DECODING PERCEIVED SPEECH FROM HIGH-FREQUENCY EEG PHASE PATTERNS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieving segments of continuous speech from scalp electroencephalography (EEG) identifies what a listener heard among candidate audio segments. Existing methods mainly exploit low-frequency EEG related to slower changes in speech. High-frequency EEG (> 50 Hz) contains timing cues associated with the speech fundamental frequency and rapid envelope fluctuations, but its low signal-to-noise ratio has limited its use for this task. We ask whether these signals independently support multi-candidate retrieval, complement low-frequency decoding, and retain segment-discriminative information in their phase structure. We train separate matching branches for low- and high-frequency EEG with speech features at corresponding time scales, fuse their scores, and compare waveform, phase, two-bin phase, and power representations of the same high-frequency signal. For 17 unseen listeners, the high-frequency branch attains 5.62% Top-1 accuracy in five-second retrieval from an average of more than 1,000 candidates per listener. Score fusion raises full-pool Top-1 accuracy from 8.65% for the low-frequency branch to 15.39%. Continuous phase and even a one-bit phase representation retain substantial retrieval performance after discarding amplitude scale. These results establish the complementary value of high-frequency EEG and the segment-level utility of its rapid phase-related timing structure. Code is available at `https://anonymous.4open.science/r/HFEEG-decoding-2203/`.

1 INTRODUCTION

Perceived-speech decoding uses neural signals recorded while a listener hears speech to reconstruct or identify the heard material. It connects neural activity to speech content and offers a way to study the neural processing of perceived language. Low-frequency neural phase patterns were shown to discriminate heard sentences (Luo & Poeppel, 2007), providing early evidence of a relation between neural timing and speech structure. The task also informs research toward non-invasive speech brain–computer interfaces (Défossez et al., 2023). Related neural speech-tracking analyses have been used to assess residual language processing in patients with disorders of consciousness (Gui et al., 2020), selective attention in adults with attention-deficit/hyperactivity disorder (ADHD) (Levy et al., 2025), and word-level processing in children with ADHD (Luo et al., 2023).

Scalp EEG is comparatively inexpensive and can be recorded continuously without surgery. Previous decoding work has predominantly used its low-frequency component, progressing from envelope tracking and reconstruction (Crosse et al., 2016; de Cheveigné et al., 2018) to short-window matching (Monesi et al., 2020; Accou et al., 2021a) and retrieval from multiple speech candidates. Défossez et al. (2023) aligned brain recordings with pretrained wav2vec 2.0 representations through cross-modal contrastive learning to retrieve corresponding speech segments. Wang et al. (2026) extended retrieval to acoustic, self-supervised speech, and language representations.

These tasks impose different discrimination demands. Temporal response functions predict EEG from lagged speech features; reverse reconstruction estimates an envelope from EEG, while canonical correlation analysis scores correspondence between projected EEG and speech time courses (Ding & Simon, 2012b; Crosse et al., 2016; de Cheveigné et al., 2018). Multi-candidate retrieval must instead rank the synchronized window ahead of many alternatives, making the candidate pool part of the evaluation difficulty. Representative earlier retrieval systems sample brain recordings at 100–120 Hz, with Nyquist limits of 50–60 Hz (Défossez et al., 2023; Wang et al., 2026). A sepa-

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.

rately sampled HF pathway can test whether information above that range contributes to the same retrieval task.

Speech-related neural activity is not confined to the slow fluctuations tracked by low-frequency EEG. Frequency-following response (FFR) research examines neural timing locked to periodic sound structure: both speech fundamental frequency and rapid envelope changes can evoke corresponding responses, including in scalp recordings of continuous natural speech (Forte et al., 2017; Maddox & Lee, 2018; Aiken & Picton, 2008; Canneyt et al., 2021; Bachmann et al., 2021; Coffey et al., 2016; Bidelman, 2018; Kulasingham et al., 2020; Schüller et al., 2023; Coffey et al., 2019). These rapid cues could help distinguish speech segments. Their scalp signal-to-noise ratio is low, however, and conventional measurements often average repeated short stimuli or accumulate long continuous recordings (Bachmann et al., 2021; Coffey et al., 2016; Whitham et al., 2007; Coffey et al., 2019). Recent models have achieved above-chance binary short-window EEG–speech matching from high-frequency-related inputs (Thornton et al., 2024b; Puffay et al., 2022). Whether a single, unaveraged high-frequency window can identify the heard segment among many candidates, and whether its scores improve a low-frequency system, remains to be established.

Volume conduction attenuates and spatially smooths cortical potentials at the scalp (Srinivasan et al., 1998); distributed sources may partially cancel (Ahlfors et al., 2010), and muscle activity can contaminate higher EEG frequencies (Whitham et al., 2007). These measurement properties make short-window retrieval a direct test of whether stimulus-specific HF structure remains accessible without trial averaging.

High-frequency speech responses also span distinct measurements and time scales. In continuous-speech scalp EEG, brainstem-related activity has been recovered in the 100–300-Hz range with a response delay of approximately 9 ms (Forte et al., 2017; Maddox & Lee, 2018). Magnetoencephalography studies report cortical responses near 30–35 ms for fundamental frequencies in the 70–150-Hz range; rapid higher-harmonic envelope modulation can also drive responses in that band (Kulasingham et al., 2020; Schüller et al., 2023). Intracranial studies often measure changes in high-gamma amplitude or power associated with speech events (Nourski et al., 2009; Golumbic et al., 2013; Mesgarani & Chang, 2012; Mesgarani et al., 2014; Oganian & Chang, 2019). These findings motivate a direct comparison of the information accessible from a rapidly varying scalp waveform, its phase-related timing, and its power under the same segment-retrieval objective.

We address three linked questions. Can short high-frequency EEG windows retrieve the heard natural-speech segment from a large candidate pool for listeners unseen during training? Do high-frequency scores improve the target ranking of a low-frequency system under the same queries and candidate sets? Which properties of the high-frequency waveform retain information that distinguishes segments? To answer them, we pair low-frequency EEG with frozen pretrained speech features and high-frequency EEG with auditory subband amplitudes that preserve rapid changes. We learn the two EEG–audio correspondences separately and evaluate independent retrieval and score fusion. From the same high-pass waveform, we also derive continuous phase, two-bin phase, and log power to compare the retrieval information accessible from each representation.

Our contributions are threefold. First, high-frequency EEG alone reaches 5.62% full-pool Top-1 accuracy for 17 unseen listeners, each with about 1,099 candidate five-second segments on average, versus an approximately 0.091% chance expectation. Second, fusion of independently trained branch scores increases Top-1 from 8.65% for low-frequency EEG alone to 15.39% under the identical retrieval protocol. Third, continuous phase and its one-bit, two-bin coding retain substantial performance without amplitude scale and outperform log power, implicating rapid phase-related temporal structure. Appendix A reviews the relevant literature in detail.

2 METHODS

2.1 FREQUENCY-SPECIFIC EEG–AUDIO ALIGNMENT

Figure 1 shows the retrieval framework. EEG and the actual audio played while each participant listened to natural speech were aligned by stimulus time and divided into non-overlapping five-second windows. The low-frequency (LF) branch receives 64-channel EEG filtered to 0.5–40 Hz and sampled at 100 Hz (64×500), paired with layer-9 wav2vec 2.0 features interpolated from

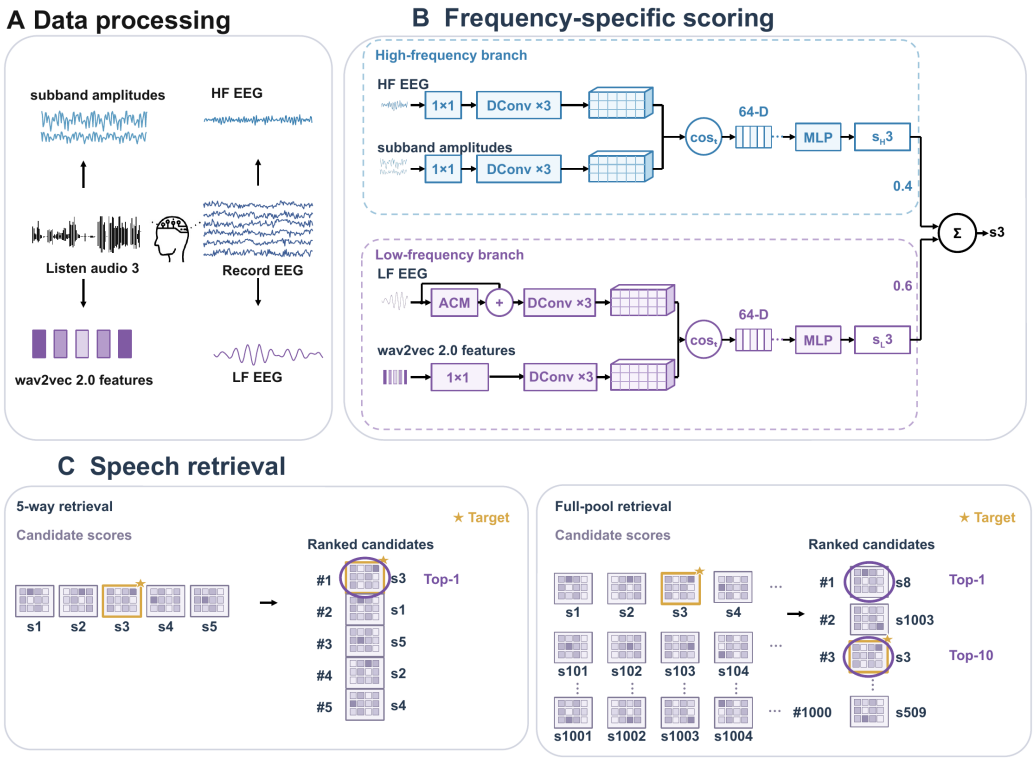

Figure 1: Two-branch EEG–audio segment retrieval for continuous speech. (A) Synchronized five-second windows pair low-frequency (LF) EEG with frozen wav2vec 2.0 layer-9 features and high-frequency (HF) EEG with 28 compressed, rectified auditory subband amplitudes. (B) Each branch uses three dilated convolution layers (DConv$\times$3); the LF EEG encoder also uses adaptive channel modulation (ACM). $\cos_t$ is channel-wise temporal cosine similarity, 64-D denotes the matching vector, and an MLP outputs a score. Independently trained scores are fused with a 0.6/0.4 weight selected on the source validation set. (C) Fused scores rank five or all candidates; the star marks the synchronized target. Candidate order is illustrative.

approximately 50 to 100 Hz (768×500). The high-frequency (HF) branch receives EEG high-pass filtered at 50 Hz and sampled at 1024 Hz (64×5120), paired with 28 auditory subband amplitude trajectories from the same audio (28×5120).

For EEG query X and candidate audio window A_j, the LF and HF branches independently output scalar match scores $s_L(j)$ and $s_H(j)$; larger scores indicate better matches. The branches are trained separately and share no parameters. The main evaluation ranks candidates using s_L, s_H, or $s_F(j) = 0.6s_L(j) + 0.4s_H(j)$. Fusion weights were chosen exclusively on the SparrKULee source validation set and fixed before testing. All three conditions use identical queries and candidate sets, so HF-only retrieval tests independent discrimination and the fusion gain tests complementarity to LF ranking.

In the LF EEG encoder, adaptive channel modulation (ACM) precedes three dilated convolution layers; a channel projection and three dilated layers encode the LF speech features. Separate projection and dilated-convolution encoders process HF EEG and subband amplitudes. Each side of each branch produces a 64-channel latent sequence. We calculate temporal cosine similarity separately for each channel and map the resulting 64-dimensional match vector to a scalar through a multilayer perceptron. Each training query contains its synchronized audio and four mismatches from the same recording. Separate five-class cross-entropy objectives train the two branches. AdamW uses a 10^{-3} learning rate, 10^{-2} weight decay, and gradient-norm cap 5; the best validation-loss checkpoint is saved during at most 100 epochs, with early stopping after ten unimproved epochs. Appendix B gives the transformations, exact scoring equations, and full implementation details.

2.2 DATASETS AND PREPROCESSING

SparrKULee (Accou et al., 2024a), containing 85 participants and 666 recordings, is the source dataset. Participants are disjoint across training ($n = 54$), validation ($n = 14$), and test ($n = 17$), producing 75,020, 17,412, and 18,680 eligible five-second queries, respectively. Participant-level splitting matters because random windows from the same continuous EEG record are temporally correlated and can leak listener identity across splits (Xu et al., 2026; Brookshire et al., 2024). Etard, Weissbart, and two self-collected Chinese-story EEG datasets are used only for external evaluation of frozen models; they do not select checkpoints or fusion weights. The two self-collected cohorts include 12 participants in total, all of whom gave written informed consent before acquisition. Continuous EEG undergoes artifact interpolation, multichannel Wiener filtering, common-average reference, and stimulus-trigger correction before filter-affected edges are removed. Each dataset is aligned to stimulus time and cut into non-overlapping five-second windows; the LF and HF pathways are sampled at 100 and 1024 Hz. Full dataset coverage, preprocessing, ethics information, and query denominators appear in Appendix B.5 and Table S1.

The source recordings cover 154.75 hours of eligible Dutch audiobook and podcast listening. The public Etard target has 20 distinct listeners and 3,477 queries from English and Dutch stories; Weissbart has 13 listeners and 5,304 queries from English stories. For these transfers, source models are retrained using the 56 named electrodes common to the source and target recordings, then frozen for target evaluation. Each self-collected cohort contains six listeners hearing the same 25 Chinese stories, with 1,493 eligible queries per listener. Its frozen HF source model is retrained using 55 common electrodes. Reporting these target-specific query and sensor denominators makes the full-pool chance rates and the scope of each frozen transfer explicit.

2.3 SPEECH REPRESENTATIONS

For LF speech input we use frozen `GroNLP/wav2vec2-dutch-base`, a wav2vec 2.0 model (Baevski et al., 2020) with twelve Transformer encoder layers and 768-dimensional hidden states. Following the layer choice of Wang et al. (2026), we extract frame-wise states from layer 9 and interpolate them to 100 Hz, yielding 768×500 features per window. The HF input is a 28-band amplitude representation of the synchronized audio. For the 48-kHz source recordings, we subtract the mean, construct gammatone-like filters equally spaced on the ERB scale from 50 to 5000 Hz, rectify and compress each subband output with a power of 0.6, and apply anti-aliased polyphase resampling to 1024 Hz. The resulting window has shape 28×5120. Audio feature-extraction parameters are fixed during EEG–audio training.

2.4 LOW-FREQUENCY ENCODER BASELINES

We adapt four representative LF EEG encoders under a common input and retrieval protocol. ConvConcatNet combines raw EEG, earlier outputs, and attention-transformed representations over six rounds of convolution and concatenation (Wang et al., 2026). The adapted BrainMagick encoder retains gated and residual temporal convolutions but disables subject conditioning (Défossez et al., 2023). VLAAI has four convolutional blocks with skip connections (Accou et al., 2023); AWaveNet uses non-causal dilated residual blocks and cross-layer links (Van Dyck et al., 2023). All models receive the same 0.5–40 Hz EEG, frozen wav2vec features, five-second query segmentation, and candidate sets. Only the EEG encoder changes: the speech encoder, temporal cosine matching, and scoring head remain shared design choices across these comparisons.

2.5 HIGH-FREQUENCY REPRESENTATIONS AND INPUT ANALYSES

From the same high-pass waveform $x_H(t)$ we construct four inputs: waveform W, continuous phase Φ, two-bin phase Q, and log power P. Let $z(t) = x_H(t) + i\mathcal{H}\{x_H(t)\}$ be the analytic signal, with Hilbert transform $\mathcal{H}$ and phase $\phi(t) = \arg z(t)$. The continuous-phase input concatenates $\cos\phi$ and $\sin\phi$, doubling the channels per electrode. The two-bin representation divides phase into half-cycle bins and encodes

$$q_2(t) = 2\,\mathbf{1}\{\cos\phi(t) \geq 0\} - 1, \qquad P(t) = \log\big(|z(t)|^2 + \epsilon\big). \tag{1}$$

Here ϵ is 10^{-12} times the median, across channels, of average training-record power. Within each window, Φ and P are standardized channel-wise; Q remains ± 1. A separate HF scoring model

is trained for each input. When $|z(t)| > 0$, $\cos\phi(t) = x_H(t)/|z(t)|$, so Q is mathematically equivalent to the high-pass waveform's sign. Its interpretation here is two-bin quantization of phase. The comparison measures information accessible under this modeling protocol; the main fusion experiment uses the raw waveform W branch.

We also examine input timing, frequency, and electrodes. For timing, a fixed random EEG–stimulus offset is sampled per continuous recording/trial and a new model is trained in each condition. HF Q uses nominal offsets of ± 3, ± 10, and ± 30 ms, rounded to samples on the 1024-Hz grid. LF waveform controls use ± 10 and ± 30 ms on the 100-Hz grid; ± 3 ms is omitted because it rounds to zero there. For frequency, we split 50–500 Hz into nine adjacent 50-Hz bands, extract Q for each band, and train separate models. For spatial input, we apply common-average reference to all 64 electrodes, retain one prespecified six-electrode frontal, central, temporal, parietal, or occipital group, zero the remaining positions, and train each Q model separately. This compares sensor subsets after full-array referencing, not a physical six-electrode system or source localization. All conditions retain the primary participant split and query construction. They assess what can be learned from each training condition, rather than test-time perturbations of one frozen model.

2.6 Retrieval and statistical analysis

Each five-second EEG window X_i queries its synchronized audio window A_i. In five-way retrieval, four distinct-time audio windows from the same recording serve as mismatches; target rank 1 is Top-1 and nominal chance is 20%. In full-pool retrieval, all valid audio windows for that test participant are candidates; we report the proportion of queries ranking the target first or within the top ten. LF, HF, and fusion scores are evaluated on the same queries and candidate sets. Accuracies are counted per query and pooled over test queries. In the SparrKULee main test, the mean full pool has approximately 1,099 candidates per participant, yielding an approximately 0.091% query-pooled random Top-1 expectation.

The independent unit for statistical analysis is the test participant. Paired conditions are compared by two-sided paired Student t tests; omnibus comparisons across the four representations, nine bands, and five electrode groups use Friedman repeated-measures tests. For chance comparisons, each participant's accuracy is differenced from that participant's corresponding random rate and tested by a two-sided one-sample t test. Chance is 20% for five-way retrieval, $1/K$ for full-pool Top-1, and $\min(10, K)/K$ for Top-10 at query-specific pool size K. We report Benjamini–Hochberg (BH) adjusted q values and participant-equal paired mean differences with 95% confidence intervals. An outcome is described as above chance only when its difference is positive and $q < 0.05$. The six chance tests spanning three metrics in each of the two self-collected datasets are adjusted together.

3 Results and Discussion

3.1 High-frequency EEG supports large-pool retrieval and improves low-frequency ranking

On 17 SparrKULee test participants unseen in training, the HF branch reaches 56.15% five-way Top-1, 5.62% full-pool Top-1, and 16.69% full-pool Top-10 (Figure 2). The average full pool contains about 1,099 candidates; the query-pooled random Top-1 expectation is about 0.091%. At the participant level, HF full-pool Top-1 exceeds each participant's chance rate by 5.79 percentage points on average (95% CI [3.08, 8.50], BH-adjusted $q = 0.0007$). Thus, a separately learned HF score identifies synchronized speech within a large pool.

The LF branch achieves 69.55%, 8.65%, and 27.43% on the same three metrics, numerically exceeding the four adapted reference encoders under matched inputs and retrieval (Figure 2). Against ConvConcatNet, its participant-paired full-pool Top-1 difference is 0.84 percentage points (95% CI [0.24, 1.44], $q = 0.0089$). Source-validation-weighted fusion reaches 76.88%, 15.39%, and 38.43%, respectively. Relative to LF alone, the participant-paired full-pool Top-1 gain is 6.68 points (95% CI [3.68, 9.68], $q = 0.0003$); relative to HF alone it is 9.30 points (95% CI [4.40, 14.19], $q = 0.0010$). The pooled accuracies count all queries, whereas paired mean differences give equal weight to participants. The HF-only result and fusion improvement together show complementarity between the two complete scoring systems.

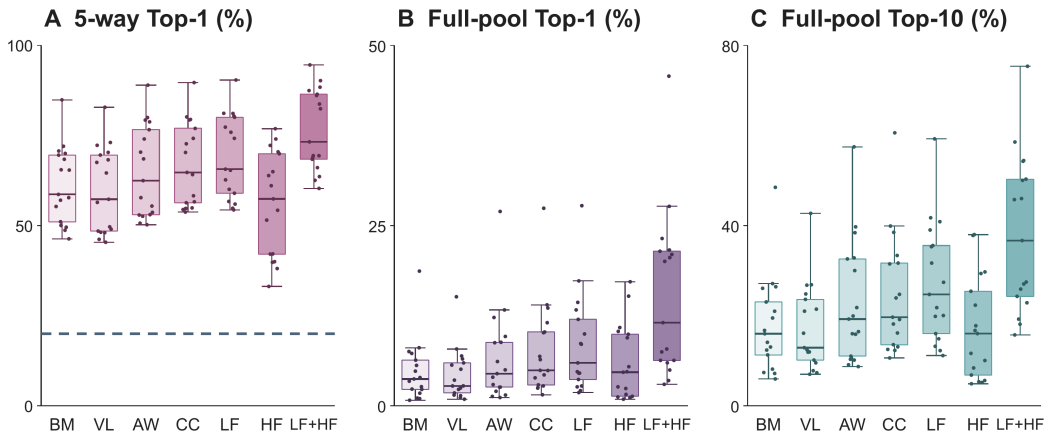

Figure 2: Segment retrieval for 17 unseen listeners: (A) five-way Top-1, (B) full-pool Top-1, and (C) full-pool Top-10 (all in percent). BM, VL, AW, and CC denote BrainMagick, VLAAI, AWaveNet, and ConvConcatNet. LF and HF are the primary low- and high-frequency branches; LF+HF is validation-weighted score fusion. The dashed line in A marks 20% chance. Each dark point is one participant; box centers are medians, boxes span the interquartile range, and whiskers extend to the most distant observations within 1.5 interquartile ranges. Query-pooled accuracies in the text need not equal boxplot medians.

The full-pool task requires the synchronized segment to outrank every other eligible audio window for that listener, rather than four sampled mismatches. Its Top-10 result also tests whether the target moves into the small group of highest-ranked candidates: fusion raises that measure from 27.43% to 38.43% on the same queries. Because the branches are optimized separately and combined only at the score level, this gain does not depend on a jointly retrained encoder. The fusion weight is fixed using source validation data; the identical weight is carried into the public-dataset transfer tests below.

3.2 Phase and two-bin phase retain HF retrieval information

We next compare amplitude variation and rapid timing structure in the HF waveform. Separately trained models receive raw waveform W, continuous phase Φ, two-bin phase Q, or log power P (Figure 3). Their full-pool Top-1 accuracies are 5.62%, 4.89%, 4.34%, and 0.27%, respectively. Five-way Top-1 is 56.15%, 54.96%, 52.12%, and 27.48%; full-pool Top-10 is 16.69%, 15.51%, 13.47%, and 2.16%. Phase and even its one-bit, two-bin form retain substantial discrimination after amplitude scale is discarded, whereas log power alone yields much lower retrieval performance.

All three metrics have the same numerical ordering, with W, Φ, and Q exceeding P. Participant-paired full-pool Top-1 differences versus P are 5.59, 4.89, and 4.42 percentage points, respectively (BH-adjusted $q = 0.0009$, 0.0010, and 0.0025). Under these independently trained representations, phase-related timing structure remains useful when amplitude scale is removed.

This comparison probes a different signal property from high-gamma power-envelope tracking. Intracranial high-gamma amplitude and power have been related to auditory objects and speech features (Nourski et al., 2009; Golumbic et al., 2013; Mesgarani & Chang, 2012; Mesgarani et al., 2014; Oganian & Chang, 2019), and slow changes in scalp high-gamma power can track the speech envelope (Synigal et al., 2020). Here, Φ and Q retain within-window phase timing while discarding amplitude scale; P preserves power variation. The result therefore establishes that the trained retrieval system can access substantially more segment information from the phase-derived inputs than from log power under matched candidate evaluation. Because Q is also the sign of the high-pass waveform wherever analytic amplitude is nonzero, the result does not by itself identify a unique neural phase-coding mechanism or source.

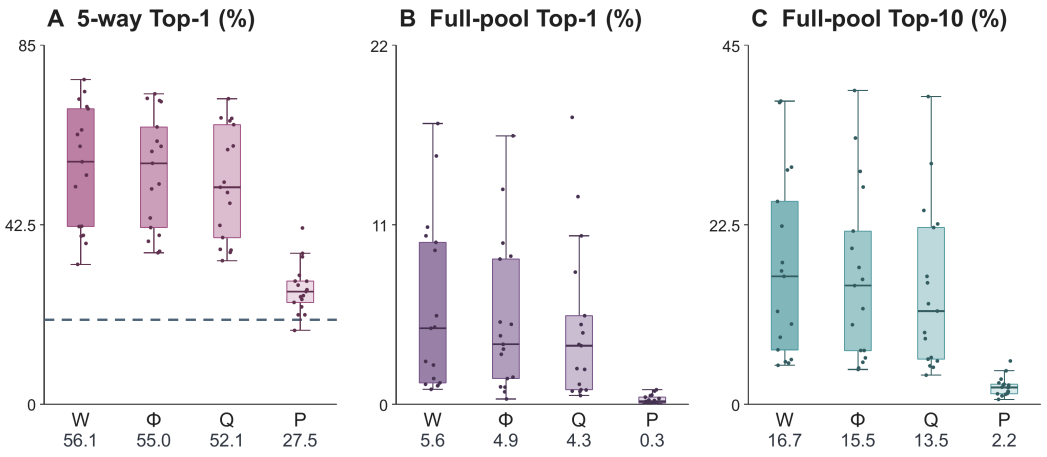

Figure 3: Retrieval with independently trained HF EEG input representations. W: raw high-pass waveform above approximately 50 Hz; Φ: sine–cosine continuous phase; Q: two-bin phase; P: log power. Panels show (A) five-way Top-1, (B) full-pool Top-1, and (C) full-pool Top-10. The dashed line in A is 20% chance. Boxplot conventions follow Figure 2; numbers beneath conditions are accuracies pooled across all test queries, rather than boxplot medians.

3.3 Timing, frequency, and sensor analyses

3.3.1 Timing

If rapid temporal structure contributes to retrieval, millisecond-scale EEG–audio alignment should affect the information accessible to the model. We therefore retrain a two-bin-phase model for each training-time offset condition (Figure 4A). With no offset and with fixed per-recording/trial offsets sampled from ± 3, ± 10, and ± 30 ms, full-pool Top-1 is 4.34%, 2.69%, 1.39%, and 0.13%, respectively. Relative to zero offset, participant-paired changes at ± 3 and ± 10 ms are -1.68 percentage points (95% CI $[-2.68, -0.68]$, $q = 0.0031$) and -3.10 points ($[-4.41, -1.78]$, $q = 0.0002$). At ± 30 ms, full-pool Top-1 no longer differs significantly from chance ($q = 0.4643$). The LF waveform control retains 8.65%, 7.51%, and 5.06% full-pool Top-1 at zero, ± 10, and ± 30 ms, showing a more gradual numerical decrease over the corresponding nominal offsets. All conditions use independent training; the full three-metric distributions appear in Figure S1.

Offsets are sampled independently for each continuous recording or trial and held fixed within it, so the model is trained with differing EEG–audio alignments across records rather than one global delay. At the largest tested offset, HF two-bin-phase Top-1 approaches the approximately 0.091% random expectation, whereas the LF waveform retains 5.06% under its corresponding offset condition. The contrast links the rapidly varying HF input to the precision of the EEG–stimulus time axis and informs the external-recording comparison below.

3.3.2 Frequency

Nine independently trained 50-Hz bands isolate where the HF two-bin-phase representation supports retrieval (Figure 4B). Full-pool Top-1 is 2.45% at 50–100 Hz, 1.19% at 100–150 Hz, and 0.51% at 150–200 Hz. The 50–100-Hz band is strongest among those tested, significantly exceeding each of the other eight in participant-paired comparisons (within-family adjusted $q \leq 0.0417$). Against participant-specific chance, the 50–100, 100–150, 150–200, and 250–300-Hz bands remain significant (BH-adjusted $q = 0.0060$, 0.0020, 0.0085, and 0.0109). Thus, retrievable phase-bin cues concentrate in the lowest tested HF band but extend to selected higher frequencies. Figure S2 shows five-way Top-1 and full-pool Top-10 as well.

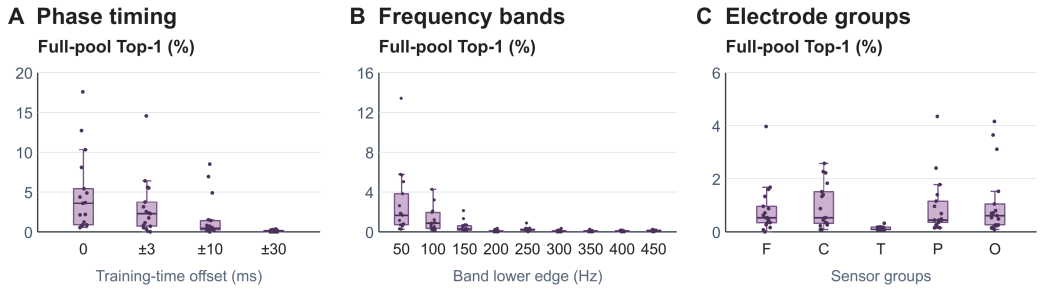

Figure 4: HF two-bin-phase retrieval under different input conditions, shown as full-pool Top-1 (%). (A) Training-time offsets; (B) nine adjacent 50-Hz bands, labeled by lower band edge; (C) six-electrode frontal (F), central (C), temporal (T), parietal (P), and occipital (O) groups. Each condition is trained independently. Dark points are unseen participants ($n = 17$); boxplot conventions follow Figure 2. LF timing controls, all three metrics, and electrode positions are in Appendix C.

3.3.3 SENSORS

After common-average referencing all 64 electrodes, we independently retrain HF Q models using five prespecified six-electrode subsets (Figure 4C). Frontal, central, temporal, parietal, and occipital full-pool Top-1 scores are 0.77%, 0.96%, 0.12%, 0.83%, and 0.99%, respectively. Each exceeds its participant-specific chance rate (BH-adjusted $q = 0.0059$, 0.0008, 0.0170, 0.0062, and 0.0071). The temporal subset yields the lowest score. These are sensor comparisons after a reference computed from the full array, not source localization or performance expected from a physically six-electrode recording. Figure S3 provides electrode positions, LF controls, and all three metrics.

3.4 CROSS-DATASET RETRIEVAL AND SCORE COMPLEMENTARITY

To assess independent continuous-speech recordings, we retrain LF and HF models on the 56 electrodes shared by SparrKULee and two public datasets, freeze them, and evaluate Etard and Weissbart (Figure 5A). On Etard, LF, HF, and fusion five-way Top-1 are 56.77%, 36.01%, and 60.80%. On Weissbart, they are 52.34%, 36.82%, and 57.11%. HF supplies above-chance retrieval in both target datasets, and fixed-weight fusion improves LF ranking.

Full-pool chance levels differ across these four datasets, so Figure 5 uses five-way Top-1 for a common 20% reference; full-pool distributions are shown in Figure S4. HF full-pool Top-1 is 3.68% on Etard (chance 0.575%, $q = 0.0002$) and 1.58% on Weissbart (chance 0.245%, $q = 0.0112$), each above chance. On Etard, LF and fusion full-pool Top-1 are 8.89% and 12.08%, a participant-paired gain of 3.04 points (95% CI [1.43, 4.65], $q = 0.0009$). On Weissbart, they are 4.51% and 6.26%, a paired gain of 1.75 points (95% CI [0.14, 3.37], $q = 0.0357$).

The retrieval setting also differs from earlier high-frequency EEG decoding benchmarks. Puffay et al. (2022) evaluated five-second fundamental-frequency-related binary EEG–speech matching on two male-narrated stories, obtaining above-chance accuracy; Thornton et al. (2024b) reported 65.64% accuracy for three-second binary matching from 70–220-Hz EEG with a 100-model ensemble. Such tasks choose between one target and one mismatch. By contrast, our principal test ranks roughly 1,099 candidate windows per unseen listener, and LF, HF, and fusion share each query and its full candidate pool. Previous multi-candidate systems based on lower-rate brain recordings also differ in candidate counts and representation design (Défossez et al., 2023; Wang et al., 2026); their reported Top-10 values are therefore context rather than matched-condition performance baselines. The adapted LF encoder comparisons in Figure 2 provide the within-protocol reference for this study.

3.5 SHARED STIMULI WITH DIFFERENT TRIGGER RECORDS

The two self-collected datasets use the same 25 Chinese stories but different EEG–audio trigger records. Hardware-EEG logs an event at audio playback, whereas Matlab-EEG aligns via MATLAB

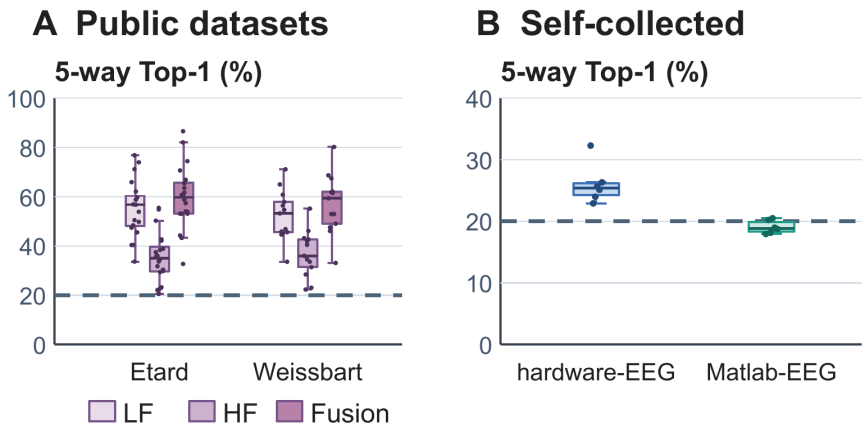

Figure 5: Five-way Top-1 (%) on four external EEG datasets. (A) Public Etard ($n = 20$) and Weissbart ($n = 13$) datasets show LF, HF, and fixed-weight fusion. (B) Two self-collected Chinese-story datasets each include six participants hearing the same 25 stories; hardware-EEG uses playback-time hardware triggers, while Matlab-EEG uses MATLAB playback-event records. Both evaluate a frozen HF model. Dark points denote participants, boxes and whiskers follow Figure 2, and the dashed line marks 20% chance. The two panels have different vertical scales. Full-pool results appear in Appendix C.

playback-event records (Herrmann, 2025). Each independent cohort contains six participants, each contributing 25 trials and 1,493 five-second queries. Both are evaluated at their supplied common time origin, without an additional test-time EEG shift, using an HF model retrained on 55 shared source electrodes and then frozen (Figure 5B).

Hardware-EEG obtains 26.04% five-way Top-1, 6.04 percentage points above the 20% chance rate ($t(5) = 4.50$, 95% CI [2.59, 9.50], BH-adjusted $q = 0.0302$); Matlab-EEG obtains 19.09% ($q = 0.0884$). Hardware-EEG also exceeds chance in full-pool Top-1 and Top-10, reaching 0.368% and 2.121% ($q = 0.0302$ and 0.0375). Matlab-EEG reaches 0.022% and 0.558%, neither above chance; its Top-1 difference from chance is negative ($q = 0.0375$), and its Top-10 chance test gives $q = 0.0654$. Participants are the independent statistical units, and the six chance tests across the two datasets and three metrics are BH-adjusted together. Full distributions appear in Figure S5.

With shared story stimuli, query construction, and frozen evaluation, the contrast emphasizes the EEG–audio synchronization chain. It accords with the timing-offset analysis: access to HF temporal cues depends on alignment quality. The cohorts contain different participants, however, and the trigger-to-sound-output latency and jitter have not been measured. Thus the observed difference cannot be attributed wholly to trigger type; that explanation requires a within-participant, same-device controlled test. As larger multimodal EEG resources emerge (Chen et al., 2026), acquisition protocols should anchor EEG to the actual audio output and quantify fixed delay, trial-to-trial jitter, and long-record drift through audio loopback or equivalent checks. A hardware event alone does not establish its timing precision relative to sound.

4 Conclusion

High-frequency scalp EEG independently supports continuous-speech segment retrieval for unseen listeners and contributes information complementary to a low-frequency scoring system. On 17 SparrKULee test listeners, HF EEG achieves 5.62% Top-1 from about 1,099 candidates per listener; score fusion increases full-pool Top-1 from 8.65% to 15.39%. Frozen tests on two public datasets reproduce HF retrievability and fusion gains. Continuous and two-bin phase retain substantial retrieval ability without amplitude scale and outperform log power. Timing, band, and sensor analyses map the input conditions for these cues. The two datasets using identical stories and the controlled timing offsets identify verified EEG–audio synchronization as a priority for future high-frequency speech-decoding recordings.

AI USE STATEMENT

Generative AI tools assisted with organizing literature leads, drafting and restructuring text, writing code, translating this manuscript into English, and preparing its LaTeX format. These tools did not participate in experimental design or the formulation of the research hypotheses. The authors are responsible for the study design, EEG data collection, experiment execution, result verification, bibliographic accuracy, and final argument.

ETHICS STATEMENT

The collection and use of the two self-collected EEG datasets were approved by the relevant institutional ethics committee. All 12 participants provided written informed consent before data collection. The committee name and approval identifier are withheld in this double-blind version and can be supplied after review; participant-level results use anonymous identifiers.

REPRODUCIBILITY STATEMENT

Appendix B describes the network and scoring equations, training configuration, data processing, and query counts. Appendix C provides extended result distributions and statistical tests. An anonymized code repository is linked in the abstract.

REFERENCES

Bernd Accou, Mohammad Jalilpour Monesi, Jair Montoya, Hugo Van hamme, and Tom Francart. Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network, 2021a. 2020 28th European Signal Processing Conference (EUSIPCO), pp. 1175–1179. DOI: https://doi.org/10.23919/EUSIPCO47968.2020.9287417.

Bernd Accou, Jonas Vanthornhout, Hugo Van hamme, and Tom Francart. Decoding of the speech envelope from EEG using the VLAAI deep neural network, 2023. Scientific Reports, 13:812. DOI: https://doi.org/10.1038/s41598-022-27332-2.

Bernd Accou, Lies Bollens, Marlies Gillis, Wendy Verheijen, Hugo Van hamme, and Tom Francart. SparrKULee: A Speech-Evoked Auditory Response Repository from KU Leuven, Containing the EEG of 85 Participants, 2024a. Data, 9(8):94. DOI: https://doi.org/10.3390/data9080094.

Bernd Accou, Hugo Van hamme, and Tom Francart. CLASH: Contrastive learning through alignment shifting to extract stimulus information from EEG, 2024b. arXiv:2302.01924v2. DOI: https://doi.org/10.48550/arXiv.2302.01924.

Bernd Accou et al. Predicting speech intelligibility from EEG in a non-linear classification paradigm, 2021b. Journal of Neural Engineering, 18(6):066008. DOI: https://doi.org/10.1088/1741-2552/ac33e9.

Seppo P. Ahlfors, Jooman Han, Fa-Hsuan Lin, Thomas Witzel, John W. Belliveau, Matti S. Hämäläinen, and Eric Halgren. Cancellation of EEG and MEG signals generated by extended and distributed sources, 2010. Human Brain Mapping, 31(1):140–149. DOI: https://doi.org/10.1002/hbm.20851.

Steven J. Aiken and Terence W. Picton. Envelope and spectral frequency-following responses to vowel sounds, 2008. Hearing Research, 245:35–47. DOI: https://doi.org/10.1016/j.heares.2008.08.004.

M. Florencia Assaneo and David Poeppel. The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm, 2018. Science Advances, 4:eaao3842. DOI: https://doi.org/10.1126/sciadv.aao3842.

Florine L. Bachmann, Ewen N. MacDonald, and Jens Hjortkjær. Neural Measures of Pitch Processing in EEG Responses to Running Speech, 2021. Frontiers in Neuroscience, 15:738408. DOI: https://doi.org/10.3389/fnins.2021.738408.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations, 2020. Advances in Neural Information Processing Systems, 33:12449–12460.

Gavin M. Bidelman. Subcortical sources dominate the neuroelectric auditory frequency-following response to speech, 2018. NeuroImage, 175:56–69. DOI: https://doi.org/10.1016/j.neuroimage.2018.03.060.

Lies Bollens et al. Contrastive representation learning with transformers for robust auditory EEG decoding, 2025. Scientific Reports, 15:28744. DOI: https://doi.org/10.1038/s41598-025-13646-4.

Geoffrey Brookshire, Jake Kasper, Nicholas M. Blauch, Yunan Charles Wu, Ryan Glatt, David A. Merrill, Spencer Gerrol, Keith J. Yoder, Colin Quirk, and Ché Lucero. Data leakage in deep learning studies of translational EEG, 2024. Frontiers in Neuroscience, 18:1373515. DOI: https://doi.org/10.3389/fnins.2024.1373515.

Jana Van Canneyt, Jan Wouters, and Tom Francart. Neural tracking of the fundamental frequency of the voice: The effect of voice characteristics, 2021. European Journal of Neuroscience, 53(11):3640–3653. DOI: https://doi.org/10.1111/ejn.15229.

Sitong Chen, Beiqianyi Li, Cuilin He, Dongyang Li, Mingyang Wu, Xinke Shen, Song Wang, Xuetao Wei, Xindi Wang, Haiyan Wu, and Quanying Liu. An EEG Dataset for Multimodal Semantic Alignment and Neural Decoding during Reading and Listening, 2026. Scientific Data, 13:148. DOI: https://doi.org/10.1038/s41597-025-06466-8.

Emily B. J. Coffey, Sibylle C. Herholz, Alexander M. P. Chepesiuk, Sylvain Baillet, and Robert J. Zatorre. Cortical contributions to the auditory frequency-following response revealed by MEG, 2016. Nature Communications, 7:11070. DOI: https://doi.org/10.1038/ncomms11070.

Emily B. J. Coffey, Trent Nicol, Travis White-Schwoch, Bharath Chandrasekaran, Jennifer Krizman, Erika Skoe, Robert J. Zatorre, and Nina Kraus. Evolving perspectives on the sources of the frequency-following response, 2019. Nature Communications, 10:5036. DOI: https://doi.org/10.1038/s41467-019-13003-w.

Michael J. Crosse, Giovanni M. Di Liberto, Adam Bednar, and Edmund C. Lalor. The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli, 2016. Frontiers in Human Neuroscience 10:604. DOI: https://doi.org/10.3389/fnhum.2016.00604.

Alain de Cheveigné, Daniel D. E. Wong, Giovanni M. Di Liberto, Jens Hjortkjær, Malcolm Slaney, and Edmund Lalor. Decoding the auditory brain with canonical component analysis, 2018. NeuroImage, 172:206–216. DOI: https://doi.org/10.1016/j.neuroimage.2018.01.033.

Nai Ding and Jonathan Z. Simon. Emergence of neural encoding of auditory objects while listening to competing speakers, 2012a. Proceedings of the National Academy of Sciences, 109(29):11854–11859. DOI: https://doi.org/10.1073/pnas.1205381109.

Nai Ding and Jonathan Z. Simon. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening, 2012b. Journal of Neurophysiology, 107:78–89. DOI: https://doi.org/10.1152/jn.00297.2011.

Nai Ding, Lucia Melloni, Hang Zhang, Xing Tian, and David Poeppel. Cortical tracking of hierarchical linguistic structures in connected speech, 2016. Nature Neuroscience, 19(1):158–164. DOI: https://doi.org/10.1038/nn.4186.

Keith B. Doelling, Luc H. Arnal, Oded Ghitza, and David Poeppel. Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing, 2014. NeuroImage, 85 (Part 2):761–768. DOI: https://doi.org/10.1016/j.neuroimage.2013.06.035.

Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings, 2023. Nature Machine Intelligence 5:1097–1107. DOI: https://doi.org/10.1038/s42256-023-00714-5.

O. Etard and T. Reichenbach. Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise, 2019. Journal of Neuroscience, 39(29):5750–5759. DOI: https://doi.org/10.1523/JNEUROSCI.1828-18.2019.

O. Etard and T. Reichenbach. EEG Dataset for 'Decoding of selective attention to continuous speech from the human auditory brainstem response' and 'Neural Speech Tracking in the Theta and in the Delta Frequency Band Differentially Encode Clarity and Comprehension of Speech in Noise', 2022. Zenodo, version 2.0.0. DOI: https://doi.org/10.5281/zenodo.7778289.

O. Etard, M. Kegler, C. Braiman, A. E. Forte, and T. Reichenbach. Decoding of selective attention to continuous speech from the human auditory brainstem response, 2019. NeuroImage, 200:1–11. DOI: https://doi.org/10.1016/j.neuroimage.2019.06.029.

Antonio Elia Forte, Octave Etard, and Tobias Reichenbach. The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention, 2017. eLife, 6:e27203. DOI: https://doi.org/10.7554/eLife.27203.

Zhen Fu, Xihong Wu, and Jing Chen. Using frequency-following responses (FFRs) to evaluate the auditory function of frequency-modulation (FM) discrimination, 2017. Applied Informatics, 4:10. DOI: https://doi.org/10.1186/s40535-017-0040-7.

Zhen Fu, Hongying Yang, Fei Chen, Xihong Wu, and Jing Chen. Brainstem encoding of frequency-modulated sweeps is relevant to Mandarin concurrent-vowels identification for normal-hearing and hearing-impaired listeners, 2019. Hearing Research, 380:123–136. DOI: https://doi.org/10.1016/j.heares.2019.06.005.

Elana M. Zion Golumbic et al. Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a 'Cocktail Party', 2013. Neuron, 77:980–991. DOI: https://doi.org/10.1016/j.neuron.2012.12.037.

Joachim Gross, Nienke Hoogenboom, Gregor Thut, Philippe Schyns, Stefano Panzeri, Pascal Belin, and Simon Garrod. Speech Rhythms and Multiplexed Oscillatory Sensory Coding in the Human Brain, 2013. PLOS Biology, 11(12):e1001752. DOI: https://doi.org/10.1371/journal.pbio.1001752.

P. Gui, Y. Jiang, D. Zang, et al. Assessing the depth of language processing in patients with disorders of consciousness, 2020. Nature Neuroscience, 23:761–770. DOI: https://doi.org/10.1038/s41593-020-0639-1.

Björn Herrmann. Enhanced neural speech tracking through noise indicates stochastic resonance in humans, 2025. eLife, 13:RP100830. DOI: https://doi.org/10.7554/eLife.100830.

Mary F. Howard and David Poeppel. Discrimination of Speech Stimuli Based on Neuronal Response Phase Patterns Depends on Acoustics But Not Comprehension, 2010. Journal of Neurophysiology, 104:2500–2511. DOI: https://doi.org/10.1152/jn.00251.2010.

Mary F. Howard and David Poeppel. The neuromagnetic response to spoken sentences: Comodulation of theta band amplitude and phase, 2012. NeuroImage, 60:2118–2127. DOI: https://doi.org/10.1016/j.neuroimage.2012.02.028.

Christoph Kayser, Marcelo A. Montemurro, Nikos K. Logothetis, and Stefano Panzeri. Spike-Phase Coding Boosts and Stabilizes Information Carried by Spatial and Temporal Spike Patterns, 2009. Neuron, 61(4):597–608. DOI: https://doi.org/10.1016/j.neuron.2009.01.008.

Christoph Kayser, Caroline Wilson, Houman Safaai, Shuzo Sakata, and Stefano Panzeri. Rhythmic Auditory Cortex Activity at Multiple Timescales Shapes Stimulus–Response Gain and Background Firing, 2015. Journal of Neuroscience, 35(20):7750–7762. DOI: https://doi.org/10.1523/JNEUROSCI.0268-15.2015.

Joshua P. Kulasingham, Christian Brodbeck, Alessandro Presacco, Stefanie E. Kuchinsky, Samira Anderson, and Jonathan Z. Simon. High gamma cortical processing of continuous speech in younger and older listeners, 2020. NeuroImage, 222:117291. DOI: https://doi.org/10.1016/j.neuroimage.2020.117291.

O. Levy, S. Libman Hackmon, Y. Zvilichovsky, A. Korisky, A. Bidet-Caulet, J. B. Schweitzer, and E. Zion Golumbic. Selective attention and sensitivity to auditory disturbances in a virtually real classroom, 2025. eLife, 13:RP103235, Version of Record. DOI: https://doi.org/10.7554/eLife.103235.3.

C. Luo, Y. Gao, J. Fan, Y. Liu, Y. Yu, and X. Zhang. Compromised word-level neural tracking in the high-gamma band for children with attention deficit hyperactivity disorder, 2023. Frontiers in Human Neuroscience, 17:1174720. DOI: https://doi.org/10.3389/fnhum.2023.1174720.

Huan Luo and David Poeppel. Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex, 2007. Neuron, 54:1001–1010. DOI: https://doi.org/10.1016/j.neuron.2007.06.004.

Ross K. Maddox and Adrian K. C. Lee. Auditory Brainstem Responses to Continuous Natural Speech in Human Listeners, 2018. eNeuro, 5:ENEURO.0441-17.2018. DOI: https://doi.org/10.1523/ENEURO.0441-17.2018.

Nima Mesgarani and Edward F. Chang. Selective cortical representation of attended speaker in multi-talker speech perception, 2012. Nature, 485:233–236. DOI: https://doi.org/10.1038/nature11020.

Nima Mesgarani, Connie Cheung, Keith Johnson, and Edward F. Chang. Phonetic Feature Encoding in Human Superior Temporal Gyrus, 2014. Science, 343:1006–1010. DOI: https://doi.org/10.1126/science.1245994.

Mohammad Jalilpour Monesi, Bernd Accou, Jair Montoya-Martinez, Tom Francart, and Hugo Van hamme. An LSTM Based Architecture to Relate Speech Stimulus to EEG, 2020. ICASSP, pp. 941–945. DOI: https://doi.org/10.1109/ICASSP40776.2020.9054000.

Kirill V. Nourski, Richard A. Reale, Hiroyuki Oya, Hiroto Kawasaki, Christopher K. Kovach, Haiming Chen, Matthew A. Howard III, and John F. Brugge. Temporal Envelope of Time-Compressed Speech Represented in the Human Auditory Cortex, 2009. Journal of Neuroscience, 29(49):15564–15574. DOI: https://doi.org/10.1523/JNEUROSCI.3065-09.2009.

Yulia Oganian and Edward F. Chang. A speech envelope landmark for syllable encoding in human superior temporal gyrus, 2019. Science Advances, 5:eaay6279. DOI: https://doi.org/10.1126/sciadv.aay6279.

Yulia Oganian, Katsuaki Kojima, Assaf Breska, Chang Cai, Anne Findlay, Edward F. Chang, and Srikantan S. Nagarajan. Phase Alignment of Low-Frequency Neural Activity to the Amplitude Envelope of Speech Reflects Evoked Responses to Acoustic Edges, Not Oscillatory Entrainment, 2023. Journal of Neuroscience, 43:3909–3921. DOI: https://doi.org/10.1523/JNEUROSCI.1663-22.2023.

Hyojin Park, Robin A. A. Ince, Philippe G. Schyns, Gregor Thut, and Joachim Gross. Frontal Top-Down Signals Increase Coupling of Auditory Low-Frequency Oscillations to Continuous Speech in Human Listeners, 2015. Current Biology, 25(12):1649–1653. DOI: https://doi.org/10.1016/j.cub.2015.04.049.

Jonathan E. Peelle, Joachim Gross, and Matthew H. Davis. Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension, 2013. Cerebral Cortex, 23(6):1378–1387. DOI: https://doi.org/10.1093/cercor/bhs118.

Corentin Puffay, Jana Van Canneyt, Jonas Vanthornhout, Hugo Van hamme, and Tom Francart. Relating the fundamental frequency of speech with EEG using a dilated convolutional network, 2022. Interspeech 2022, pp. 4038–4042. DOI: https://doi.org/10.21437/INTERSPEECH.2022-315.

Lars Riecke, Elia Formisano, Bettina Sorger, Deniz Başkent, and Etienne Gaudrain. Neural Entrainment to Speech Modulates Speech Intelligibility, 2018. Current Biology, 28(2):161–169.e5. DOI: https://doi.org/10.1016/j.cub.2017.11.033.

Alina Schüller, Achim Schilling, Patrick Krauss, Stefan Rampp, and Tobias Reichenbach. Attentional Modulation of the Cortical Contribution to the Frequency-Following Response Evoked by Continuous Speech, 2023. Journal of Neuroscience, 43(44):7429–7440. DOI: https://doi.org/10.1523/JNEUROSCI.1247-23.2023.

Ramesh Srinivasan, Paul L. Nunez, and Richard B. Silberstein. Spatial filtering and neocortical dynamics: estimates of EEG coherence, 1998. IEEE Transactions on Biomedical Engineering, 45(7):814–826. DOI: https://doi.org/10.1109/10.686789.

Shyanthony R. Synigal, Emily S. Teoh, and Edmund C. Lalor. Including Measures of High Gamma Power Can Improve the Decoding of Natural Speech From EEG, 2020. Frontiers in Human Neuroscience, 14:130. DOI: https://doi.org/10.3389/fnhum.2020.00130.

Francois D. Szymanski, Neil C. Rabinowitz, Cesare Magri, Stefano Panzeri, and Jan W. H. Schnupp. The Laminar and Temporal Structure of Stimulus Information in the Phase of Field Potentials of Auditory Cortex, 2011. Journal of Neuroscience, 31(44):15787–15801. DOI: https://doi.org/10.1523/JNEUROSCI.1416-11.2011.

Mike Thornton, Jonas Auernheimer, Constantin Jehn, Danilo Mandic, and Tobias Reichenbach. Detecting Gamma-Band Responses to the Speech Envelope for the ICASSP 2024 Auditory EEG Decoding Signal Processing Grand Challenge, 2024a. IEEE ICASSP Workshops, 55–56. DOI: https://doi.org/10.1109/ICASSPW62465.2024.10626244.

Mike D. Thornton, Danilo P. Mandic, and Tobias J. Reichenbach. Decoding Envelope and Frequency-Following EEG Responses to Continuous Speech Using Deep Neural Networks, 2024b. IEEE Open Journal of Signal Processing, 5:700–716. DOI: https://doi.org/10.1109/OJSP.2024.3378593.

B. M. S. L. Van Dyck, L. Yang, and M. M. Van Hulle. Decoding Auditory EEG Responses Using an Adapted Wavenet, 2023. ICASSP, pp. 1–2. DOI: https://doi.org/10.1109/ICASSP49357.2023.10095420.

Jonas Vanthornhout, Lien Decruy, Jan Wouters, Jonathan Z. Simon, and Tom Francart. Speech Intelligibility Predicted from Neural Entrainment of the Speech Envelope, 2018. Journal of the Association for Research in Otolaryngology, 19(2):181–191. DOI: https://doi.org/10.1007/s10162-018-0654-z.

Bo Wang, Xiran Xu, Boda Xiao, Linze Zheng, Xihong Wu, Heping Cheng, and Jing Chen. Hierarchical Decoding of Perceived Speech From Non-Invasive Brain Recordings, 2026. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 34:2349–2360. DOI: https://doi.org/10.1109/TNSRE.2026.3688564.

Rui Wang, Marcos Perreau-Guimaraes, Claudio Carvalhaes, and Patrick Suppes. Using phase to recognize English phonemes and their distinctive features in the brain, 2012. Proceedings of the National Academy of Sciences, 109(50):20685–20690. DOI: https://doi.org/10.1073/pnas.1217500109.

H. Weissbart, K. D. Kandylaki, and T. Reichenbach. Cortical tracking of surprisal during continuous speech comprehension, 2020. Journal of Cognitive Neuroscience, 32(1):155–166. DOI: https://doi.org/10.1162/jocn_a_01467.

H. Weissbart, K. D. Kandylaki, and T. Reichenbach. EEG Dataset for 'Cortical Tracking of Surprisal during Continuous Speech Comprehension', 2022. Zenodo, version 2.0.0. DOI: https://doi.org/10.5281/zenodo.7775260.

Emma M. Whitham et al. Scalp electrical recording during paralysis: quantitative evidence that EEG frequencies above 20 Hz are contaminated by EMG, 2007. Clinical Neurophysiology, 118:1877–1888. DOI: https://doi.org/10.1016/j.clinph.2007.04.027.

Anna Wilsch, Toralf Neuling, Jonas Obleser, and Christoph S. Herrmann. Transcranial alternating current stimulation with speech envelopes modulates speech comprehension, 2018. NeuroImage, 172:766–774. DOI: https://doi.org/10.1016/j.neuroimage.2018.01.038.

Xiran Xu, Bo Wang, Boda Xiao, Yadong Niu, Yiwen Wang, Xihong Wu, Heping Cheng, and Jing Chen. The impacts of temporal autocorrelations on EEG decoding, 2026. Biomedical Signal Processing and Control, 113(B):108783. DOI: https://doi.org/10.1016/j.bspc.2025.108783.

Benedikt Zoefel, Alan Archer-Boyd, and Matthew H. Davis. Phase Entrainment of Brain Oscillations Causally Modulates Neural Responses to Intelligible Speech, 2018. Current Biology, 28(3):401–408.e5. DOI: https://doi.org/10.1016/j.cub.2017.11.071.

A Extended Related Work

Speech-tracking research offers evidence from both neural phase structure and activity magnitude. Low-frequency phase patterns can distinguish sentences and depend on acoustic structure and comprehension (Luo & Poeppel, 2007; Peelle et al., 2013; Howard & Poeppel, 2010). High-frequency evidence includes intracranial high-gamma amplitude or power related to speech features and FFR timing locked to periodic sound structure (Nourski et al., 2009; Golumbic et al., 2013; Mesgarani & Chang, 2012; Mesgarani et al., 2014; Oganian & Chang, 2019; Coffey et al., 2019). We derive continuous phase, two-bin phase, and log power from one high-pass waveform and compare their utility for short-window retrieval. This representation comparison does not reproduce a particular high-gamma-envelope reconstruction method.

Envelope reconstruction, common representation spaces, and nonlinear encoders have progressively established EEG–speech correspondences (Crosse et al., 2016; de Cheveigné et al., 2018; Monesi et al., 2020; Accou et al., 2021a). BrainMagick and hierarchical decoding align brain activity with pretrained or multilevel speech representations for multi-candidate retrieval (Défossez et al., 2023; Wang et al., 2026); high-frequency-related models and combined frequency inputs have improved binary matching (Thornton et al., 2024b; Puffay et al., 2022). Our two branches pair EEG with speech features at the corresponding time scale and assess independent HF retrieval and its gain to LF ranking using the same queries and candidate sets. Sections 2 and B describe the models and training setup.

A.1 Low-frequency speech tracking and phase evidence

Low-frequency speech tracking concerns the temporal correspondence between neural activity and slower speech structure. Luo & Poeppel (2007) compared single-trial theta-band (4–8 Hz) phase trajectories in listeners' magnetoencephalography with sentence-specific templates. These trajectories discriminated sentences, with weaker discrimination when intelligibility decreased. Phase locking between speech and auditory-cortex activity increases during comprehension (Peelle et al., 2013), and 2–9-Hz EEG phase has been used to identify English phonemes and distinctive features (Wang et al., 2012). Later work relates these responses to envelope edges, attentional selection, and linguistic hierarchy (Doelling et al., 2014; Ding & Simon, 2012a; Ding et al., 2016; Oganian et al., 2023); external rhythmic interventions have tested the influence of relative phase on comprehension (Riecke et al., 2018; Wilsch et al., 2018; Zoefel et al., 2018). Low-frequency phase and amplitude may co-vary (Howard & Poeppel, 2012), and frontal modulation and auditory–motor coupling change with intelligibility and syllabic rate (Park et al., 2015; Assaneo & Poeppel, 2018). Neural envelope-tracking thresholds under different noise levels correlate with behavioral speech-intelligibility thresholds (Vanthornhout et al., 2018). Reversed speech also evokes discriminable phase patterns (Howard & Poeppel, 2010); phase can therefore reflect acoustic structure while also being modulated by comprehension and attention.

Neural speech tracking has also been applied to residual language processing in patients unable to perform behavioral tests (Gui et al., 2020), selective attention and auditory-distractor sensitivity in adults with ADHD (Levy et al., 2025), and word-level processing in children with ADHD (Luo et al., 2023). These motivate possible applications; our experimental task is retrieval of the actually heard segment from audio candidates using listening-time EEG.

A.2 High-frequency responses and temporal structure

High-frequency scalp EEG is difficult to measure reliably. Volume conduction attenuates and spatially smooths cortical potentials (Srinivasan et al., 1998); activities from distributed sources may partly cancel at the scalp (Ahlfors et al., 2010); and higher-frequency recordings are vulnerable to muscle artifacts (Whitham et al., 2007). Together these factors reduce measurable signal-to-noise ratio.

FFR studies examine neural timing locked to periodic sound structure, often with stimulus fundamental frequencies around 80–500 Hz (Coffey et al., 2019). Individual EEG electrodes can track speech fundamental frequency and its rapid variation (Fu et al., 2017; 2019). Conventional paradigms repeatedly present short sounds and average hundreds to thousands of trials (Bachmann et al., 2021; Coffey et al., 2016; 2019). Brainstem-related responses have also been extracted from

scalp EEG during continuous natural speech (Forte et al., 2017; Maddox & Lee, 2018); Forte et al. (2017) reported a brainstem-related delay of about 9 ms in 100–300-Hz scalp responses. Magnetoencephalography studies observed a cortical response near 30–35 ms for 70–150-Hz fundamental frequency and showed that higher-harmonic envelope modulation could drive this band (Schüller et al., 2023); another study reported responses around 80–90 Hz with a peak latency near 40 ms in a 70–200-Hz range (Kulasingham et al., 2020). Such evidence implicates timing related to fundamental frequency and rapid envelope changes (Aiken & Picton, 2008; Canneyt et al., 2021; Bachmann et al., 2021; Coffey et al., 2016; Bidelman, 2018; Kulasingham et al., 2020; Schüller et al., 2023), typically after accumulating longer records.

High-frequency speech tracking also encompasses changes in activity strength. Intracranial high-gamma amplitude and power vary with speech events and can distinguish auditory objects and encode acoustic or phonetic features (Nourski et al., 2009; Golumbic et al., 2013; Mesgarani & Chang, 2012; Mesgarani et al., 2014; Oganian & Chang, 2019). At the scalp, a slow envelope of 70–150-Hz high-gamma power can track the speech envelope and, for some participants, improve joint reconstruction with LF EEG (Synigal et al., 2020). Low-frequency phase and gamma amplitude have carried complementary speech information in magnetoencephalography (Gross et al., 2013). Animal auditory-cortex studies suggest that local-field-potential phase can increase or stabilize stimulus information in spike patterns and contribute to laminar representation and gain changes across time scales (Kayser et al., 2009; Szymanski et al., 2011; Kayser et al., 2015). Our analysis instead measures segment-retrieval performance from scalp EEG input representations.

A.3 From reconstruction and matching to multi-candidate retrieval

Forward temporal response functions use speech features to predict EEG across lags, then compare the prediction with the recorded signal (Ding & Simon, 2012b; Crosse et al., 2016). Reverse reconstruction estimates a heard speech envelope from multichannel, lagged EEG; its correlation with candidate envelopes supports identification (Crosse et al., 2016). Canonical correlation analysis (CCA) linearly projects EEG and speech into spaces whose paired time courses are maximally correlated, making projected correlation a match score (de Cheveigné et al., 2018). Nonlinear encoders extend this shared-space approach by learning features useful for distinguishing correct from mismatched pairs.

Nonlinear models have developed along short-window EEG–speech matching and envelope reconstruction. CNN–LSTM networks have classified matching EEG and speech-envelope windows (Monesi et al., 2020); dilated convolutions model temporal correspondence (Accou et al., 2021b;a); VLAAI reconstructs the speech envelope from EEG (Accou et al., 2023). CLASH learns time-aligned EEG representations through contrastive training for envelope reconstruction (Accou et al., 2024b). A transformer-based auditory EEG encoder has combined convolution, attention, and feed-forward processing with a speech encoder for binary matching and envelope regression on unseen stories and participants (Bollens et al., 2025).

BrainMagick aligns brain recordings with wav2vec 2.0 features to retrieve speech among many candidates (Défossez et al., 2023). ConvConcatNet uses iterative convolutions and cross-layer concatenation; hierarchical decoding additionally uses acoustic, self-supervised, and language representations (Wang et al., 2026). Their 120-Hz and 100-Hz brain-recording sampling rates imply Nyquist limits of 60 and 50 Hz, respectively, below the full HF range examined here. Défossez et al. (2023) reported 25.7% Top-10 from 190 candidates (about 5% chance), while Wang et al. (2026) reported 54.3% mean Top-10 from 104–208 candidates per participant. Those values reflect different protocols and are not matched-condition performance comparisons.

Most earlier high-frequency-related decoding evidence comes from binary matching or envelope reconstruction (Thornton et al., 2024b;a; Synigal et al., 2020; Puffay et al., 2022). Puffay et al. (2022) used dilated convolutions for fundamental-frequency-related binary EEG–speech matching, obtaining above-50% accuracy on five-second windows from two male-narrated stories. Thornton et al. (2024b) reported 65.64% for three-second binary matching with a 100-model ensemble using only 70–220-Hz EEG. Combined LF and HF inputs have also aided reconstruction or binary matching (Thornton et al., 2024b; Synigal et al., 2020; Puffay et al., 2022). Large-pool retrieval presents a different test: the target must outrank many mismatches. We evaluate HF retrieval for unseen listeners

among more than a thousand candidates on average, its gain to LF ranking under identical candidate sets, and the information retained by continuous and two-bin phase.

B Extended Methods

B.1 Branch inputs and encoder transformations

Let $b \in \{L, H\}$ denote the low- and high-frequency branches. In the source-dataset main experiment, $X_L \in \mathbb{R}^{64\times 500}$ and $X_H \in \mathbb{R}^{64\times 5120}$. Candidate A_j has speech representations $F_L(A_j) \in \mathbb{R}^{768\times 500}$ and $F_H(A_j) \in \mathbb{R}^{28\times 5120}$. The first is frozen wav2vec 2.0 layer-9 output, and the second is a fixed-parameter auditory-subband amplitude representation. In the equations, P_0 and P_1 are trainable 1×1 input and terminal channel projections; $D_{\mathbf{d}}$ denotes three dilated convolution layers with dilation sequence $\mathbf{d}$. Superscripts E and A denote EEG and audio sides. All branch- and side-specific encoder parameters are distinct.

The LF EEG first undergoes adaptive channel modulation. If Pool_t pools over time and $g_{64\to 8\to 64}$ is a two-layer channel mapping, then

$$w(X_L) = g_{64\to 8\to 64}(\mathrm{Pool}_t(P_{\mathrm{ACM}}(X_L))), \tag{B1}$$

$$X'_L = X_L + \alpha[X_L \odot w(X_L)], \qquad \alpha_{\mathrm{init}} = 0.1. \tag{B2}$$

The product $\odot$ broadcasts channel weights and α is trainable. A three-layer convolution with kernel size 3 and dilations $(1, 3, 9)$ encodes the modulated EEG; the LF audio encoder uses an input projection, dilations $(2, 6, 18)$, and a terminal projection:

$$U_L = D^{E,L}_{(1,3,9)}(X'_L), \qquad V_{L,j} = \left[P^{A,L}_1 \circ D^{A,L}_{(2,6,18)} \circ P^{A,L}_0\right](F_L(A_j)). \tag{B3}$$

The HF EEG and subband amplitudes have independent projections and three-layer encoders, both yielding 64-channel latent sequences. Compositions act right to left:

$$U_H = \left[P^{E,H}_1 \circ D^{E,H}_{(2,6,18)} \circ P^{E,H}_0\right](X_H), \tag{B4}$$

$$V_{H,j} = \left[P^{A,H}_1 \circ D^{A,H}_{(2,6,18)} \circ P^{A,H}_0\right](F_H(A_j)). \tag{B5}$$

B.2 Matching vectors, scalar scores, and retrieval

For $U_b, V_{b,j} \in \mathbb{R}^{64\times T_b}$, with $T_L = 500$ and $T_H = 5120$, temporal cosine similarity in each latent channel forms $c_{b,j} \in \mathbb{R}^{64}$:

$$c_{b,j,k} = \frac{\langle U_b[k,:], V_{b,j}[k,:]\rangle}{\|U_b[k,:]\|_2 \, \|V_{b,j}[k,:]\|_2}, \qquad k = 1, \ldots, 64. \tag{B6}$$

Each scoring head maps 64 to 64 dimensions, applies a sigmoid, and maps 64 to one scalar. Writing the two affine maps as $M_{b,1}$ and $M_{b,2}$ gives

$$s_b(j) = M_{b,2}(\sigma(M_{b,1}(c_{b,j}))), \qquad b \in \{L, H\}. \tag{B7}$$

Each score depends only on its own branch's EEG and candidate-audio features. The single-branch models are trained independently, with fusion applied only to their scores for the same candidate. For candidate set $\mathcal{C}$,

$$s_F(j) = 0.6 s_L(j) + 0.4 s_H(j), \tag{B8}$$

$$\hat{j}_b = \arg\max_{j\in\mathcal{C}} s_b(j), \qquad b \in \{L, H, F\}. \tag{B9}$$

B.3 Scores in the different experiments

The main LF condition uses s_L, the HF condition uses s_H trained on raw high-pass waveform W, and fusion uses s_F. The four LF baselines replace only the EEG encoder and still produce scores through Equations (B6)–(B7). Representation models trained independently on W, Φ, Q, and P yield $s^{(W)}_H$, $s^{(\Phi)}_H$, $s^{(Q)}_H$, and $s^{(P)}_H$, all paired with the same type of auditory-subband input. The

Table S1: Five natural-speech EEG datasets and eligible five-second query windows. Training, validation, and test windows are participant-disjoint where those splits apply. EEG hours are the cumulative duration of eligible recordings.

Dataset	Language	Subjects/records	EEG (h)	Train	Val.	Test
SparrKULee	Dutch	85/666	154.75	75,020	17,412	18,680
Etard	En./Du.	20/132	4.95	–	–	3,477
Weissbart	English	13/195	7.52	–	–	5,304
hardware-EEG	Chinese	6/150	12.54	–	–	8,958
Matlab-EEG	Chinese	6/150	12.54	–	–	8,958

phase input has 128 EEG channels; the other three have 64, and all project to 64 latent channels. Models for Q timing, bands, and electrode subsets, as well as LF timing and electrode controls, are retrained per condition. They are neither automatically fused with the other branch nor produced by substituting inputs into a frozen model at test time. The public external datasets use source-retrained models on shared electrodes, frozen to calculate s_L, s_H, and s_F; the self-collected datasets evaluate only frozen HF scores.

B.4 Implementation details

Each training query pairs its synchronized audio positive with four same-recording mismatches, and five-class cross-entropy optimizes candidate ranking. Parameters are updated after accumulating eight queries. LF and HF models train separately with AdamW, learning rate 10^{-3}, weight decay 10^{-2}, and gradient-norm cap 5. Training lasts at most 100 epochs; the checkpoint with minimum validation cross-entropy is saved after each epoch, and training stops after ten epochs without improvement. Formal models use the prespecified random seed 0. Fusion weights are selected on source validation data only. Experiments were implemented in PyTorch on one A800 GPU.

B.5 Datasets, preprocessing, and ethics

B.5.1 SparrKULee

The main study uses the SparrKULee natural-speech listening dataset (Accou et al., 2024a). We include 666 recordings from 85 participants listening to Dutch audiobooks or podcasts. Original EEG has 64 channels at 8192 Hz. After alignment by stimulus-synchronization events, eligible recordings total 154.75 hours, or 44.25–146.33 minutes per participant. HF EEG is resampled to 1024 Hz. Temporal autocorrelation of continuous EEG can offer stimulus-independent decodable cues (Xu et al., 2026), and random window-level splits can leak participant identity and inflate unseen-listener performance (Brookshire et al., 2024). We therefore use participant-disjoint training, validation, and test splits of 54, 14, and 17 people (470, 97, and 99 recordings), yielding 75,020, 17,412, and 18,680 complete five-second windows from the jointly valid EEG and speech intervals.

Preprocessing is applied to continuous EEG and includes LF edge handling, artifact interpolation, multichannel Wiener filtering, common-average reference, and correction using stimulus-trigger times. LF EEG is resampled to 100 Hz and limited to 0.5–40 Hz; HF EEG retains a continuous waveform high-pass filtered above approximately 50 Hz on a 1024-Hz time grid. Filter-affected edges are removed before stimulus-aligned non-overlapping five-second windows are extracted, and channels are standardized within each window. None of the four datasets below participates in model training, checkpoint selection, or fusion-weight selection; they are used solely for retrieval evaluation on target recordings.

B.5.2 Etard

The first public external dataset comprises continuous-speech EEG recordings by Etard and colleagues (Etard & Reichenbach, 2022; Etard et al., 2019; Etard & Reichenbach, 2019). We include four English and four Dutch stories in quiet, involving 18 and 15 participants, respectively; deduplication yields 20 people and 132 recordings. The released 63-channel, 1000-Hz EEG is aligned to audio stimulus onset. Matching preprocessing and dual 100/1024-Hz resampling leave 4.95 hours of

eligible EEG, 8.83–18.02 minutes per participant, and 3,477 five-second queries. We retrain source models using the 56 electrode names shared with SparrKULee and freeze them for target evaluation.

B.5.3 WEISSBART

The second public external dataset is from Weissbart and colleagues (Weissbart et al., 2022; 2020). Thirteen participants heard three English stories; we include 195 released recordings aligned to audio onset, originally at 63 channels and 1000 Hz. The same preprocessing and dual resampling yield 34.70 minutes per participant, 7.52 hours in total, and 5,304 five-second queries. Models retrained on the 56 named electrodes shared with SparrKULee are frozen for evaluation.

B.5.4 SELF-COLLECTED HARDWARE-TRIGGER EEG

This external dataset contains EEG recorded during 25 Chinese continuous-story stimuli, numbered 1–25, with hardware trigger events written at audio playback. Six participants each completed 25 trials, giving 150 recordings. The original 1000-Hz signal has 62 EEG and two electrooculography channels. After preprocessing, each participant contributes 125.36 minutes of eligible EEG, for 12.54 hours overall and 8,958 jointly valid five-second queries. A source model retrained on the 55 electrode names shared with SparrKULee is frozen for target evaluation.

B.5.5 SELF-COLLECTED MATLAB-EVENT EEG

This external dataset uses the same 25 Chinese stories but establishes EEG–audio correspondence from MATLAB playback-event records. We include the first six participants' 25 numbered story trials each, totaling 150 recordings. Original acquisition is at 1000 Hz with 62 EEG and two electrooculography channels. After preprocessing, each participant has 125.36 minutes of eligible EEG, totaling 12.54 hours and 8,958 complete five-second queries. The frozen source model is retrained on the 55 electrode names shared with SparrKULee before target evaluation.

B.5.6 ETHICS AND INFORMED CONSENT

The acquisition and use of both self-collected datasets were approved by the relevant institutional ethics committee. Its name and approval number are withheld during double-blind review and can be added after review. All 12 participants provided written informed consent before collection. Participant-level results use anonymous identifiers.

C EXTENDED RESULTS

Figures S1–S5 provide the complete participant-level distributions for the analyses discussed in Section 3.

C.1 TIMING-OFFSET ANALYSIS

Figure S1 shows five-way Top-1, full-pool Top-1, and Top-10 for the independently retrained LF waveform and HF phase-bin models.

C.2 FREQUENCY-BAND ANALYSIS

Figure S2 shows all three retrieval metrics for the nine independently trained 50-Hz bands.

C.3 SPATIAL INPUT ANALYSIS

Figure S3 shows the electrode layouts and all three retrieval metrics for the matched LF and HF six-electrode input conditions.

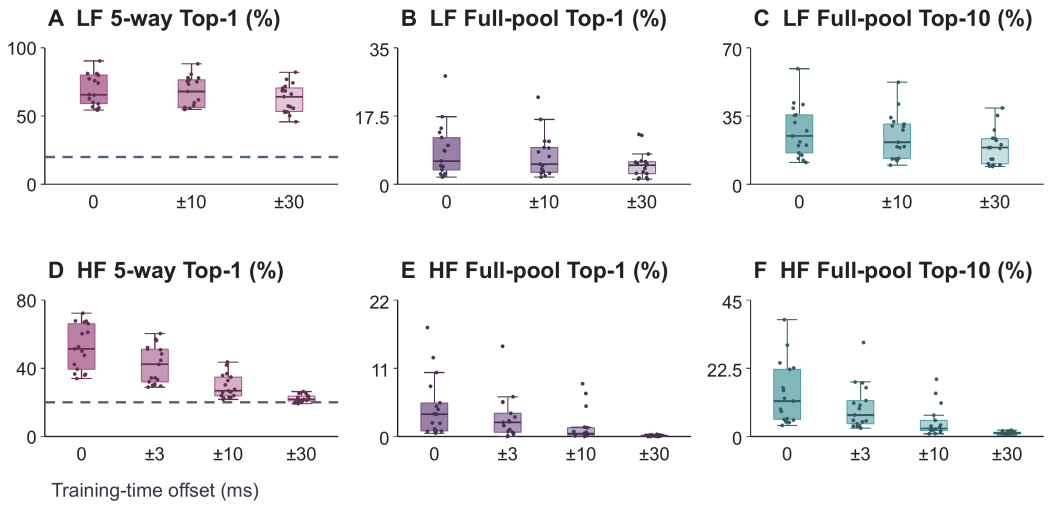

Figure S1: All three retrieval metrics under training-time EEG–audio offsets. (A–C) LF waveform five-way Top-1, full-pool Top-1, and full-pool Top-10; (D–F) the corresponding HF two-bin-phase metrics (%). LF uses 0, ± 10, and ± 30 ms; HF uses 0, ± 3, ± 10, and ± 30 ms. Nonzero offsets are sampled per recording/trial and each condition is retrained. Dashed lines in A and D mark 20% five-way chance. Dark points are unseen participants ($n = 17$); boxplot conventions follow Figure 2.

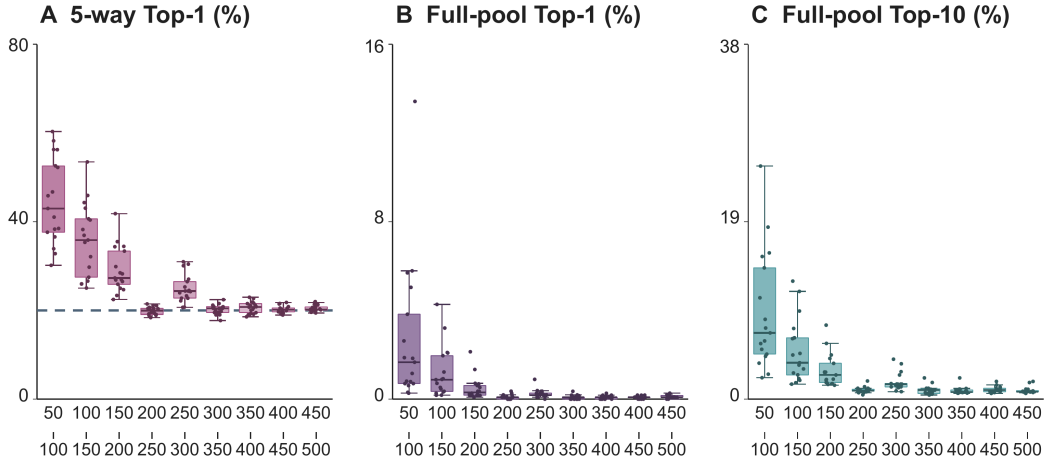

Figure S2: HF two-bin-phase retrieval across nine adjacent 50-Hz bands: (A) five-way Top-1, (B) full-pool Top-1, and (C) full-pool Top-10 (%). Paired numbers on the horizontal axis denote each band's edges; models are independently trained. The dashed line in A marks 20% chance. Dark points are unseen participants ($n = 17$); boxplot conventions follow Figure 2.

C.4 Public external datasets

Figure S4 provides all three metrics for frozen cross-dataset evaluation on Etard and Weissbart. Each target dataset builds its own complete candidate pool, so chance rates for full-pool Top-1 and Top-10 differ between datasets.

C.5 Self-collected datasets

Figure S5 shows participant-level five-way and full-pool results for both self-collected cohorts.

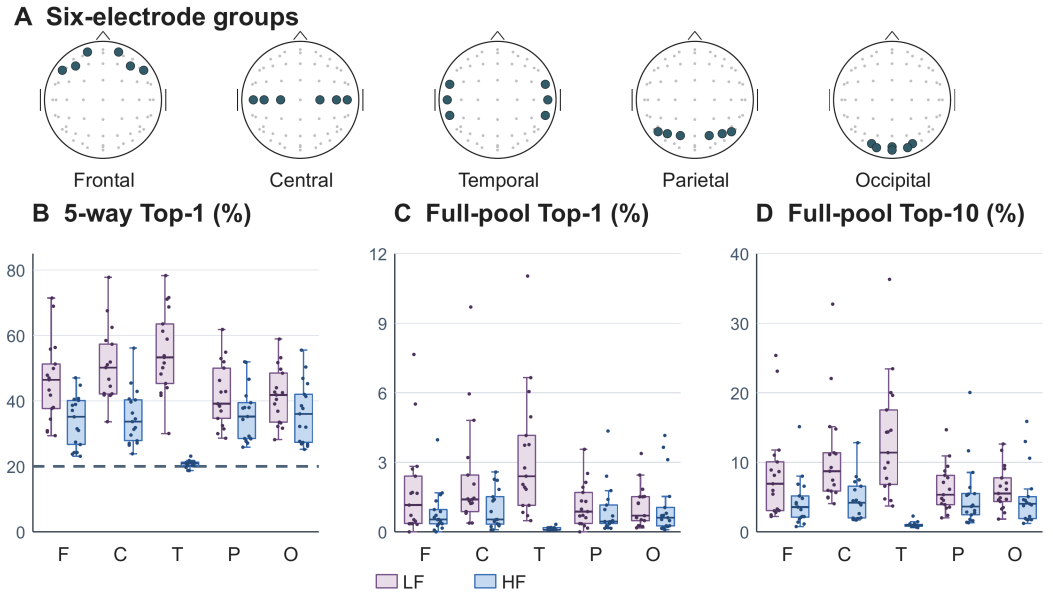

Figure S3: Five six-electrode layouts and LF/HF spatial input conditions. (A) Scalp positions; dark dots are retained sensors and gray dots the others. F, C, T, P, and O denote frontal, central, temporal, parietal, and occipital groups. (B–D) Five-way Top-1, full-pool Top-1, and full-pool Top-10 (%) for independently trained LF (experiments 060–064) and HF two-bin phase (experiments 033–037). The dashed line in B is 20% chance. Dark points are unseen participants ($n = 17$); boxplot conventions follow Figure 2. The branch inputs and encoders differ, so this compares retrieval with the same sensor subsets rather than source locations.

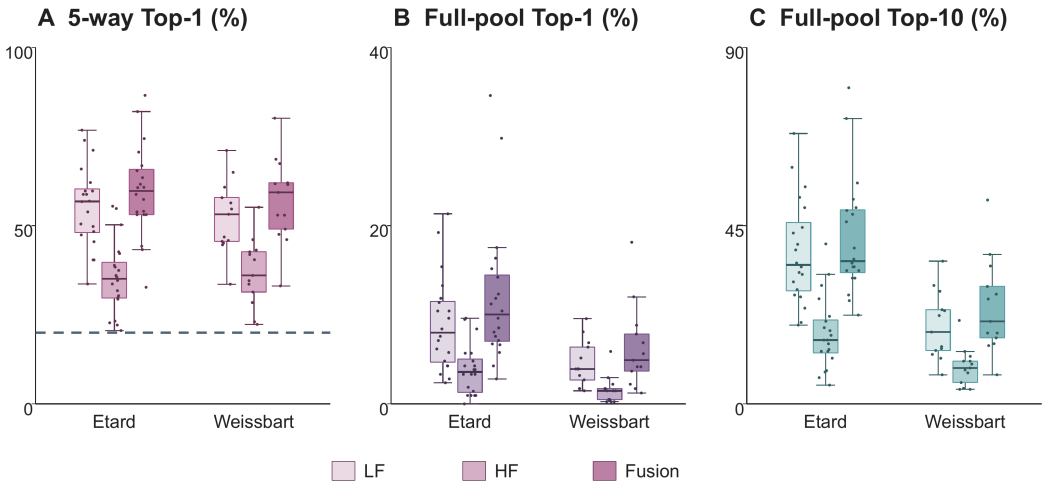

Figure S4: Frozen cross-dataset retrieval on public EEG recordings: (A) five-way Top-1, (B) full-pool Top-1, and (C) full-pool Top-10 (%). LF, HF, and source-validation-weighted fusion are shown for Etard ($n = 20$) and Weissbart ($n = 13$). The dashed line in A marks 20% chance. Each dark point is a target participant; boxplot conventions follow Figure 2. The two datasets' full-pool chance levels differ.

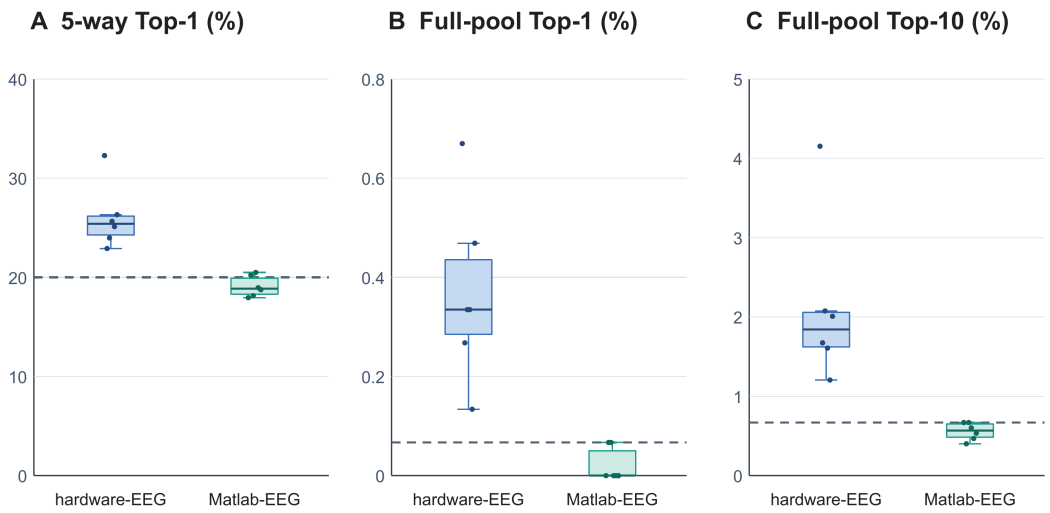

Figure S5: Frozen HF retrieval for two self-collected Chinese-story EEG datasets: (A) five-way Top-1, (B) full-pool Top-1, and (C) full-pool Top-10 (%). Hardware-EEG and Matlab-EEG each contain six participants hearing the same 25 stories, with 25 trials and 1,493 queries per participant. Each dark point is a participant; boxplot conventions follow Figure 5. Dashed lines mark chance rates of 20%, $1/1493 = 0.067\%$, and $10/1493 = 0.670\%$. Both datasets use their supplied EEG–audio time origin without additional test-time offsets.