcal{T}}_i$ from $\mathcal{X}_i$. It consists of a vector-quantized speech autoencoder, a neuro-to–language alignment module, and an ASR model, which are trained or fine-tuned in three separate stages.

Speech Vector-Quantized Autoencoder. We build a speech vector-quantized autoencoder to learn compact speech embeddings. Given a speech waveform $\mathcal{Y}_i$, we first compute its mel-spectrogram $\mathcal{M}_i$. An encoder composed of convolutional layers, downsampling layers, and self-attention blocks compresses M_i into continuous latent embeddings $S_{\mathcal{Y}_i}$. A learnable vocabulary maps each latent vector to its nearest entry under ℓ_2 distance after normalization, producing the quantized embeddings $S^q_{\mathcal{Y}_i}$. A decoder, roughly symmetric to the encoder, then reconstructs the spectrogram as $\tilde{\mathcal{M}}_i$. The autoencoder is trained with spectrogram reconstruction and vector-quantization objectives:

$$\mathcal{L}_{\text{vqae}} = \frac{1}{N}\sum_{i=1}^{N}\Big[\big\|\mathcal{M}_i - \tilde{\mathcal{M}}_i\big\|_2^2 + \alpha_1\big\|\text{sg}[S_{\mathcal{Y}_i}] - S^q_{\mathcal{Y}_i}\big\|_2^2 + \alpha_2\big\|S_{\mathcal{Y}_i} - \text{sg}[S^q_{\mathcal{Y}_i}]\big\|_2^2\Big]. \tag{3}$$

Here, the first term encourages the reconstructed spectrogram $\tilde{\mathcal{M}}_i$ to match the ground-truth $\mathcal{M}_i$. The second term updates the vocabulary vectors by moving them toward the encoder latents, while the last term encourages the encoder outputs to stay close to the selected vocabulary entries, preventing unstable switching across entries during training. $\text{sg}[\cdot]$ denotes the stop-gradient operator, which is used with the straight-through estimator trick for training the quantizer.

Neuro-to-Language Alignment. We introduce an MLP projector to map the neural embeddings $E_{\mathcal{X}_i}$ into the pre-quantization speech embedding space, yielding $\tilde{S}_{\mathcal{Y}_i}$. The predicted continuous embeddings are quantized and subsequently decoded into the reconstructed mel-spectrogram $\tilde{\mathcal{M}}_i$. We jointly train another neural signal embedder and the projector while keeping the speech vector-quantized autoencoder fixed. The alignment objective is defined as:

$$\begin{aligned}\mathcal{L}_{\text{contr}} &= -\frac{1}{2N}\sum_{i=1}^{N}\left[\log\frac{\exp(\tilde{S}_{\mathcal{Y}_i}\cdot S^\top_{\mathcal{Y}_i}/\tau)}{\sum_{j=1}^{N}\exp(\tilde{S}_{\mathcal{Y}_i}\cdot S^\top_{\mathcal{Y}_j}/\tau)} + \log\frac{\exp(\tilde{S}_{\mathcal{Y}_i}\cdot S^\top_{\mathcal{Y}_i}/\tau)}{\sum_{j=1}^{N}\exp(\tilde{S}_{\mathcal{Y}_j}\cdot S^\top_{\mathcal{Y}_i}/\tau)}\right], \\ \mathcal{L}_{\text{align}} &= \frac{1}{N}\sum_{i=1}^{N}\left[\left\|\tilde{\mathcal{M}}_i - \mathcal{M}_i\right\|_2^2 + \beta_1\left\|\tilde{S}_{\mathcal{Y}_i} - S_{\mathcal{Y}_i}\right\|_2^2\right] + \beta_2\mathcal{L}_{\text{contr}}.\end{aligned} \tag{4}$$

After training, the reconstructed mel-spectrogram $\tilde{\mathcal{M}}_i$ is fed into the pretrained Whisper model (Radford et al., 2023) for text transcription. Its autoregressive language prior helps recover linguistic content from incomplete neural predictions. To further encourage text generation, we optimize next-token prediction against the ground-truth transcript $\mathcal{T}_i$, while retaining the alignment objective:

$$\mathcal{L}_{\text{text}} = -\sum_{i=1}^{N}\sum_{j=1}^{L_i}\frac{1}{NL_i}\log\mathcal{P}\left(\mathcal{T}_{i,j} \mid \mathcal{T}_{i,<j}, \tilde{\mathcal{M}}_i\right) + \gamma\mathcal{L}_{\text{align}}. \tag{5}$$

Here, $\mathcal{P}(\cdot)$ is the probability that Whisper assigns to the correct next token given the input and previous tokens, while $\mathcal{T}_i = [\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_{L_i}]$ denotes the i-th text with L_i tokens in total. During this stage, we update the projector and Whisper encoder while keeping the neural signal embedder, speech autoencoder, and Whisper decoder frozen.

3.4 Joint Speech Reconstruction

Given the acoustic predictions $\tilde{A}^q_{\mathcal{Y}_i}$ and the reconstructed transcript $\tilde{\mathcal{T}}_i$, we integrate both streams through the pretrained Fish-Speech text-to-speech (TTS) model. The acoustic predictions provide cues about how the speech is realized while the transcripts supply explicit guidance on linguistic content. Following the in-context voice cloning formulation, we use the predicted acoustic-stream codec sequence $\tilde{A}^q_{\mathcal{Y}_i}$ as the in-context speech prompt $\mathcal{C}_i$ and pair it with the reconstructed transcript for joint reconstruction. To reduce the influence of unreliable acoustic predictions, we further construct the prompt $\mathcal{C}_i$ by retaining some selected temporal segments. We also consider linguistic-stream prompt refinement, which combines the code sequences generated from an additional text-to-speech process of the reconstructed transcript with selected predictions of acoustic-stream residual codebooks. Conditioned on $\tilde{\mathcal{T}}_i$ and $\mathcal{C}_i$, the TTS model autoregressively generates a new codec

sequence $Z_i = \{z_{i,t}\}_{t=1}^{T''}$. The joint reconstruction is formulated as:

$$q\left(Z_i \mid \tilde{\mathcal{T}}_i, \mathcal{C}_i\right) = \prod_{t=1}^{T''} \mathcal{Q}\left(z_{i,t} \mid z_{i,<t}, \tilde{\mathcal{T}}_i, \mathcal{C}_i\right), \quad \tilde{Z}_i \sim q\left(Z_i \mid \tilde{\mathcal{T}}_i, \mathcal{C}_i\right), \quad \tilde{\mathcal{Y}}_i = D(\tilde{Z}_i), \tag{6}$$

where L_i is the generated sequence length, $z_{i,t}$ contains the codec tokens at time t, and $D(\cdot)$ denotes the Fish-Speech codec decoder. Through this process, the in-context speech prompt provides brain-derived acoustic characteristics of the perceived speech, while the transcript provides explicit guidance for speech generation. The pretrained TTS model further incorporates prosodic and temporal priors to integrate these complementary sources of information. Both reconstruction streams, the TTS model, and the codec remain frozen during the joint speech reconstruction process.

4 Experiment Settings

Datasets and Preprocessing. We conduct experiments on the Brennan EEG (Brennan & Hale, 2019) and Gwilliams MEG (Gwilliams et al., 2023) datasets, both collected during continuous English narrative listening. Neural recordings undergo notch and band-pass filtering, downsampling, sentence-level segmentation, and temporal alignment with the corresponding speech. Each segment is further processed with robust scaling and normalization. The resulting neural-speech pairs are split into training, validation, and test sets at an 8:1:1 ratio. Further details are provided in Appendix B.1.

Baselines. We compare against open-vocabulary speech reconstruction methods evaluated on EEG/MEG, excluding fMRI-based methods (Liu et al., 2025) due to substantial differences in spatiotemporal resolution, methods using invasive recordings (Wairagkar et al., 2025), and closed-vocabulary retrieval approaches (Défossez et al., 2023). We first construct a vanilla baseline that regresses mel-spectrograms from neural recordings and reconstructs waveforms using BigVGAN-v2 (Lee et al., 2023a). We also compare with FESDE (Lee et al., 2024), which maps autoencoded EEG embeddings to VITS-style speech latents (Kim et al., 2021) via a normalizing-flow connector and reconstructs speech with a self-trained HiFi-GAN decoder (Kong et al., 2020). Finally, we include the acoustic-only variant of *MindVoice*, without linguistic recovery, as an additional baseline. Further details are provided in Appendix B.2.

Evaluation Metrics. We evaluate reconstructed speech along four complementary dimensions. For *spectral fidelity*, we report mel-spectrogram MSE (mel-MSE) and Pearson correlation (mel-r). For *linguistic fidelity*, reconstructed speech is transcribed using Qwen3-ASR (Shi et al., 2026) and evaluated against ground-truth transcripts using WER and baseline-rescaled BERTScore-F1 (Zhang et al., 2020). For *representation similarity*, we compute cosine similarity between reconstructed and reference speech embeddings extracted by HuBERT (Hsu et al., 2021) and WavLM (Chen et al., 2022). For *articulatory fidelity*, we report SPARC-r (Cho et al., 2024), the mean Pearson correlation across 12 inferred articulatory trajectories after dynamic time warping alignment. These metrics form our main evaluation panel; additional metrics and results are provided in Appendices B.3 and C.1.

Implementation Details. We have provided detailed implementation settings and training configs in Appendix B.4. All results are reported across five random seeds unless stated otherwise.

5 Results

5.1 Quantitative Experimental Results

As shown in Table 1, ***MindVoice* substantially outperforms existing baselines in speech intelligibility, learned speech representations, and articulatory similarity across both EEG and MEG**. Specifically, compared with the strongest existing baseline, *MindVoice* achieves markedly more faithful recovery of linguistic content. WER is reduced from 91.0% to 38.1% on EEG and from 98.7% to 44.3% on MEG, while normalized ASR-BERTScore-F1 increases by 0.592 (CI [0.565, 0.618]) and 0.277 (CI [0.258, 0.296]), respectively. Consistent improvements are also observed in pretrained speech representation spaces: HuBERT similarity increases by 0.257 (CI [0.253, 0.260]) on EEG and 0.397 (CI [0.395, 0.399]) on MEG, while WavLM similarity increases by 0.799 (CI [0.793, 0.804]) and 0.628 (CI [0.627, 0.630]), respectively. Moreover, SPARC-r improves by 0.320 (CI [0.306, 0.334]) on EEG and 0.184 (CI [0.176, 0.192]) on MEG, indicating greater similarity to the reference speech in articulatory representations. All CIs are 95% and computed across five seeds.

Table 1: **Quantitative Comparison between MindVoice and baseline methods**. MindVoice-A denotes the acoustic-stream-only variant. We report the mean over five training seeds.

Data	Method	Spectral		Intelligibility		Representation		Articulation
		mel-MSE ↓	mel-r ↑	WER ↓	ASR-BERT ↑	HuBERT ↑	WavLM ↑	SPARC-r ↑
EEG	Vanilla	0.254	0.470	91.0%	0.150	0.404	0.065	0.429
	FESDE	0.178	0.550	98.2%	0.166	0.612	0.007	0.296
	MindVoice-A	**0.168**	**0.555**	102.3%	0.064	0.581	0.199	0.435
	MindVoice	0.287	0.382	**38.1%**	**0.758**	**0.868**	**0.863**	**0.749**
MEG	Vanilla	0.177	0.685	117.8%	0.058	0.446	0.215	0.478
	FESDE	0.164	0.490	98.7%	0.206	0.474	0.133	0.267
	MindVoice-A	**0.110**	**0.725**	98.8%	0.199	0.755	0.540	0.636
	MindVoice	0.329	0.342	**44.3%**	**0.483**	**0.871**	**0.843**	**0.661**

The comparison with *MindVoice-A*, our acoustic-stream-only variant, further reveals an important trade-off between spectral fidelity and speech intelligibility. Although *MindVoice-A* achieves the best spectral correspondence on both datasets, its reconstructed speech remains poorly intelligible. Incorporating the reconstructed linguistic content and pretrained TTS priors substantially improves intelligibility and high-level speech similarity in the complete *MindVoice*, although its frame-level spectral metrics become lower. This result highlights that **spectral similarity alone does not necessarily translate into intelligible reconstruction; the complete *MindVoice* prioritizes linguistically and perceptually meaningful speech characteristics over exact frame-wise spectral matching**.

5.2 Visualization and Qualitative Results

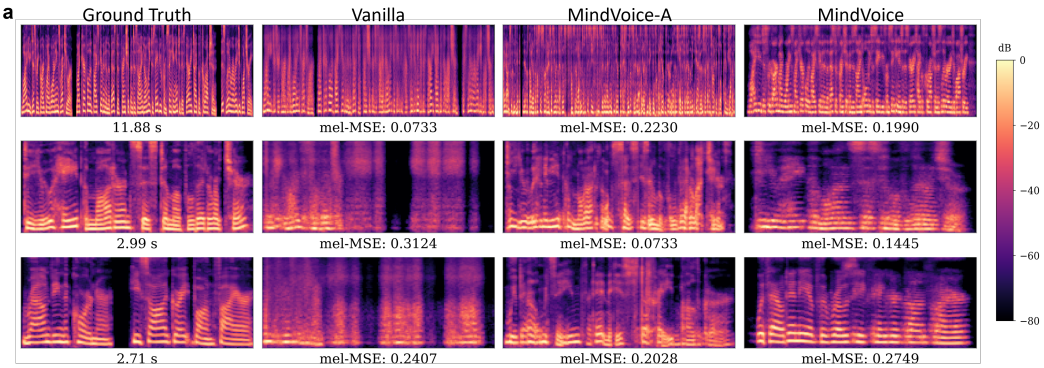
Figure 3 further illustrates the trade-off between spectral fidelity and linguistic intelligibility observed in Table 1. Acoustic-centric reconstruction, including *Vanilla* and *MindVoice-A*, achieves closer correspondence to the reference spectrogram in terms of spectral energy distribution. However, pointwise spectral correspondence alone does not appear to fully explain the differences in intelligibility, which may also depend on how local spectrotemporal patterns are organized into recognizable phonetic and lexical structures. This suggests that **spectral fidelity and linguistic intelligibility reflect complementary aspects of reconstructed speech.**

The transcription results further support this interpretation. By explicitly recovering linguistic content and using it to guide speech generation, *MindVoice* more consistently preserves the overall sentence structure, key content words, and local grammatical relations. Although the resynthesized speech may deviate more from the reference in frame-wise spectral energy, its local acoustic patterns form more coherent linguistic sequences. **Explicit linguistic recovery therefore improves the preservation of what was perceived, even when exact spectral correspondence is reduced.**

5.3 Ablation Studies

We further investigate where the gains of *MindVoice* originate. As shown in Table 2, the acoustic-only *MindVoice-A* achieves the strongest spectral fidelity but remains poorly intelligible, whereas the linguistic-only *MindVoice-L* achieves the best intelligibility at the expense of spectral correspondence. These opposing trends reveal a clear trade-off: improving linguistic recovery does not simultaneously preserve the strongest frame-wise acoustic correspondence, and vice versa. The complete *MindVoice* introduces recovered linguistic content into acoustic reconstruction and reaches a favorable balance between the two, substantially improving intelligibility over *MindVoice-A* while recovering better spectral fidelity than *MindVoice-L*. Notably, this balance yields the strongest overall performance in learned speech-representation similarity, with the highest HuBERT score and the tied-best WavLM score. These results suggest that **acoustic fidelity and linguistic intelligibility are difficult to optimize simultaneously, while *MindVoice* provides an effective trade-off that best preserves the overall deep speech representation.**

The noise controls further disentangle information decoded from neural signals from that supplied by pretrained priors. Replacing the linguistic-stream neural input with Gaussian noise increases WER from 43.5% to 99.6% and reduces ASR-BERT from 0.488 to 0.203, even though the pretrained ASR and TTS priors remain available. This indicates that linguistic recovery critically depends on informative neural input rather than generative priors alone. Replacing the acoustic-stream input



b

(1)	GT:	Do you hate the modern system of literature?
	Vanilla:	The brain and mind system of the creature. (WER=87.5%, ASR-BERT=0.277)
	Ours-A:	Do you have any other more concise and profound words of care? (WER=112.5%, ASR-BERT=0.072)
	Ours:	Who hate the modern system of literature? (WER=25.0%, ASR-BERT=0.697)
(2)	GT:	Yes, I suppose one could see it thus.
	Vanilla:	Yeah. (WER=100.0%, ASR-BERT=0.182)
	Ours-A:	Yes, I searched one cookie policy. (WER=62.5%, ASR-BERT=0.061)
	Ours:	Yes, and pose one could see it thus. (WER=25.0%, ASR-BERT=0.518)
(3)	GT:	The author leaned back with slight sigh if both exhausted satisfied by his effort.
	Vanilla:	We are believed to have lost sight of his voice exhaustion that was caused by his efforts. (WER=100.0%, ASR-BERT=-0.017)
	Ours-A:	He ought to lean back with a slight sigh, if both exhausted and satisfied by his efforts. (WER=50.0%, ASR-BERT=0.564)
	Ours:	Doctor leaned back with slight sigh, as both exhausted satisfied by his effort. (WER=21.4%, ASR-BERT=0.780)

Figure 3: **Mel-spectrogram and transcription results**. **a**, comparison of mel-spectrograms obtained by different methods with the ground-truth. **b**, selected transcription results. More results can be found in Appendix C.

Table 2: **Ablation results on the MEG dataset**. MindVoice-L denotes the variant using only the linguistic stream with text-to-speech process, while Noise-A and Noise-L replace the MEG inputs to the acoustic and linguistic streams with Gaussian noise, respectively. Results are reported from a single run.

Method	Spectral		Intelligibility		Representation		Articulation
	Mel-MSE↓	Mel-r ↑	WER↓	ASR-BERT↑	HuBERT↑	WavLM↑	SPARC-r ↑
MindVoice	0.329	0.341	43.5%	0.488	**0.871**	**0.844**	0.664
MindVoice-A	**0.109**	**0.727**	97.6%	0.207	0.759	0.550	0.640
MindVoice-L	0.379	0.306	**39.3%**	**0.516**	0.862	0.844	**0.667**
Noise-A	0.345	0.340	51.1%	0.446	0.824	0.775	0.607
Noise-L	0.518	0.302	99.6%	0.203	0.551	0.438	0.323

with noise likewise reduces HuBERT, WavLM, and SPARC-r, showing that acoustic reconstruction also relies on signal-derived information. Together, these controls demonstrate that **neural signals provide essential information for both linguistic and acoustic reconstruction, rather than the recovered speech being driven primarily by pretrained speech priors.**

5.4 Model Interpretability and Preference

We further examine how the two reconstruction streams utilize neural signals. For Figure 4a, we perform a Grad-CAM-style (Selvaraju et al., 2017) attribution analysis for each test sample to estimate the contribution of individual MEG channels to the linguistic and acoustic reconstruction streams, and average the resulting channel-wise attribution scores over the test set. The linguistic stream exhibits high-attribution responses across a broader set of the 208 MEG channels, whereas the acoustic stream shows relatively more concentrated channel preferences. This suggests that **linguistic reconstruction draws on a more distributed pattern of neural information, while acoustic reconstruction relies more strongly on a smaller subset of channels**. Such a distinction is broadly consistent with neurophysiological evidence that higher-level linguistic representations are distributed across widespread cortical systems, beyond the auditory pathways supporting earlier acoustic and phonetic processing (Huth et al., 2016; Keshishian et al., 2023).

We next characterize the reconstruction preference of our MindVoice. As shown in Figure 4b, reconstructed sentence length closely tracks the ground-truth length, with a fitted relation of $y =$

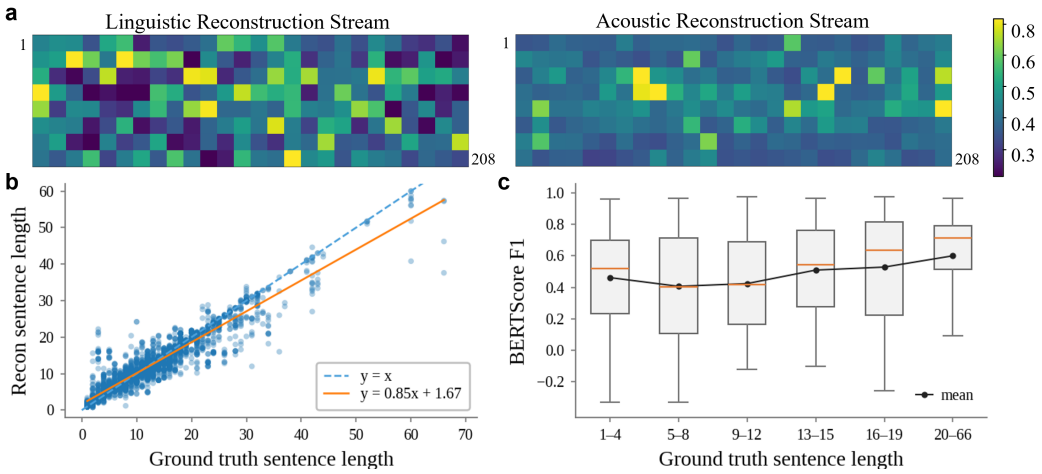


Figure 4: **Model interpretability and linguistic reconstruction characteristics**. **a**, Grad-CAM-style channel attribution maps for the linguistic and acoustic reconstruction streams, averaged over the test set; each pixel corresponds to one of 208 MEG channels. **b**, regression between ground-truth and reconstructed sentence lengths (word counts), with the identity line shown for reference. **c**, distributions of baseline-rescaled BERTScore-F1 across ground-truth word-counts groups; black markers and line denote the group means.

$0.85x + 1.67$. The slope close to the identity line indicates that the model largely preserves sentence-length structure, with only a mild compression tendency for longer utterances. Figure 4c further shows that linguistic reconstruction quality remains robust across sentence lengths and tends to increase for longer utterances. **Together, these results suggest that the linguistic stream recovers structured sentence-level information from MEG signals, rather than merely producing plausible linguistic outputs through pretrained priors**.

6 Discussion and Limitations

Our results highlight a clear distinction between acoustic fidelity and linguistic intelligibility in non-invasive speech reconstruction. Acoustic-centric reconstruction can preserve spectral characteristics while remaining poorly intelligible, whereas explicit linguistic recovery substantially improves recovery of what was perceived. The complete *MindVoice* balances these complementary objectives and achieves the strongest similarity in learned speech-representation spaces. These findings suggest that intelligible reconstruction benefits from explicitly recovering linguistic information rather than relying on it to emerge implicitly from acoustic reconstruction. Ablation and interpretability analyses further show that both linguistic and acoustic reconstruction depend on informative neural signals. Noise controls degrade the corresponding reconstruction quality, while distinct channel-attribution patterns indicate different preferences over the available neural information. The linguistic stream additionally preserves sentence-level structure, supporting the view that non-invasive neural signals contain recoverable speech information at multiple levels. Two limitations remain. First, linguistic intelligibility and spectral fidelity exhibit a clear trade-off, and jointly improving both remains challenging. Second, the linguistic and acoustic streams are not fully disentangled: each still retains information associated with the other. Thus, they should be viewed as emphasizing different aspects of speech rather than representing strictly independent factors.

7 Conclusion

We present *MindVoice*, a framework for intelligible speech reconstruction from non-invasive neural signals through explicit linguistic recovery. By introducing recovered linguistic content into acoustic reconstruction, *MindVoice* substantially improves speech intelligibility while retaining complementary acoustic information. Our results demonstrate that explicit linguistic recovery is a key ingredient for closing the intelligibility gap in non-invasive neural speech reconstruction and provide a promising direction toward practical speech brain–computer interfaces.

REFERENCES

Mokhles M Abdulghani, Wilbur L Walters, and Khalid H Abed. Imagined speech classification using eeg and deep learning. *Bioengineering*, 10(6):649, 2023.

Bernd Accou, Jonas Vanthornhout, Hugo Van hamme, and Tom Francart. Decoding of the speech envelope from eeg using the vlaai deep neural network. *Scientific Reports*, 13(1):812, 2023.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pp. 173–182. PMLR, 2016.

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

Sercan Ö Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, et al. Deep voice: Real-time neural text-to-speech. In *International conference on machine learning*, pp. 195–204. PMLR, 2017.

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

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. Seamlessm4t: Massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*, 2023a.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023b.

Jonathan R Brennan and John T Hale. Hierarchical structure guides rapid linguistic predictions during naturalistic listening. *PloS one*, 14(1):e0207741, 2019.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518, 2022.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 6255–6271, 2025.

Cheol Jun Cho, Peter Wu, Tejas S Prabhune, Dhruv Agarwal, and Gopala K Anumanchipalli. Coding speech through vocal tract kinematics. *IEEE Journal of Selected Topics in Signal Processing*, 18 (8):1427–1440, 2024.

Matthew H Davis and Ingrid S Johnsrude. Hierarchical processing in spoken language comprehension. *Journal of Neuroscience*, 23(8):3423–3431, 2003.

Xabier de Zuazo, Ibon Saratxaga, and Eva Navas. Megconformer: Conformer-based meg decoder for robust speech and phoneme classification. *arXiv preprint arXiv:2512.01443*, 2025.

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

Florian Destoky, Morgane Philippe, Julie Bertels, Marie Verhasselt, Nicolas Coquelet, Marc Vander Ghinst, Vincent Wens, Xavier De Tiège, and Mathieu Bourguignon. Comparing the potential of meg and eeg to uncover brain tracking of speech temporal envelope. *Neuroimage*, 184:201–213, 2019.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

Stéphane d'Ascoli, Corentin Bel, Jérémy Rapin, Hubert Banville, Yohann Benchetrit, Christophe Pallier, and Jean-Rémi King. Towards decoding individual words from non-invasive brain recordings. *Nature Communications*, 16(1):10521, 2025.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, and Shiliang Zhang. Funasr: A fundamental end-to-end speech recognition toolkit. In *Proceedings of Interspeech*, pp. 1593–1597, 2023.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *Proc. Interspeech 2020*, pp. 5036–5040, 2020.

Laura Gwilliams, Graham Flick, Alec Marantz, Liina Pylkkänen, David Poeppel, and Jean-Rémi King. Introducing meg-masc a high-quality magneto-encephalography dataset for evaluating natural speech processing. *Scientific data*, 10(1):862, 2023.

Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature reviews neuroscience*, 8(5):393–402, 2007.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Hangrui Hu, Xinfa Zhu, Ting He, Dake Guo, Bin Zhang, Xiong Wang, Zhifang Guo, Ziyue Jiang, Hongkun Hao, Zishan Guo, et al. Qwen3-tts technical report. *arXiv preprint arXiv:2601.15621*, 2026.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 (7600):453–458, 2016.

Menoua Keshishian, Serdar Akkol, Jose Herrero, Stephan Bickel, Ashesh D Mehta, and Nima Mesgarani. Joint, distributed and hierarchically organized encoding of linguistic features in the human auditory cortex. *Nature human behaviour*, 7(5):740–753, 2023.

Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33:8067–8077, 2020.

Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International conference on machine learning*, pp. 5530–5540. PMLR, 2021.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033, 2020.

Erin M Kunz, Benyamin Abramovich Krasa, Foram Kamdar, Donald T Avansino, Nick Hahn, Seonghyun Yoon, Akansha Singh, Samuel R Nason-Tomaszewski, Nicholas S Card, Justin J Jude, et al. Inner speech in motor cortex and implications for speech neuroprostheses. *Cell*, 188(17): 4658–4673, 2025.

Edmund C Lalor and John J Foxe. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European journal of neuroscience*, 31(1):189–193, 2010.

Gilad Landau, Miran Özdogan, Gereon Elvers, Francesco Mantegna, Pratik Somaiya, Dulhan Jayalath, Luisa Kurth, Teyun Kwon, Brendan Shillingford, Greg Farquhar, et al. The 2025 pnpl competition: Speech detection and phoneme classification in the libribrain dataset. *arXiv preprint arXiv:2506.10165*, 2025.

Jihwan Lee, Aditya Kommineni, Tiantian Feng, Kleanthis Avramidis, Xuan Shi, Sudarsana Reddy Kadiri, and Shrikanth Narayanan. Toward fully-end-to-end listened speech decoding from eeg signals. In *Interspeech*, pp. 1500–1504, 2024.

Jihwan Lee, Tiantian Feng, Aditya Kommineni, Sudarsana Reddy Kadiri, and Shrikanth Narayanan. Enhancing listened speech decoding from eeg via parallel phoneme sequence prediction. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5. IEEE, 2025.

Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. Bigvgan: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*, 2023a.

Young-Eun Lee, Seo-Hyun Lee, Sang-Ho Kim, and Seong-Whan Lee. Towards voice reconstruction from eeg during imagined speech. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pp. 6030–6038, 2023b.

Hao Li, Yuan Fang, Xueliang Zhang, Fei Chen, and Guanglai Gao. Cross-attention-guided wavenet for eeg-to-mel spectrogram reconstruction. In *Proceedings of Interspeech*, 2024.

Jiawei Li, Chunxu Guo, Chao Zhang, Edward F Chang, and Yuanning Li. High-fidelity neural speech reconstruction through an efficient acoustic-linguistic dual-pathway framework. *Elife*, 14: RP109400, 2026.

Jilong Li, Zhenxi Song, Jiaqi Wang, Meishan Zhang, Honghai Liu, Min Zhang, and Zhiguo Zhang. Brainecho: Semantic brain signal decoding through vector-quantized spectrogram reconstruction for whisper-enhanced text generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2762–2778, 2025.

Shijia Liao, Yuxuan Wang, Tianyu Li, Yifan Cheng, Ruoyi Zhang, Rongzhi Zhou, and Yijin Xing. Fish-speech: Leveraging large language models for advanced multilingual text-to-speech synthesis. *arXiv preprint arXiv:2411.01156*, 2024.

Kaylo T Littlejohn, Cheol Jun Cho, Jessie R Liu, Alexander B Silva, Bohan Yu, Vanessa R Anderson, Cady M Kurtz-Miott, Samantha Brosler, Anshul P Kashyap, Irina P Hallinan, et al. A streaming brain-to-voice neuroprosthesis to restore naturalistic communication. *Nature neuroscience*, 28(4): 902–912, 2025.

Che Liu, Changde Du, Xiaoyu Chen, and Huiguang He. Reverse the auditory processing pathway: Coarse-to-fine audio reconstruction from human brain activity. *Tsinghua Science and Technology*, 2025.

Bob McMurray, McCall E Sarrett, Samantha Chiu, Alexis K Black, Alice Wang, Rebecca Canale, and Richard N Aslin. Decoding the temporal dynamics of spoken word and nonword processing from eeg. *NeuroImage*, 260:119457, 2022.

Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F Chang. Phonetic feature encoding in human superior temporal gyrus. *Science*, 343(6174):1006–1010, 2014.

Sean L Metzger, Kaylo T Littlejohn, Alexander B Silva, David A Moses, Margaret P Seaton, Ran Wang, Maximilian E Dougherty, Jessie R Liu, Peter Wu, Michael A Berger, et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976): 1037–1046, 2023.

James A O'sullivan, Alan J Power, Nima Mesgarani, Siddharth Rajaram, John J Foxe, Barbara G Shinn-Cunningham, Malcolm Slaney, Shihab A Shamma, and Edmund C Lalor. Attentional selection in a cocktail party environment can be decoded from single-trial eeg. *Cerebral cortex*, 25 (7):1697–1706, 2015.

Miran Özdogan, Gilad Landau, Gereon Elvers, Dulhan Jayalath, Pratik Somaiya, Francesco Mantegna, Mark Woolrich, and Oiwi Parker Jones. Libribrain: Over 50 hours of within-subject meg to improve speech decoding methods at scale. *arXiv preprint arXiv:2506.02098*, 2025.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Sparsh Rastogi, Harsh Dadwal, Khushboo Modi, Jatin Bedi, and Jasmeet Singh. Towards sentence level imagined speech generation from eeg signals. In *Proc. Interspeech 2025*, pp. 5558–5562, 2025.

Josef P Rauschecker and Sophie K Scott. Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature neuroscience*, 12(6):718–724, 2009.

Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech: Fast, robust and controllable text to speech. *Advances in neural information processing systems*, 32, 2019.

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *International Conference on Learning Representations*, 2020.

Motoshige Sato, Kenichi Tomeoka, Ilya Horiguchi, Kai Arulkumaran, Ryota Kanai, and Shuntaro Sasai. Scaling law in neural data: Non-invasive speech decoding with 175 hours of eeg data. *arXiv preprint arXiv:2407.07595*, 2024.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Rini A Sharon and Hema A Murthy. The "sound of silence" in eeg—cognitive voice activity detection. In *Proceedings of Interspeech*, pp. 2767–2771, 2020.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, et al. Qwen3-asr technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Maitreyee Wairagkar, Nicholas S Card, Tyler Singer-Clark, Xianda Hou, Carrina Iacobacci, Lee M Miller, Leigh R Hochberg, David M Brandman, and Sergey D Stavisky. An instantaneous voice-synthesis neuroprosthesis. *Nature*, 644(8075):145–152, 2025.

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

Xiran Xu, Yujie Yan, Xihong Wu, and Jing Chen. Shine: Sequential hierarchical integration network for eeg and meg. *arXiv preprint arXiv:2602.23960*, 2026.

Zihao Xu, Yanru Bai, Ran Zhao, Hongmei Hu, Guangjian Ni, and Dong Ming. Decoding selective auditory attention with eeg using a transformer model. *Methods*, 204:410–417, 2022.

Yiqian Yang, Hyejeong Jo, Yiqun Duan, Qiang Zhang, Jinni Zhou, Xuming Hu, Won Hee Lee, Renjing Xu, and Hui Xiong. Mad: Multi-alignment meg-to-text decoding. *arXiv preprint arXiv:2406.01512*, 2024.

Yiqian Yang, Yiqun Duan, Qiang Zhang, Hyejeong Jo, Jinni Zhou, Won Hee Lee, Renjing Xu, and Hui Xiong. Neuspeech: Decode neural signal as speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6636–6640. IEEE, 2026.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

A PRETRAINED SPEECH MODELS

This section describes the pretrained speech models used in *MindVoice*, distinguishing their original training from their subsequent use or adaptation in our framework.

A.1 FISH-SPEECH S1-MINI

Model and codec architecture. We use OpenAudio S1-mini, referred to as Fish-Speech S1-mini in this work, together with its released speech codec. S1-mini is a 0.5-billion-parameter distilled version of the 4-billion-parameter S1 model. Its codec follows a Descript Audio Codec-like architecture augmented with Transformer modules. The released codec operates at 44.1 kHz and combines convolutional downsampling, temporal Transformer processing, residual vector quantization, and waveform decoding. Its overall downsampling factor is 2048, corresponding to approximately 21.5 codec frames per second.

Codebook composition and residual quantization. The codec contains one semantic-dominant codebook, indexed by $k = 0$, followed by $K = 9$ residual codebooks. The first codebook contains 4096 entries, whereas each residual codebook contains 1024 entries. Thus, each codec frame is represented by ten discrete indices. Codebook lookup is performed in an 8-dimensional space, with learned projections mapping between this space and the 1024-dimensional quantization latent space.

The first codebook is assigned an explicit semantic role through the codec's semantic-distillation design. A training-only prediction module maps its quantized output to wav2vec speech representations, providing content-oriented supervision. Subsequent codebooks progressively encode the remaining quantization residual to preserve additional acoustic detail. Thus, the designation *semantic-dominant* reflects targeted semantic supervision rather than simply the codebook's position in the residual stack. This distinction concerns the emphasis of the learned representations, not a strict separation between linguistic and acoustic information.

Autoregressive speech generation. The speech generator uses a dual-autoregressive architecture. A slow Transformer models dependencies across codec frames and predicts the first-codebook token, while a fast Transformer autoregressively predicts the remaining codebooks within each frame. Let $z_{i,t}^{(k)}$ denote the k-th codebook index contained in $z_{i,t}$, and let $h_{i,t}^{\mathrm{FS}}$ denote the slow Transformer's contextual state. The generation process used in joint reconstruction can be expressed as:

$$p_{\mathrm{FS}}\left(Z_i \mid \widetilde{T}_i, C_i\right) = \prod_{t=1}^{|Z_i|}\left[p_{\mathrm{FS,slow}}\left(z_{i,t}^{(0)} \mid Z_{i,<t}, \widetilde{T}_i, C_i\right) \times \prod_{k=1}^{K} p_{\mathrm{FS,fast}}\left(z_{i,t}^{(k)} \mid z_{i,t}^{(<k)}, h_{i,t}^{\mathrm{FS}}\right)\right], \quad (7)$$

where $z_{i,t}^{(<k)}$ contains the preceding codebook indices within the same frame, and $|Z_i|$ is the number of generated codec frames. Start and end tokens are omitted for clarity.

Usage in MindVoice. In *MindVoice*, the Fish-Speech codec and TTS model remain frozen. The acoustic stream learns to predict the codec indices from neural recordings, while joint reconstruction conditions the pretrained generator on the recovered transcript $\widetilde{T}_i$ and the codec-based speech prompt C_i. The generated sequence is then converted into speech using the pretrained codec decoder $D(\cdot)$.

A.2 WHISPER

Architecture. We use the Whisper-base.en (Radford et al., 2023), an encoder–decoder Transformer with approximately 74 million parameters. It contains six encoder layers and six decoder layers, with hidden width 512 and eight attention heads per layer. Its original audio frontend processes 16 kHz waveforms into 80-channel log-mel spectrograms using a 25 ms analysis window and a 10 ms hop. A 30-second audio segment therefore contains 3000 spectrogram frames. The audio encoder begins with two one-dimensional convolutional layers with kernel size three and strides one and two, respectively, followed by GELU activations, sinusoidal positional embeddings, and bidirectional Transformer blocks. The strided frontend reduces a 30-second input to 1500 encoder positions. The text decoder uses causal self-attention, cross-attention to the encoder outputs, and learned positional embeddings. It autoregressively predicts text tokens from the encoded audio and previously generated tokens.

Original pretraining. Whisper was trained on 680,000 hours of audio paired with transcripts collected from the web, using multilingual and multitask weak supervision. Its tasks include speech transcription and speech-to-English translation, with special tokens specifying language, task, and timestamp information. The decoder is trained through teacher-forced next-token prediction. Suppressing task-control tokens for clarity, the transcription component of this objective is:

$$\mathcal{L}_{\mathrm{Wh,pre}} = -\mathbb{E}_{(M_i,T_i)\sim\mathcal{D}_{\mathrm{Wh}}}\left[\sum_{j=1}^{L_i}\log p_{\mathrm{Wh}}\left(T_{i,j} \mid T_{i,<j}, M_i\right)\right], \tag{8}$$

where $\mathcal{D}_{\mathrm{Wh}}$ denotes the original speech–text pretraining distribution, rather than the neural-recording dataset used in *MindVoice*.

Adaptation in MindVoice. In our linguistic reconstruction stream, Whisper receives the reconstructed mel-spectrogram $\widetilde{M}_i$, rather than the spectrogram M_i of the original speech. During text-supervised adaptation, we optimize the projector and Whisper encoder using transcript next-token supervision together with the alignment objective. The neural signal embedder, speech vector-quantized autoencoder, and Whisper decoder remain frozen. This adaptation is intended to accommodate the distribution of neural-predicted spectrograms while retaining the pretrained decoder's speech-to-text generation prior. Consequently, Whisper provides an audio-conditioned linguistic prior, rather than an independent text-only language model.

B Additional Experiment Settings

B.1 Data Preprocessing

MEG-MASC. We retained the 208 MEG sensor channels and applied a 50 Hz notch filter followed by a 1–58 Hz band-pass filter to the continuous recordings, which were then resampled to 200 Hz. We constructed utterances by grouping consecutive word events with the same story, sound, and sequence identifiers. Each neural segment extended from the onset of the first word to the offset of the last word. The corresponding speech segment was extracted using the timestamps in the source audio, and its transcript was obtained by concatenating the annotated words in order. Multichannel audio was averaged to mono and resampled to 16 kHz. Each MEG segment was normalized independently for each channel using the median and interquartile range of its first 100 samples (0.5 s), or all available samples for shorter segments. A zero interquartile range was replaced by one. The standardized values were clipped to $[-10, 10]$ and divided by 10. We rejected empty segments, segments containing non-finite values, and segments with more than 20% zero-valued entries; these checks were applied both before and after normalization. Empty or non-finite speech segments were also excluded. The resulting dataset contained 33,124 paired neural-speech samples.

Brennan. The EEG recordings were processed at their original sampling rate of 500 Hz with a 60 Hz notch filter and a 0.5–99 Hz band-pass filter, followed by resampling to 200 Hz. For each recording, we centered each channel by its temporal median and scaled it by its interquartile range. The resulting values were clipped to $[-10, 10]$, and the entire recording was divided by its maximum absolute value when this value was nonzero. Inputs were standardized to 62 channels by retaining the first 62 channels when more were available and adding zero-valued channels when fewer were available. The stimulus audio was converted to a mono waveform at 16 kHz. We used WhisperX with Whisper large-v2 to obtain transcripts and refined their timestamps through forced alignment. The resulting segment boundaries were used to extract paired EEG and speech segments. Empty EEG segments were discarded. Each EEG segment was stored using 2,400 temporal samples (12 s): longer segments were truncated, and shorter segments were zero-padded. Its valid length was recorded so that storage padding could be removed during loading. The resulting dataset contained 6,811 samples.

Train-Val-Test Split. We split each dataset into mutually disjoint training, validation, and test sets at an 8:1:1 ratio. The split is performed only once and remains fixed across all experiments with different random seeds, without re-splitting. Model parameters are learned exclusively from the training set, the validation set is used for model selection, and the test set is reserved solely for final evaluation. Strict separation is maintained among the three subsets throughout the entire experimental pipeline, ensuring that no data leakage occurs.

Speech representations. For the linguistic stream, we extracted 80-bin log-Mel spectrograms from the 16 kHz waveforms using the Whisper-base.en feature extractor. From its padded output, we retained the first 2,400 frames, corresponding to a 24 s window at a 10 ms frame hop, and used bilinear interpolation to reduce the temporal dimension to 1,200 frames. This yielded a fixed 80×1200 representation for the speech autoencoder. For the acoustic stream, waveforms were resampled to 44.1 kHz and encoded by the frozen Fish Speech DAC codec. Each speech segment was represented by one semantic codebook with 4,096 entries and nine residual codebooks with 1,024 entries each. The token frame rate was $44{,}100/2{,}048 \approx 21.53$ Hz.

Valid lengths and padding. We retained the valid temporal lengths of neural signals and codec sequences. Variable-length inputs were padded within each mini-batch, with corresponding masks identifying valid positions. Codec tokens introduced by batch padding were excluded from the acoustic loss, and temporal masks restricted neural-to-speech reconstruction losses to valid positions. Transcripts were tokenized using the Whisper tokenizer, with padded text positions ignored in the text-generation loss.

B.2 BASELINES

We describe the two comparison methods and four diagnostic variants used to examine acoustic reconstruction, linguistic recovery, and dependence on neural input. We use the same acoustic and linguistic stream definitions as in the main text.

Vanilla. Vanilla directly regresses Mel spectrograms from EEG/MEG recordings. Its neural encoder consists of temporal convolutions, a spatial convolution that aggregates recording channels, learnable positional embeddings, and a four-layer Transformer with a hidden dimension of 256 and eight attention heads. The encoded sequence is temporally interpolated and mapped to 80-bin Mel spectrograms through an MLP prediction head. The model is trained with supervised spectrogram reconstruction, with padded target frames masked out. A pretrained BigVGAN-v2 vocoder converts the predicted spectrograms into speech waveforms. This baseline evaluates direct acoustic reconstruction without an explicit intermediate transcript.

FESDE. We implement a FESDE-style reconstruction model for both EEG and MEG. A neural encoder extracts features from the recordings, while an auxiliary decoder reconstructs the neural input to regularize the representation. During training, a speech posterior encoder maps the target linear spectrogram to a latent distribution. A connector predicts neural-conditioned prior statistics, and a normalizing flow aligns the speech posterior latents with this prior through a KL-divergence objective. A jointly trained HiFi-GAN-style generator reconstructs the waveform. The training objective combines neural reconstruction, Mel reconstruction, KL divergence, adversarial learning, and discriminator feature matching. At inference, the neural-conditioned prior is mapped through the inverse flow and waveform generator, bypassing the speech posterior encoder. This baseline reconstructs speech through a continuous latent representation without explicit text recovery.

MindVoice-A. MindVoice-A retains only the acoustic reconstruction stream. Given a neural recording, it predicts the semantic-dominant codec tokens and the nine residual-codebook token sequences. These predictions are passed directly to the frozen Fish-Speech codec decoder to obtain the waveform. The linguistic stream and the autoregressive TTS generation stage are omitted. This variant measures the reconstruction quality attainable from neural-to-codec prediction alone.

MindVoice-L. MindVoice-L retains the linguistic reconstruction stream and the pretrained TTS model. Neural recordings are mapped to the speech autoencoder's latent space, decoded into Mel spectrograms, and transcribed by the adapted Whisper model. The predicted transcript is then supplied to Fish-Speech for speech generation without an acoustic prompt. This variant evaluates reconstruction driven by recovered linguistic content and pretrained speech-generation priors. It reuses the trained linguistic stream and frozen TTS model.

Noise-A. Noise-A replaces the neural input to the acoustic stream with independent standard Gaussian noise, while the linguistic stream continues to receive the original neural recording. Noise is sampled from $\mathcal{N}(0, 1)$. The acoustic input retains the original channel count. Codec tokens

predicted from noise are combined with the transcript recovered from the original neural signal for TTS synthesis. Noise input is applied at inference with all model parameters fixed.

Noise-L. Noise-L replaces the neural input to the linguistic stream with independent standard Gaussian noise, while the acoustic stream receives the original recording. The noise-derived transcript is combined with codec tokens predicted from the original neural signal for TTS synthesis. Noise input is applied at inference with all model parameters fixed.

B.3 Evaluation Metrics

Let y_i and $\widehat{y}_i$ denote the reference and reconstructed waveforms for example i, and let T_i denote its reference transcript. Audio features are evaluated on mono waveforms resampled to 16 kHz. For linguistic evaluation, we transcribe the final reconstructed waveform using Qwen3-ASR-0.6B with the language set to English, yielding $\widehat{T}_i$. Thus, linguistic metrics assess the generated speech rather than the intermediate transcript supplied to TTS. Unless stated otherwise, we compute each score per example and report its arithmetic mean over the evaluation set. WER and CER aggregate edit counts over the corpus, while BLEU aggregates corpus-level n-gram counts. The arrows $\uparrow$ and $\downarrow$ indicate whether higher or lower values are preferable.

B.3.1 Main Evaluation Metrics

Mel-spectrogram mean squared error (mel-MSE; $\downarrow$). We extract 80-bin log-Mel features using the Whisper base.en feature extractor, with a 400-sample Hann window and a 160-sample hop (25 ms and 10 ms, respectively). We retain Whisper's standard log compression and scaling: for Mel power P, $L = \log_{10}(\max(P, 10^{-10}))$ and $M = (\max(L, L_{\max} - 8) + 4)/4$, where $L_{\max}$ is computed separately for each waveform. Feature extraction uses the default 30 s window. We compare the first $F_i = \min(3000, \lceil |y_i|/160 \rceil, \lceil |\widehat{y}_i|/160 \rceil)$ frames, excluding the remaining padding. For the resulting $M_i, \widehat{M}_i \in \mathbb{R}^{80 \times F_i}$,

$$\text{mel-MSE}_i = \frac{1}{80F_i} \sum_{b=1}^{80} \sum_{t=1}^{F_i} \left(M_{i,b,t} - \widehat{M}_{i,b,t} \right)^2 . \tag{9}$$

These evaluation features retain their original temporal resolution; they are not the temporally compressed features used to train the speech autoencoder. No dynamic time warping (DTW) is applied to the Mel metrics.

Mel-spectrogram correlation (mel-r; $\uparrow$). Using the same feature matrices and valid frames, we flatten both matrices and compute Pearson correlation:

$$\text{mel-}r_i = \rho\Big(\text{vec}(M_i), \text{vec}(\widehat{M}_i) \Big), \qquad \rho(u, v) = \frac{\sum_j (u_j - \bar{u})(v_j - \bar{v})}{\sqrt{\sum_j (u_j - \bar{u})^2} \sqrt{\sum_j (v_j - \bar{v})^2}}. \tag{10}$$

The implementation returns zero when there are fewer than two entries or either standard deviation is below 10^{-12}. This convention also applies to the trajectory correlations below.

Word error rate (WER; $\downarrow$). For edit-based and lexical metrics, we case-fold transcripts, convert typographic apostrophes to straight apostrophes, remove ASCII punctuation, and collapse whitespace. WER is computed with `jiwer` using word-level minimum-edit alignment:

$$\text{WER} = \frac{\sum_i (S_i + D_i + I_i)}{\sum_i N_i}, \tag{11}$$

where S_i, D_i, and I_i are substitution, deletion, and insertion counts, and N_i is the number of reference words. We report WER as a percentage. Insertions can make WER exceed 100%.

ASR-BERT ($\uparrow$). We compute BERTScore using layer 17 of RoBERTa-large on the original ASR and reference strings, without the lexical normalization above or inverse-document-frequency weighting. Let u_j and v_k be unit-normalized contextual token embeddings of the predicted and reference

transcripts, respectively. Greedy token matching gives

$$P_{\mathrm{B}} = \frac{1}{n}\sum_{j=1}^{n}\max_{k} u_j^\top v_k, \quad R_{\mathrm{B}} = \frac{1}{m}\sum_{k=1}^{m}\max_{j} u_j^\top v_k, \quad F_{\mathrm{B}} = \frac{2P_{\mathrm{B}}R_{\mathrm{B}}}{P_{\mathrm{B}}+R_{\mathrm{B}}}. \tag{12}$$

The main-text ASR-BERT score is the baseline-rescaled F1:

$$\text{ASR-BERT}_i = \frac{F_{\mathrm{B},i} - b_F}{1 - b_F}, \qquad b_F = 0.83122575. \tag{13}$$

Predictions that become empty after lexical normalization receive a rescaled score of zero. Other scores are not clipped, so negative values are possible.

HuBERT similarity (↑). We use the final hidden states of `hubert-base-ls960`. The checkpoint's feature extractor preprocesses each waveform, and attention masks exclude padding from temporal mean pooling. Writing $h_t(y)$ for a valid hidden state and $\mathcal{V}(y)$ for the valid frame set,

$$e(y) = \frac{1}{|\mathcal{V}(y)|}\sum_{t\in\mathcal{V}(y)} h_t(y), \qquad s_i = \frac{e(y_i)^\top e(\widehat{y}_i)}{\max(\|e(y_i)\|_2\|e(\widehat{y}_i)\|_2, 10^{-12})}. \tag{14}$$

Reference and predicted embeddings are extracted independently from their full waveforms. Waveforms longer than 20 s are encoded in contiguous chunks of at most 20 s, whose pooled embeddings are averaged with weights proportional to their waveform sample counts.

WavLM similarity (↑). We apply the same masked pooling, chunk aggregation, and cosine-similarity calculation using `wavlm-base-sv`. The score uses the final hidden states of `WavLMModel`, rather than a separate speaker-verification classification head. HuBERT and WavLM provide complementary comparisons in learned speech-representation spaces.

Articulatory correlation (SPARC-r; ↑). The pretrained SPARC feature extractor infers 12-dimensional articulatory trajectories at 50 Hz from each waveform. These are inferred trajectories, not articulatory measurements collected from the participants. For dimension d, we compute σ_d from the concatenated trajectories of the distinct reference audio files in the evaluation set, with a minimum value of 10^{-4}. Both reference and predicted trajectories are divided by σ_d. We then perform multivariate DTW using Euclidean distance on these normalized trajectories. If $\mathcal{P}_i = \{(p_\ell, q_\ell)\}_{\ell=1}^{L_i}$ is the alignment path and $A_i, \widehat{A}_i$ are the normalized trajectories,

$$\text{SPARC-}r_i = \frac{1}{12}\sum_{d=1}^{12}\rho\Big((A_{i,p_\ell,d})_{\ell=1}^{L_i}, (\widehat{A}_{i,q_\ell,d})_{\ell=1}^{L_i}\Big). \tag{15}$$

For short waveforms that fail the extractor's filtering requirements, we reflection-pad to 0.5 s and crop the resulting trajectory to the original duration. If full-utterance extraction fails for a longer waveform, we extract contiguous 8 s chunks and concatenate their trajectories.

B.3.2 Additional Evaluation Metrics

Mel-cepstral distortion (MCD; ↓). We extract WORLD spectral envelopes at a 5 ms frame period using DIO, StoneMask, and CheapTrick, with a DIO pitch range of 50–500 Hz. We convert the envelopes to order-25 Mel-cepstral coefficients using `pysptk`, with the frequency-warping parameter returned by `mcepalpha` for 16 kHz. The zeroth coefficient is excluded. Euclidean DTW aligns the remaining 25-dimensional sequences. With alignment path $\mathcal{Q}$,

$$\mathrm{MCD} = \frac{10\sqrt{2}}{|\mathcal{Q}|\ln 10}\sum_{(p,q)\in\mathcal{Q}}\sqrt{\sum_{d=1}^{25}(c_{p,d} - \widehat{c}_{q,d})^2}. \tag{16}$$

MCD is reported in decibels and uses both full utterances, allowing differences in speaking rate through DTW.

UTMOS (↑). We use the pretrained `utmos22_strong` model from SpeechMOS to estimate the perceptual quality of each complete generated waveform at 16 kHz. This is a reference-free, automatic MOS prediction; it is not a human listening-test score.

SPARC normalized error and similarity. Using the same normalized trajectories and DTW path as SPARC-r, we additionally compute

$$\text{SPARC-NRMSE}_i\ (\downarrow) = \sqrt{\frac{1}{12L_i}\sum_{\ell=1}^{L_i}\sum_{d=1}^{12}(A_{i,p_\ell,d} - \widehat{A}_{i,q_\ell,d})^2}, \tag{17}$$

$$\text{SPARC-Sim}_i\ (\uparrow) = \exp(-\text{SPARC-NRMSE}_i). \tag{18}$$

The exponential is applied per example before averaging. These scores measure trajectory error and its monotonic similarity transform, respectively.

Character error rate (CER; ↓). CER uses the same corpus-level edit-count formula as WER, with characters as the alignment units and reference characters as the denominator. Normalized spaces between words are retained as characters. Like WER, CER can exceed one because of insertions.

Exact match and token F1 (↑). Exact match is the fraction of normalized transcript pairs that are identical:

$$\text{EM} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}[\text{norm}(\widehat{T}_i) = \text{norm}(T_i)]. \tag{19}$$

For token F1, transcripts are tokenized by whitespace. If O_i is the multiset overlap count, and n_i, m_i are the predicted and reference token counts, then $F_{1,i} = 2O_i/(n_i + m_i)$. Two empty sequences receive one; pairs with no overlap otherwise receive zero. The final token F1 is the mean across examples.

BLEU-1 and BLEU-4 (↑). We compute corpus BLEU using NLTK on the normalized, whitespace-tokenized transcripts, with one reference per example:

$$\text{BLEU}_K = \text{BP}\exp\left(\sum_{n=1}^{K} w_n \log p_n\right), \qquad \text{BP} = \min(1, \exp(1 - r/c)), \tag{20}$$

where p_n is clipped corpus n-gram precision, and r, c are the total reference and candidate lengths. BLEU-1 uses $w_1 = 1$; BLEU-4 uses $w_1 = \cdots = w_4 = 1/4$. We apply NLTK smoothing method 1, which adds 0.1 to zero match counts; the library returns zero when no unigram matches exist. Empty candidates receive zero. Scores are stored on the $[0, 1]$ scale.

ROUGE-1-F1 and ROUGE-L-F1 (↑). We use `rouge-score` without stemming and average sentence-level F1 scores. Following lexical normalization, the package applies its own lowercase alphanumeric tokenizer. ROUGE-1 uses unigram multiset overlap. ROUGE-L uses the length ℓ_i of the longest common subsequence:

$$P_{L,i} = \frac{\ell_i}{n_i}, \qquad R_{L,i} = \frac{\ell_i}{m_i}, \qquad F_{L,i} = \frac{2P_{L,i}R_{L,i}}{P_{L,i} + R_{L,i}}. \tag{21}$$

Empty-token cases receive zero under the package's convention. ROUGE and token F1 therefore have different tokenization and empty-input conventions.

Additional BERTScore statistics (↑). Besides the main ASR-BERT score, we retain raw BERTScore precision, recall, and F1, together with rescaled precision and recall. For $s \in \{P, R, F\}$, rescaling is $s' = (s - b_s)/(1 - b_s)$, with

$$b_P = 0.83150584, \qquad b_R = 0.83149410, \qquad b_F = 0.83122575. \tag{22}$$

For empty normalized predictions, each raw statistic is set to its corresponding baseline, producing zero after rescaling. All six BERTScore statistics are averaged per example; the rescaled F1 is the ASR-BERT metric already defined above.

Acoustic-token diagnostics. For the acoustic stream, we additionally record codec-token classification accuracy on valid target frames. For codebook k,

$$\mathrm{Acc}_k = \frac{\sum_i \sum_{t \in \mathcal{V}_i} \mathbf{1}[\widehat{c}_{i,k,t} = c_{i,k,t}]}{\sum_i |\mathcal{V}_i|}. \tag{23}$$

We report the first-codebook accuracy, pooled accuracy over the nine residual codebooks, pooled accuracy over all ten codebooks, and the mean of per-example accuracies. Padding is excluded. The pooled score counts individual codebook tokens, rather than requiring all ten tokens at a frame to match simultaneously. These are intermediate diagnostics specific to codec prediction, separate from the 23 waveform and transcript metrics above.

B.4 Implementation Details

Neural encoders and input formatting. We implement *MindVoice* in PyTorch and train separate models for Brennan and MEG-MASC. The acoustic and linguistic streams have independent parameters. Both use temporal convolutions with kernel widths 5 and 3, stride 2, and output widths 64 and 128, followed by a spatial convolution spanning all recording channels and producing 256-dimensional features. Each convolution is followed by batch normalization and ELU. Learnable positional embeddings and a four-layer Transformer then encode the sequence, with eight attention heads, a feed-forward dimension of 1,024, and dropout 0.1. The acoustic encoder additionally uses a learned subject embedding and masks padded positions in self-attention. The preprocessed 200 Hz neural signals are subsampled by a factor of two before entering either stream. The linguistic stream uses a fixed 24 s input window, obtained by truncating or zero-padding to 4,800 samples before subsampling to 2,400 samples. Brennan retains at most the 12 s of valid EEG stored during preprocessing. The acoustic stream operates on variable-length inputs, with valid-length interpolation to the codec frame grid. Its inference output length is $\lceil d_i f_c \rceil$, where d_i is the valid neural duration and $f_c = 44{,}100/2{,}048$ Hz is the codec frame rate.

Speech autoencoder. The Mel-domain VQ autoencoder takes 80×1200 inputs and downsamples both axes by four, producing an $8 \times 20 \times 300$ latent tensor. Its codebook contains 2,048 eight-dimensional entries; latent vectors and codebook entries are normalized before nearest-neighbor assignment. The objective is Mel reconstruction MSE plus one half of the quantizer loss, whose codebook and commitment terms have relative weights 1 and 0.25. We train it for 100 epochs with batch size 16 and learning rate 2×10^{-4}. These trainings use full precision, AdamW with weight decay 0.01, cosine learning-rate decay, and gradient clipping at norm 1.0. The autoencoder is frozen throughout subsequent neural training.

Acoustic-stream training. We use the frozen Fish-Speech codec distributed with OpenAudio S1-mini. The acoustic model predicts one 4,096-entry semantic-dominant codebook and nine 1,024-entry residual codebooks. The embedding of the predicted first-codebook token conditions residual prediction. During training, this conditioning token is replaced by its target with probability 0.1 independently at each position; inference uses predictions throughout. We use label-smoothed cross-entropy with smoothing 0.05 and equal weights for the first-codebook loss and the mean residual-codebook loss. Optimization uses AdamW with learning rate 2×10^{-4}, weight decay 0.01, three warm-up epochs followed by cosine decay, mixed precision, and gradient clipping at norm 1.0. Batch sizes are 24 for MEG-MASC and 32 for Brennan, with at most 30 and 40 epochs, respectively. We retain the checkpoint with the highest validation codec-token accuracy, excluding padding, and stop after eight epochs without improvement.

Neuro-to-language alignment. The linguistic projector combines a feature projection, a temporal convolution, and a residual MLP to produce the $8 \times 20 \times 300$ speech-latent tensor. With the speech autoencoder fixed, we train the neural encoder and projector using latent MSE, Mel MSE and contrastive loss. The first two terms have weight 1, the contrastive term has weight 0.05 and temperature 0.07. This stage uses AdamW with learning rate 4×10^{-4}, weight decay 0.01, cosine decay, batch size 16, mixed precision, and gradient clipping at norm 1.0. Training lasts at most 40 epochs, with patience 5 and checkpoint selection by the lowest total validation loss. Reconstruction losses exclude invalid temporal positions.

Table 3: **Additional lexical evaluation**. Values are means across five training seeds.

Data	Method	CER(%) ↓	EM(%) ↑	Token F1 ↑	BLEU-1 ↑	BLEU-4 ↑	ROUGE-1 ↑	ROUGE- L↑
EEG	Vanilla	79.10	0.73	0.115	0.046	0.020	0.115	0.110
	FESDE	97.32	0.00	0.031	0.000	0.000	0.031	0.031
	MindVoice-A	84.00	0.15	0.051	0.027	0.002	0.051	0.048
	MindVoice	**33.77**	**59.53**	**0.792**	**0.849**	**0.829**	**0.792**	**0.789**
MEG	Vanilla	96.02	0.21	0.078	0.090	0.026	0.078	0.074
	FESDE	97.29	0.06	0.022	0.000	0.000	0.022	0.022
	MindVoice-A	71.20	8.07	0.306	0.250	0.130	0.307	0.302
	MindVoice	**32.63**	**29.41**	**0.583**	**0.631**	**0.575**	**0.584**	**0.582**

Text calibration and decoding. We initialize text calibration from the selected alignment checkpoint and use Whisper-base.en. Only the projector, the last four Whisper encoder layers, and the encoder's final layer normalization are updated. The neural encoder, speech autoencoder, and Whisper decoder remain frozen; the neural encoder's batch-normalization statistics are also fixed. Learning rates are 10^{-4} for the projector and 10^{-5} for the trainable Whisper parameters. The objective combines transcript cross-entropy with 0.05 times the alignment objective. We use AdamW with weight decay 0.01, cosine decay, batch size 16, mixed precision, gradient clipping at norm 1.0, at most 40 epochs, and patience 5. The checkpoint is selected by total validation loss. Reconstructed Mel features are interpolated from 1,200 to 2,400 frames. Frames beyond the valid neural duration are set to -1, and 600 additional frames with this value are appended to obtain Whisper's 80×3000 input. An attention mask identifies valid frames. Text is decoded separately for each neural observation with deterministic beam search: five beams, at most 64 new tokens, length penalty 1.2, repetition penalty 2.0, and repeated-bigram blocking. The decoding configuration is fixed before test evaluation.

Joint speech reconstruction. We use the pretrained OpenAudio S1-mini Fish-Speech generator and its codec without further updates. The predicted transcript is supplied both as the transcript associated with the in-context acoustic prompt and as the target text. Generation uses top-$p = 0.5$, temperature 0.5, and repetition penalty 1.1. One second of codec-encoded silence is appended to each acoustic prompt. The generated sequence is limited to $\lceil 1.15 L_i \rceil$ codec frames, where L_i is the full brain-derived acoustic sequence length before prompt cropping; shorter outputs are not padded. Decoded waveforms are resampled to 16 kHz.

Linguistic-stream prompt refinement. We first generate a codec sequence from the predicted text without an acoustic prompt, then resample its token indices to the brain-derived sequence length by uniform nearest-neighbor indexing. We replace only its first residual-codebook row with the corresponding brain-predicted row, retaining the text-generated semantic-dominant row and remaining eight residual rows. This refined sequence is used as an in-context prompt for a subsequent TTS generation pass. For MEG-MASC, we further select the prompt that consists of the central 3 s of the predicted acoustic sequence, or the whole sequence when shorter. The dataset-specific prompt policies and generation settings are selected using validation data and fixed before test evaluation. Both reconstruction streams and all pretrained generation components remain frozen during joint speech reconstruction.

C Supplementary Experimental Results

C.1 Comprehensive Quantitative Comparison

We extend the main-text comparison of Vanilla, FESDE, *MindVoice-A*, and *MindVoice* with the remaining 16 waveform and transcript metrics defined in Appendix B.3. Tables 3–5 cover lexical agreement, additional BERTScore statistics, perceptual, cepstral, and articulatory measures. All transcript metrics are computed from ASR transcriptions of the final generated speech.

Linguistic reconstruction. Table 3 shows that *MindVoice* achieves the best results across all additional lexical metrics on both datasets. Compared with *MindVoice-A*, CER decreases from 84.00% to 33.77% on EEG and from 71.20% to 32.63% on MEG, while exact match increases to 59.53% and 29.41%, respectively. The improvements in BLEU and ROUGE further indicate better agreement with the reference content at both word and sequence levels.

Table 4: **Additional BERTScore statistics**. Values are means across five training seeds. P_{B}, R_{B}, and F_{B} are raw scores; P'_{B} and R'_{B} are baseline-rescaled scores. Rescaled recall can be negative.

Data	Method	$P_{\mathrm{B}} \uparrow$	$R_{\mathrm{B}} \uparrow$	$F_{\mathrm{B}} \uparrow$	$P'_{\mathrm{B}} \uparrow$	$R'_{\mathrm{B}} \uparrow$
EEG	Vanilla	0.876	0.839	0.857	0.264	0.042
	FESDE	0.897	0.826	0.859	0.386	−0.031
	MindVoice-A	0.851	0.833	0.842	0.118	0.012
	MindVoice	**0.960**	**0.958**	**0.959**	**0.765**	**0.750**
MEG	Vanilla	0.858	0.826	0.841	0.157	−0.033
	FESDE	**0.918**	0.821	0.866	**0.513**	−0.064
	MindVoice-A	0.871	0.859	0.865	0.234	0.164
	MindVoice	0.917	**0.909**	**0.913**	0.507	**0.460**

Table 5: **Additional perceptual, cepstral, and articulatory evaluation.** Values are means across five training seeds. MCD and SPARC measures use DTW alignment.

Data	Method	MCD(dB) ↓	UTMOS ↑	SPARC-NRMSE ↓	SPARC-Sim ↑
EEG	Vanilla	10.10	1.275	0.819	0.443
	FESDE	11.29	1.449	1.108	0.334
	MindVoice-A	**10.02**	1.266	0.923	0.399
	MindVoice	11.90	**3.455**	**0.693**	**0.510**
MEG	Vanilla	9.44	1.257	0.798	0.452
	FESDE	10.69	1.493	1.001	0.369
	MindVoice-A	**8.37**	1.596	**0.743**	**0.489**
	MindVoice	9.15	**3.892**	0.748	0.486

Semantic precision and coverage. The additional BERTScore statistics in Table 4 show that *MindVoice* achieves the highest recall and raw F1 on both datasets. Although FESDE has slightly higher precision on MEG, its low recall indicates limited coverage of the reference content. MindVoice maintains comparable precision while substantially improving recall, suggesting more complete semantic reconstruction.

Perceptual quality and acoustic fidelity. As shown in Table 5, *MindVoice* achieves the highest UTMOS on both datasets, suggesting improved perceptual quality according to this automatic estimator. It also obtains the lowest SPARC-NRMSE and highest SPARC-Sim on EEG, whereas *MindVoice-A* retains a small advantage on these measures for MEG. Meanwhile, *MindVoice-A* achieves the lowest MCD on both datasets. These results indicate that the improvements in linguistic reconstruction and predicted perceptual quality do not uniformly translate into closer cepstral or articulatory correspondence with the reference speech.

C.2 Results Across Five Runs

Table 6 reports the seven main evaluation metrics for each of the five MindVoice runs on Brennan EEG and MEG-MASC.

C.3 Ground-Truth Input Control Experiments

We examine how acoustic and linguistic inputs affect speech reconstruction by replacing the acoustic-token prompt and the supplied text at inference time. In Table 7, *Pred.* denotes the corresponding neural-stream prediction, and *GT* denotes codec tokens extracted from the target reference waveform or its reference transcript. *Mism.* denotes reference tokens or text from a different evaluation example, assigned through a fixed one-to-one mapping. All outputs are evaluated against the original target waveform and transcript. The five GT and mismatched-input controls use frozen models with a fixed random seed for the neural-stream artifacts and generation. Conditions using GT inputs are diagnostic controls with reference information supplied at inference time.

Role of linguistic input. Within the control protocol, fixing GT acoustic tokens and replacing predicted text with GT text reduces WER from 18.74% to 5.06% on EEG and from 52.21% to 16.69% on MEG. Conversely, pairing GT acoustic tokens with mismatched text yields WERs of 93.91%

Table 6: **Individual results across five MindVoice runs**. Each row reports one run.

Data	Run	mel-MSE ↓	mel-r ↑	WER(%) ↓	ASR-BERT ↑	HuBERT ↑	WavLM ↑	SPARC-r ↑
EEG	1	0.290	0.378	40.35	0.736	0.865	0.857	0.736
	2	0.285	0.383	39.61	0.747	0.869	0.863	0.745
	3	0.286	0.387	34.49	0.789	0.870	0.868	0.763
	4	0.286	0.386	36.65	0.770	0.872	0.867	0.758
	5	0.289	0.377	39.51	0.747	0.866	0.861	0.742
MEG	1	0.330	0.341	41.23	0.502	0.871	0.843	0.670
	2	0.329	0.341	43.48	0.488	0.871	0.844	0.664
	3	0.328	0.343	47.19	0.460	0.872	0.844	0.653
	4	0.330	0.341	44.46	0.483	0.872	0.843	0.660
	5	0.329	0.342	44.99	0.482	0.869	0.841	0.660

Table 7: **Ground-truth input controls.** Pred., GT, and Mism. denote predicted, target ground-truth, and mismatched reference inputs, respectively.

Data	Acoustic	Linguistic	mel-MSE ↓	mel-r ↑	WER(%) ↓	ASR-BERT ↑	HuBERT ↑	WavLM ↑	SPARC-r ↑
EEG	Pred.	Pred.	0.290	0.378	40.35	0.736	0.865	0.857	0.736
	GT	GT	**0.185**	**0.531**	**5.06**	**0.921**	**0.947**	**0.962**	**0.911**
	GT	Pred.	0.199	0.501	18.74	0.767	0.924	0.929	0.823
	Pred.	GT	0.278	0.405	8.08	0.863	0.874	0.865	0.809
	GT	Mism.	0.245	0.408	93.91	0.153	0.841	0.812	0.480
	Mism.	GT	0.240	0.421	45.55	0.632	0.865	0.860	0.674
MEG	Pred.	Pred.	0.329	0.341	43.48	0.488	0.871	0.844	0.664
	GT	GT	**0.268**	**0.412**	**16.69**	**0.690**	**0.926**	**0.924**	**0.821**
	GT	Pred.	0.281	0.385	52.21	0.445	0.902	0.888	0.696
	Pred.	GT	0.309	0.370	25.36	0.657	0.888	0.868	0.770
	GT	Mism.	0.323	0.332	106.73	0.063	0.825	0.781	0.450
	Mism.	GT	0.320	0.335	37.79	0.601	0.862	0.844	0.675

and 106.73%, respectively. These results indicate that correct acoustic prompts alone do not ensure recovery of the target content when the supplied text is incorrect.

Role of acoustic input. Within the control protocol, fixing GT text and replacing predicted acoustic tokens with GT tokens increases HuBERT similarity from 0.874 to 0.947 on EEG and from 0.888 to 0.926 on MEG, with corresponding improvements in WavLM similarity and SPARC-r. Acoustic input also affects linguistic fidelity: with GT text fixed, mismatched acoustic prompts yield higher WER than predicted prompts on both datasets (45.55% versus 8.08% on EEG; 37.79% versus 25.36% on MEG). Thus, the two inputs have interacting effects on the generated speech.

C.4 Additional Mel-spectrogram Visualization

Figures 5 and 6 further illustrate the distinction between spectral fidelity and linguistic intelligibility. Vanilla often produces diffuse spectral energy with limited temporal detail, whereas *MindVoice* exhibits more structured harmonic patterns and clearer transitions between speech segments. However, *MindVoice*-A can preserve closer correspondence to the reference spectrum, particularly in MEG examples (ii) and (iii), where it achieves substantially lower mel-MSE than *MindVoice*. Similar discrepancies appear in the EEG examples: clearer spectral structure in *MindVoice* does not necessarily yield lower reconstruction error.

C.5 Additional Transcription Examples

Additional transcription results. Figure 7 illustrates partial linguistic recovery across the selected examples. Compared with Vanilla and MindVoice-A, MindVoice generally preserves more reference words and longer meaningful phrases, yielding lower WER and higher ASR-BERT in every displayed case. Nevertheless, its outputs still contain word substitutions, repetitions, and omissions. Some examples retain most of the sentence structure despite local recognition errors, whereas others recover a substantial portion of the utterance but omit its ending. These results illustrate improved preservation of linguistic content without fully accurate transcription.

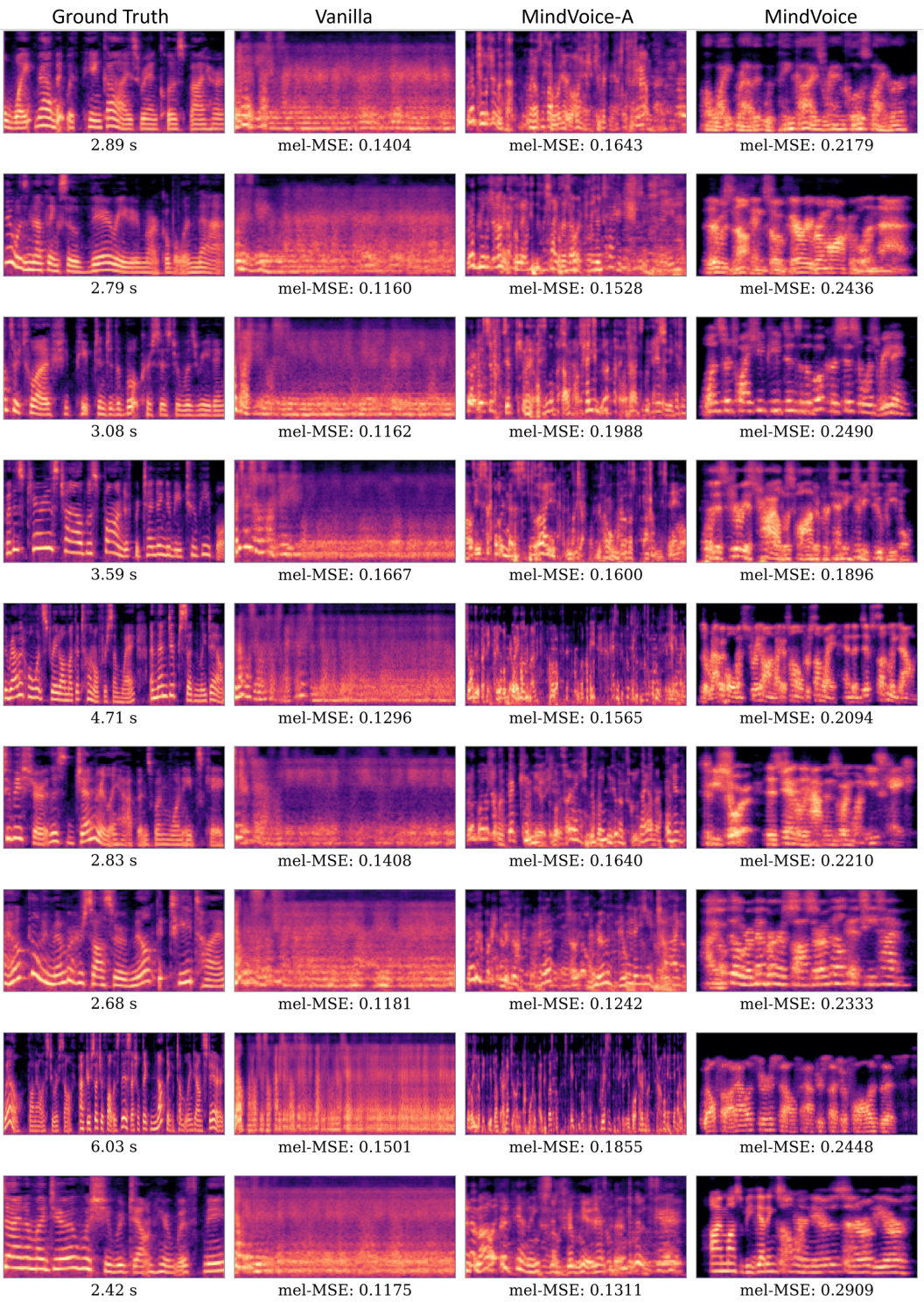


Figure 5: **Additional EEG visualization results**.

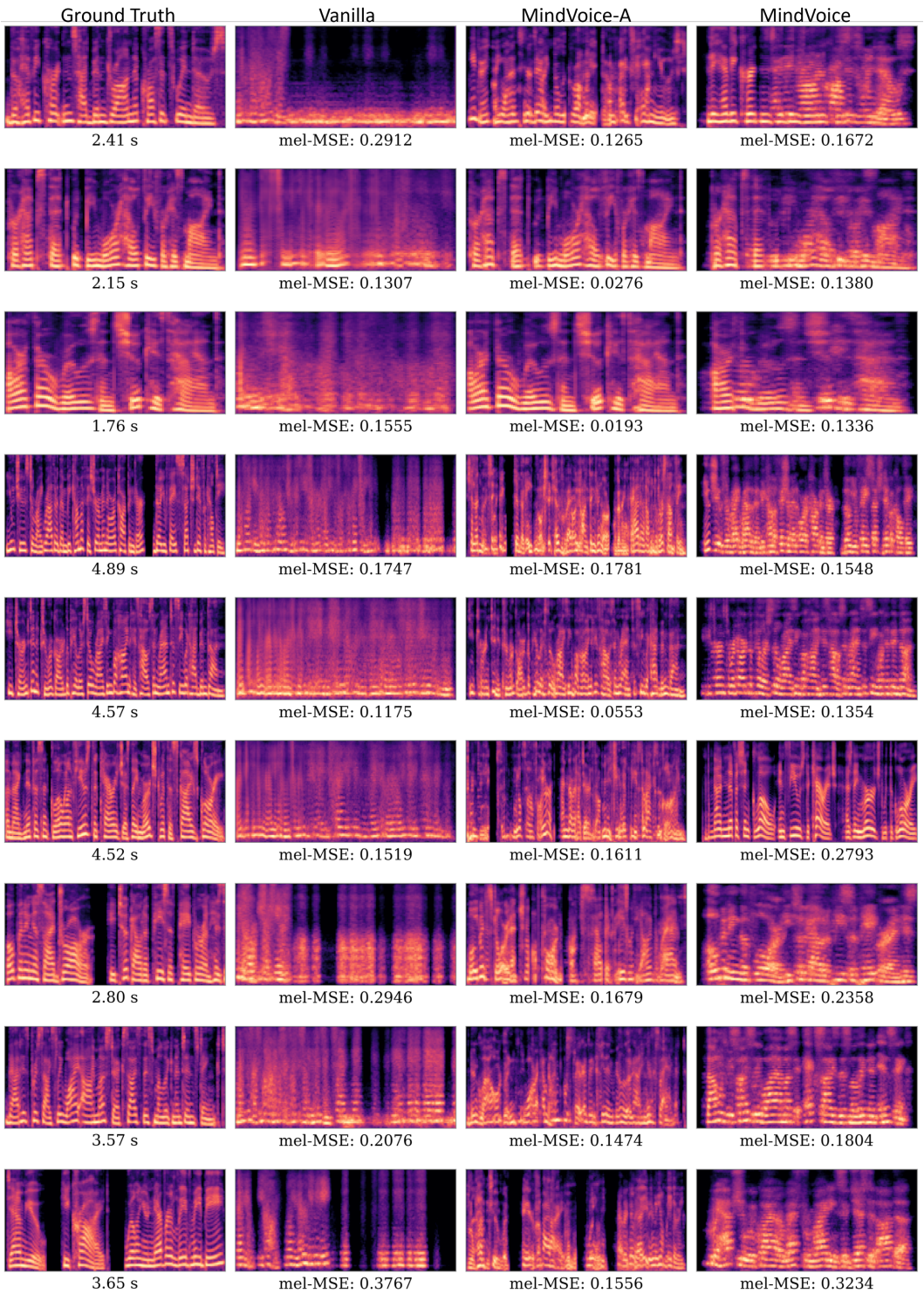


Figure 6: **Additional MEG visualization results**.

(1)	GT:	How could any writer accommodate the edicts of popular literature with the very forces of nature knocking upon his door
	Vanilla:	I'm so happy to see you. (WER=100.0%, ASR-BERT=0.113)
	MindVoice-A:	How could any writer commentate events so popularly enjoyed the very forces of nature knocking upon his door? (WER=35.0%, ASR-BERT=0.564)
	MindVoice:	Capitally writer accommodate the idios of popular literature with the very forces of nature knocking upon his door. (WER=20.0%, ASR-BERT=0.625)
(2)	GT:	Why is that more desirable than what you write today
	Vanilla:	Hmm. (WER=100.0%, ASR-BERT=0.239)
	MindVoice-A:	I have to reduce the air pressure in the way you relate to debt. (WER=130.0%, ASR-BERT=0.127)
	MindVoice:	Why is that more desirable than what you write to dead? (WER=20.0%, ASR-BERT=0.698)
(3)	GT:	After filling his pen Allan looked at his paper in the orange glow from the lantern set back in the right hand corner
	Vanilla:	Article 4 is an overlooked detail in your story from the legend said that it was just right and autumn. (WER=87.0%, ASR-BERT=-0.041)
	MindVoice-A:	After sewing his gown, I looked at his pink and the orange robe from the lantern set back in the desk's right hand corner. (WER=30.4%, ASR-BERT=0.496)
	MindVoice:	Up to filling his pen, Alan looked at his paper in the orange glow from the lantern set back in the right-hand corner. (WER=21.7%, ASR-BERT=0.874)
(4)	GT:	Emerging from between the trunks Allan set his feet softly down onto the carpet of needles and stopped
	Vanilla:	Hmm. (WER=100.0%, ASR-BERT=0.150)
	MindVoice-A:	Emerging from between the edge of the BMW, he hears shrill exclamations. Alan sits his feet softly down onto the carpet of tables and chairs. (WER=66.7%, ASR-BERT=0.406)
	MindVoice:	A martian comes between the trunks. Alan set his feet softly down onto the carpet of needles and stopped. (WER=22.2%, ASR-BERT=0.686)
(5)	GT:	They parted and Allan made his way from gate
	Vanilla:	I'm in love with this place. I'm so in love with this place. (WER=144.4%, ASR-BERT=0.049)
	MindVoice-A:	They thought they did the whole game as many minutes. (WER=100.0%, ASR-BERT=0.051)
	MindVoice:	They pottered, and Alan made his way from gate. (WER=22.2%, ASR-BERT=0.652)
(6)	GT:	Allan crouched over his desk once more pen in hand and mind blank
	Vanilla:	Hmm. (WER=100.0%, ASR-BERT=0.210)
	MindVoice-A:	Paulin crawled to his desk once more, to not turn his hand in mind blank. (WER=61.5%, ASR-BERT=0.385)
	MindVoice:	Alan crouched over his desk once more, pen in hand in mind blank blank. (WER=23.1%, ASR-BERT=0.742)
(7)	GT:	Opening a side drawer he took out a piece of paper and his
	Vanilla:	Okay, inside door. It's about a piece of paper. (WER=69.2%, ASR-BERT=0.272)
	MindVoice-A:	We can store in a drop form of sodium hydride. (WER=92.3%, ASR-BERT=-0.074)
	MindVoice:	Opening the floor, he took out a piece of paper and his. (WER=23.1%, ASR-BERT=0.600)
(8)	GT:	I want to write as others do a tale set in our real world characters who might live therein not citizens some airy beyond the stars
	Vanilla:	Oh, oh, (WER=100.0%, ASR-BERT=-0.256)
	MindVoice-A:	I'm on my way to the office. I'm going to have to get out of here. I'm going to have to get out of here. (WER=100.0%, ASR-BERT=-0.036)
	MindVoice:	Like writers, others do a tale set in our real world, characters who might live there and not citizens some airy beyond the stars. (WER=26.9%, ASR-BERT=0.666)
(9)	GT:	Before that blasted impulse came over me and I started writing like Edgar Poe though the most recent was not so morbid as my many others thank Heaven
	Vanilla:	This one it lasted in post-Apollo era, and I started wearing like a bow. But the most recent was not so weird as my many others. And then. (WER=50.0%, ASR-BERT=0.311)
	MindVoice-A:	This was to impel me really as I started writing. I should tell the most reason was not to forgive me than others. But heaven. (WER=67.9%, ASR-BERT=0.149)
	MindVoice:	Before that blasted impulse came over me, and I started writing like Edgar Poe, though the most recent. (WER=35.7%, ASR-BERT=0.632)
(10)	GT:	Well, thought Alice to herself, after such a fall as this I shall think nothing of tumbling down stairs.
	Vanilla:	I'm a little bit of (WER=89.5%, ASR-BERT=-0.092)
	MindVoice-A:	So they know that you're not a natural person. You're not a person. (WER=94.7%, ASR-BERT=0.070)
	MindVoice:	Well thought, Alice to herself after such a fall as this. (WER=42.1%, ASR-BERT=0.488)
(11)	GT:	Your writing is for you Allan not for the publishers or some imagined audience
	Vanilla:	You want to be this? I want to be this. (WER=92.9%, ASR-BERT=0.041)
	MindVoice-A:	Can Christians like in the law? I'll let you know the truth. Yeah, I'm not gonna let us. (WER=121.4%, ASR-BERT=-0.005)
	MindVoice:	Didn't bring anything, Alan, not for the publishers or some imagined audience. (WER=42.9%, ASR-BERT=0.407)
(12)	GT:	I had a perfect story in my pages my hero returned home and reclaimed his love I nearly had it tied together before
	Vanilla:	I'm a musician. I'm a musician. I'm a musician. I'm a musician. I'm a musician. (WER=95.7%, ASR-BERT=0.088)
	MindVoice-A:	I had been exploring my pages, wondering how we'd turned out every day with love. We nearly had died together before. (WER=65.2%, ASR-BERT=0.183)
	MindVoice:	I hit her up. Perfect story in my pages. My hero returned home and reclaimed his love. (WER=43.5%, ASR-BERT=0.527)
(13)	GT:	They took their seats within the car his driver set the horses trotting
	Vanilla:	You know what I mean. (WER=100.0%, ASR-BERT=-0.053)
	MindVoice-A:	We are the people that own the rights of this mobile force. (WER=100.0%, ASR-BERT=0.098)
	MindVoice:	Then car his driver set the horses trotting. (WER=46.2%, ASR-BERT=0.622)

Figure 7: **Additional transcription examples on EEG and MEG**.

C.6 Reconstruction Results

We provide reconstructed audio samples in the supplementary materials to complement the quantitative evaluation. These samples allow readers to directly assess the intelligibility and acoustic characteristics of the reconstructed speech.

D Additional Statement

We clarify the relationship between *MindVoice* and two closely related studies: BrainECHO (Li et al., 2025) and the acoustic–linguistic dual-pathway framework of Li et al. (2026). Our contribution builds on these methodological foundations while addressing a distinct combination of reconstruction objectives, neural-to-speech interfaces, and empirical questions.

D.1 Relationship to BrainECHO

BrainECHO (Li et al., 2025) establishes a three-stage approach to EEG/MEG-to-text decoding through vector-quantized spectrogram autoencoding, brain–audio latent alignment, and Whisper adaptation. Our linguistic stream shares this methodological principle, and we acknowledge BrainECHO as an important precedent rather than claiming the three-stage strategy itself as a new contribution.

The distinction lies in the reconstruction objective and the additional information pathway. BrainECHO uses reconstructed spectrograms as intermediate representations for text generation, whereas *MindVoice* uses recovered text to guide the reconstruction of speech waveforms together with a separately predicted acoustic-codec prompt. The transcript specifies linguistic content but does not uniquely determine its acoustic realization; the acoustic stream provides an additional route for conditioning synthesis on the neural observation. Consequently, our question extends beyond whether linguistic content can be decoded: we investigate how explicit linguistic recovery affects the fidelity of the *generated speech*, and whether neural-derived acoustic conditioning contributes beyond text-conditioned synthesis. MindVoice-L, which combines our linguistic stream with the same pretrained TTS model but omits acoustic prompting, provides an internal reference for this latter question. It should not be interpreted as a performance comparison with BrainECHO.

D.2 Relationship to the eLife dual-pathway framework

The framework of Li et al. (2026) demonstrates acoustic–linguistic integration for ECoG speech reconstruction, combining an LSTM/HiFi-GAN acoustic pathway with a Transformer-based linguistic pathway and voice cloning. We therefore do not claim that combining acoustic and linguistic pathways, or using voice cloning for neural speech reconstruction, is unprecedented.

MindVoice develops a codec-conditioned realization of this broader idea for EEG/MEG. Its acoustic stream directly predicts discrete indices in a frozen speech codec, and these predictions serve as in-context conditioning for a frozen autoregressive speech generator alongside the recovered transcript. Rather than requiring the predicted acoustic sequence to serve as the final waveform representation, the generator produces a new codec sequence conditioned on both inputs. This interface also permits temporal prompt selection and codebook-level refinement: selected neural-predicted residual codes can be incorporated into a text-generated prompt before synthesis. These operations are designed to retain useful acoustic evidence without passing every acoustic prediction unchanged into the final reconstruction. The distinction is thus not merely a change of recording modality or pretrained model, but a different mechanism for incorporating imperfect neural acoustic predictions into text-guided generation. Its effectiveness under the evaluated EEG/MEG conditions is an empirical question addressed by our experiments, rather than a consequence assumed from successful ECoG reconstruction.

D.3 Distinct contribution and interpretation of the evidence

Building on these precedents, *MindVoice* contributes a concrete non-invasive speech reconstruction framework and an empirical characterization of the interaction between linguistic recovery, acoustic conditioning, and pretrained synthesis. Across the evaluated EEG and MEG datasets, direct acoustic reconstruction can achieve closer frame-level spectral correspondence while preserving substantially

less linguistic content in the resulting waveform. Incorporating explicit linguistic recovery together with pretrained TTS substantially improves the linguistic fidelity of the generated speech, although it does not uniformly improve spectral metrics.

The single-stream comparison further identifies an asymmetric contribution: in the reported MEG ablation, text-only synthesis achieves lower WER, whereas adding the acoustic prompt improves selected spectral and learned-representation metrics. We interpret this as evidence of complementary but interacting objectives, not as proof that two streams always outperform one or that linguistic and acoustic information are fully disentangled. The resulting contribution is therefore both methodological and empirical: a codec-based conditioning mechanism for EEG/MEG speech reconstruction, together with evidence clarifying what explicit linguistic recovery and additional acoustic conditioning contribute to the final waveform. This extends the scope of prior neural-to-text and invasive dual-pathway approaches without claiming originality for their established component ideas.

E Broader Impacts

MindVoice advances intelligible speech reconstruction from non-invasive neural recordings by combining explicit linguistic recovery with complementary acoustic information. This framework may support investigations of speech-related neural representations and inform the development of speech brain-computer interfaces. Potential applications also warrant attention to neural-data privacy, informed consent, and the risk of misrepresenting participants' experiences. Because pretrained priors contribute to reconstruction, fluent outputs may contain inaccurate or unsupported content and should not be regarded as exact records of what a participant heard or intended. Responsible development should therefore prioritize secure data handling, transparent communication of reconstruction uncertainty, and participant control over the use and disclosure of reconstructed speech.

F AI Usage Statement

We used LLMs and AI agents primarily to assist with the implementation of standard methods and to refine the language of the manuscript for clarity and readability. All content, including AI-assisted code and text, was manually reviewed and verified by the authors.

G Ethics Statement

This work uses publicly available EEG and MEG datasets and does not involve new human-subject data collection. Neural speech reconstruction may raise privacy concerns as decoding capabilities improve. The reconstructed speech may also contain errors or hallucinations introduced by pretrained priors and should not be regarded as a faithful representation of participants' internal mental states. Future applications should therefore carefully consider informed consent, neural-data privacy, and responsible use.

H Reproducibility Statement

We provide detailed descriptions of the datasets, preprocessing, model implementation, training settings, baselines, and evaluation protocols in the main paper and supplementary material. Reconstructed speech samples are also included in the supplementary material. Our code will be publicly released to support reproducibility and future research.