論文 / 著書情報
Article / Book Information

Title	Auditory Perception as a Surrogate for Auditory Imagery in EEG-Based BCI Training: Neural Evidence and a Feasibility Study
Authors	Zhuohao Zhang, Haruto Hamada, Phurin Rangpong, Akima Connelly, Tohru Yagi
Citation	IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, Vol. 34, , pp. 3041-3052
Pub. date	2026, 6
DOI	https://doi.org/10.1109/TNSRE.2026.3703988
Creative Commons	Information is in the article.

Auditory Perception as a Surrogate for Auditory Imagery in EEG-Based BCI Training: Neural Evidence and a Feasibility Study

Zhuohao Zhang, *Graduate Student Member, IEEE*, Haruto Hamada, Phurin Rangpong, Akima Connelly, and Tohru Yagi, *Member, IEEE*

***Abstract*— Brain-computer interface (BCI) systems are commonly classified as reactive or active. Reactive BCIs rely on responses to external stimuli, simplifying decoding but limiting direct user control. In contrast, active BCIs enable intuitive control through spontaneous mental activity but impose greater signal-processing challenges. Speech BCIs, a subset of active systems, hold promise for individuals with severe neurological impairments such as locked-in syndrome (LIS), potentially enabling communication through imagined speech. However, their adoption is hindered by the demanding pre-training required to collect imagery data, which often induces fatigue, and the difficulty many users face in generating consistent, high-quality imagery. To address these limitations, we propose a hybrid BCI framework that uses passive listening tasks as a surrogate for auditory imagery tasks during model training. As a first step, this proof-of-concept feasibility study examined neural responses to five Japanese vowels (/a/, /i/, /u/, /e/, /o/) under listening and imagery conditions in a limited experimental setting with healthy male participants, using event-related potentials (ERPs), topographical mapping, and event-related spectral perturbations (ERSPs), followed by statistical analysis of shared and distinct neural patterns. In the primary three-class and binary classification analyses, classification accuracies for imagery signals obtained from models trained solely on listening data were comparable to, and in some settings higher than, those trained directly on imagery data, with both outperforming chance level. These findings indicate that passive listening may provide effective training data for active auditory imagery BCIs within the tested setting, potentially reducing cognitive demands without compromising performance. Further validation in larger and more diverse cohorts is required before clinical application.**

***Index Terms*— Auditory perception, auditory imagery, brain–computer interface, electroencephalography.**

Received 30 October 2025; revised 6 April 2026 and 6 May 2026; accepted 11 June 2026. Date of publication 15 June 2026; date of current version 24 June 2026. Associate Editor: Junfeng Sun. This work was supported in part by the Japan Science and Technology Agency (JST) BOOST program (Broadening Opportunities for Outstanding young researchers and doctoral students in STrategic areas), Japan, under Grant JPMJBS2430; and in part by JST Support for Pioneering Research Initiated by the Next Generation (SPRING), Japan, under Grant JPMJSP2106 and Grant JPMJSP2180. *(Corresponding author: Zhuohao Zhang.)*

Zhuohao Zhang, Haruto Hamada, and Tohru Yagi are with the School of Engineering, Institute of Science Tokyo, Tokyo 152-8550, Japan (e-mail: chiyo.t.aa@m.titech.ac.jp).

Phurin Rangpong and Akima Connelly are with the School of Environment and Society, Institute of Science Tokyo, Tokyo 152-8550, Japan.

This article has supplementary downloadable material available at https://doi.org/10.1109/TNSRE.2026.3703988, provided by the authors.

Digital Object Identifier 10.1109/TNSRE.2026.3703988

I. Introduction

SPEECH brain–computer interfaces (BCIs) enable communication without overt speech or movement. They hold particular promise for individuals with severe neurological impairments such as locked-in syndrome (LIS) [1]. Electroencephalography (EEG) is portable, inexpensive, and safe, making it well-suited for long-term BCI applications [2], [3], [4].

From an end-user perspective, EEG speech-imagery BCIs face practical barriers. While many EEG-based studies have attempted to classify speech imagery signals, training these systems typically requires prolonged imagery sessions to collect calibration data [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], which are cognitively demanding, induce fatigue, and reduce motivation [15], [16]. In addition, many users struggle to generate consistent, high-quality imagery, leading to degraded data reliability and model performance [17]. Reducing this training burden remains a key challenge [17], [18]. This work addresses this challenge directly by proposing a listening-based training paradigm.

To the best of our knowledge, this is the first EEG study to evaluate whether models trained on listening data can effectively classify imagery data. Cognitive neuroscience studies show that listening to and imagining the same speech activate overlapping cortical networks, particularly within auditory and language areas [19], [20], [21], [22]. Building on this evidence, we hypothesize that EEG signals recorded during passive listening could share features with those observed during imagery. If confirmed, passive listening, which requires little cognitive effort, could be exploited as training data for imagery-based BCIs. This approach retains the intuitive control of active BCIs while alleviating the heavy training demands, offering a more practical path toward real-world applications.

Another challenge arises from the intrinsic complexity of speech imagery. It is a complex process rather than a single phenomenon [23]; covert speech engages both auditory and articulatory components, often recruiting motor-related regions even in the absence of overt movement in many studies [5], [6], [7], [9], [10], [11], [14], [20]. This heterogeneity complicates feature isolation and the interpretation of decoding results [20], [24]. To address this issue, the present study emphasizes auditory features while minimizing potential confounds from articulatory processes. In addition, existing studies have primarily focused on machine learning accuracy, but often pro-

vide limited insight into the underlying neural characteristics [5], [6], [7], [8], [9], [10], [11], [12], [13], [14]. In contrast, our approach reduces the confounding influence of motor activity and directly examines EEG features, thereby contributing both to decoding performance and to a deeper understanding of auditory imagery mechanisms.

Prior EEG studies on Japanese vowels have adhered mainly to an imagery-only training–testing paradigm [12], [13], [14]. While [12] and [13] considered only two vowels (/a/, /i/)—with [12] combining EEG with fMRI currents—and [14] included five vowels but restricted analysis to binary classification, none explored the relationship between listening and imagery, nor did they systematically examine the underlying neural correspondence.

This study proposes a novel framework that uses passive listening as a surrogate for imagery during EEG-based BCI training. We collected neural data from participants during both passive listening and auditory imagery of five Japanese vowels. First, we compared the neural responses across conditions to identify shared and distinct patterns. Next, we evaluated classification performance under two paradigms: (1) models trained on *imagery* data and tested on *imagery* trials (conventional approach), and (2) models trained on *listening* data and tested on *imagery* trials (proposed approach). Our results indicate that listening-based training achieves classification accuracy comparable to imagery-based training and remains significantly above chance level.

Overall, this work provides a proof-of-concept for hybrid training in speech-imagery BCIs. By leveraging shared neural features between listening and auditory imagery, our approach addresses a key obstacle to active BCI adoption—heavy pre-training demands. It also points toward more practical interfaces for assistive communication and neurorehabilitation in individuals with severe motor or speech impairments.

II. Methods

A. Language Selection and Stimuli

Unlike other languages, Japanese consists of 50 phonemes, collectively called the Gojūon (fifty-sound chart), whose pronunciations remain invariant across lexical contexts, making vowel decoding a practical starting point for higher-level speech decoding. By contrast, foundational phonemes in languages such as English or Chinese exhibit significant variability; the pronunciation of alphabet letters can differ greatly when embedded within words. Therefore, Japanese was chosen as the target language for this study.

To further minimize potential confounds from articulatory processes, we selected female voices from Japanese anime as auditory stimuli for the five vowels (/a/, /i/, /u/, /e/, /o/) and normalized them to −23 LUFS (44.1 kHz, 16-bit, mono). Only male participants were recruited to reduce the likelihood that they would internalize the imagined sounds as their own vocal productions. This design minimized motor-related imagery and emphasized auditory-specific neural features. The audio stimuli are publicly available at:

https://booth.pm/ja/items/3171096

B. Participants

Ten healthy male volunteers aged 22–27 (mean 24.0) participated in the experiment. All participants were right-handed, as assessed using the Handedness Questionnaire [25], and had normal hearing. None of the participants reported any history of neurological, psychiatric, or musculoskeletal disorders.

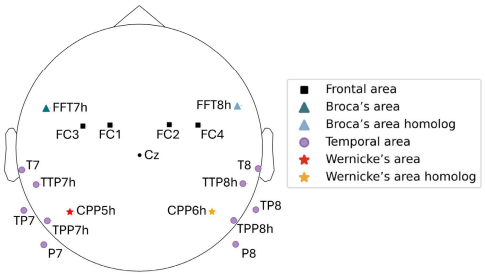

Fig. 1. Electrode positions highlighting brain regions related to audition and speech.

This study was conducted with the approval of the IRB at the Institute of Science Tokyo (# 2025120)

C. Experimental Protocol and Setup

1) Signal Acquisition: EEG activity was recorded using a commercial amplifier (Polymate Pro MP6100, Miyuki Giken, Japan). Nineteen electrodes were positioned according to the international 10–5 system, with an emphasis on auditory-related cortical regions [26], [27] (Fig. 1). Two electrooculographic (EOG) electrodes were placed to monitor ocular artifacts. In addition, surface electromyography (EMG) electrodes were positioned over the laryngeal region to verify the absence of overt articulatory movements [28].

2) Stimuli and Experimental Blocks: Data were collected across two sessions on different days, each held at the same time of day to minimize circadian variability and fatigue. Each session included two experimental blocks: *Listening* and *Auditory Imagery*. Both blocks contained five Japanese vowels (/a/, /i/, /u/, /e/, /o/) and a white-noise *Control* condition. Each stimulus was presented 54 times per block, yielding 324 trials per block and 648 trials in total.

3) Experimental Protocol: The *Imagery* paradigm was adapted from a prior auditory imagery study [12]. The *Listening* block was newly introduced to enable direct comparison between auditory perception and imagery (Fig. 2). Each trial began with a fixation cross presented for 2.0–3.0 s (randomized), preventing participants from anticipating the onset of the task phase. The final 1.0 s of this period served as the baseline. Subsequently, a 4.0 s task phase was initiated:

a) *Listening*: a 0.4 s auditory stimulus was presented at 0.0 s and repeated at 2.0 s.
b) *Imagery*: a 0.4 s auditory stimulus was presented at 0.0 s, after which participants imagined the same sound synchrony with the expected timing at 2.0 s.
c) *Control*: white noise was presented at 0.0 s, and participants remained still without performing imagery. The control condition was intended as a passive baseline rather than a temporally matched sensory control.

Each trial ended with a 3.0 s rest period indicated on screen, during which participants were allowed to blink and relax.

D. Data Processing and Analysis

EEG data (19 EEG and 2 EOG channels) were processed in EEGLAB (MATLAB) [29]. Signals were bandpass filtered between 1 and 35 Hz to suppress slow drifts and high-frequency noise while preserving auditory-related frequency bands [30]. Artifacts were removed by manual rejection of high-noise trials and independent component analysis (ICA), with eye-movement components identified and excluded using EOG references. Data were baseline-corrected using the −1 to 0 s interval and subsequently normalized with a z-score transformation for group analysis. For each participant, data from both sessions were combined. Due to excessive noise in both EEG and EMG on Day 2 for Subject 9, only Day 1 data were retained. To assess the potential impact of this reduced trial number, we additionally conducted a sensitivity analysis excluding Subject 9 entirely. Excluding Subject 9 did not materially affect the overall conclusions, and the corresponding results are reported in the Supplementary README. Details of EMG preprocessing are also included in the Supplementary Materials.

Subsequent analyses included event-related potentials (ERP), topographical mapping, and event-related spectral perturbation (ERSP) to characterize neural responses. Statistical tests were performed to quantify similarities and differences between the *Listening* and *Imagery* conditions. Finally, classification analyses were conducted to evaluate the performance of models trained on listening versus imagery data.

ERSP was computed as the average spectral power across trials:

$$ERSP(f,t) = \frac{1}{n}\sum_{k=1}^{n} |F_k(f,t)|^2 \tag{1}$$

where $F_k(f,t)$ is the spectral estimate of the k^{th} trial at frequency f and time t. Power changes were expressed relative to the baseline using a decibel (dB) transformation:

$$Power_{\_change} = 10log_{10}\frac{Power_{\_task}}{Power_{\_baseline}} \tag{2}$$

For subsequent analyses, power values were grouped into eight frequency bands:

- Theta: 4–8 Hz
- Low-Alpha: 8–10 Hz
- High-Alpha: 10–12 Hz
- Beta-I: 12–15 Hz
- Beta-II: 15–20 Hz
- Beta-III: 20–25 Hz
- Beta-IV: 25–30 Hz
- Gamma: 30–35 Hz

III. Results

A. ERP Responses

ERP analysis was first conducted to validate the recording of task-related brain activity. The Cz channel was selected for analysis. Fig. 3 shows ERP waveforms for the five vowel conditions and the *Control* condition, with *Listening* results plotted in black and *Imagery* results in blue. The y-axis denotes ERP amplitude (μV), and the x-axis denotes time from −1 to 4 s. Black dashed lines indicate the onset of external auditory stimuli, whereas yellow dashed lines in the imagery panels mark the time point at which participants were required to begin the imagery task.

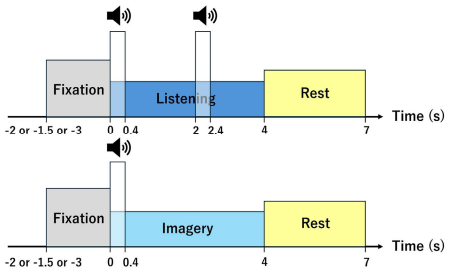

Fig. 2. Experimental protocol illustrating the listening and imagery conditions.

In the *Listening* condition, a robust N200 (N2; blue arrows) was observed approximately 200 ms after stimulus onset, reflecting auditory recognition [31]. This was followed by a P300 (P3; red arrows) at ∼300 ms, typically associated with attentional engagement and task-related processing [31]. A subsequent processing negativity (PN; purple arrows) was consistently observed across all vowels, reflecting the selective-attention demands of vowel identification [32], [33], [34].

In the *Imagery* condition, similar N200 and P300 responses were elicited after stimulus onset. PN responses were also present, suggesting sustained selective attention. Because only a single auditory stimulus was presented, no additional ERP components were observed at the 2-s mark, when participants generated imagery without external input.

In the *Control* condition, N200 and P300 components were detected; however, the typical PN was absent within the highlighted yellow interval. Instead, the waveform gradually returned to baseline. This pattern is consistent with studies suggesting that PN is related to attentional resource allocation; under passive or non-task conditions, only a smooth return of the P3 to baseline is observed without additional negative deflection [33], [34], [35], [36], [37].

These ERPs confirm effective task engagement in *Listening* and *Imagery*, whereas *Control* elicited attenuated and less sustained responses.

B. Topographical Map

1) Time-Segmented Topographical Map: After confirming successful recordings, time-segmented topographical mapping was conducted to identify task-related windows. Because the auditory stimuli lasted 0.4 s, the 4-s task period was divided into consecutive 0.5-s windows. Data from all 10 participants were averaged and filtered between 8–30 Hz to encompass alpha and beta frequency ranges.

Fig. 4(a)-(c) depict power distributions across the scalp relative to the baseline period (−1 to 0 s). The color bar indicates power changes in decibels: red denotes increases, blue denotes decreases, and green indicates no change.

In the *Listening* condition (Fig. 4(a)), power increases were predominantly observed in the bilateral temporal regions, particularly over Wernicke's area (CPP5h) and its

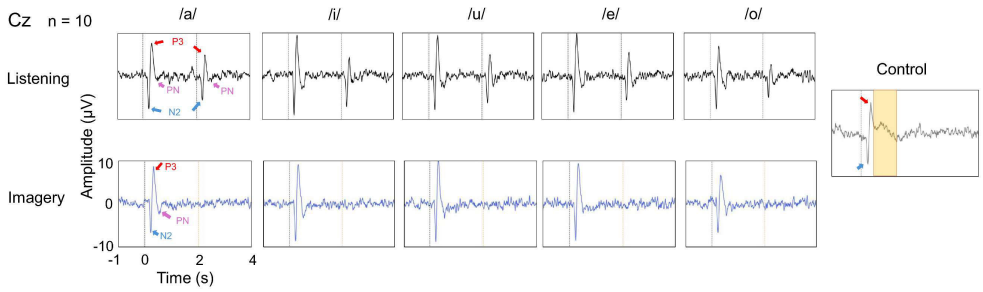

Fig. 3. ERP results showing neural responses for the related task and control conditions.

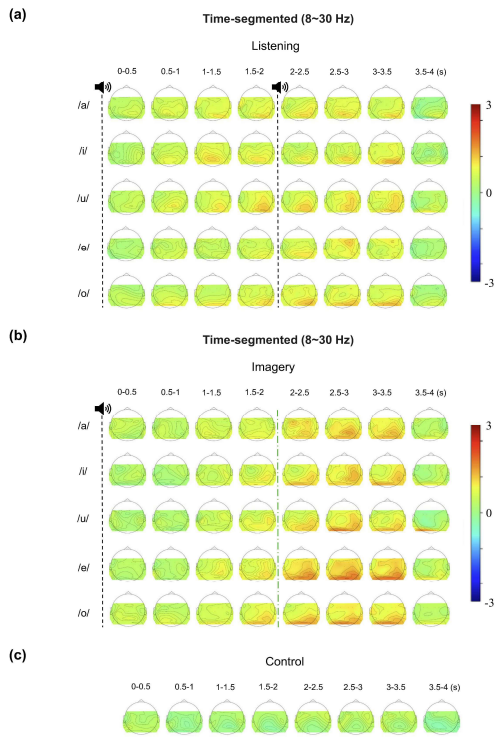

Fig. 4. Time-segmented topographical maps for three conditions: (a) Listening conditions, (b) Imagery conditions, (c) Control condition.

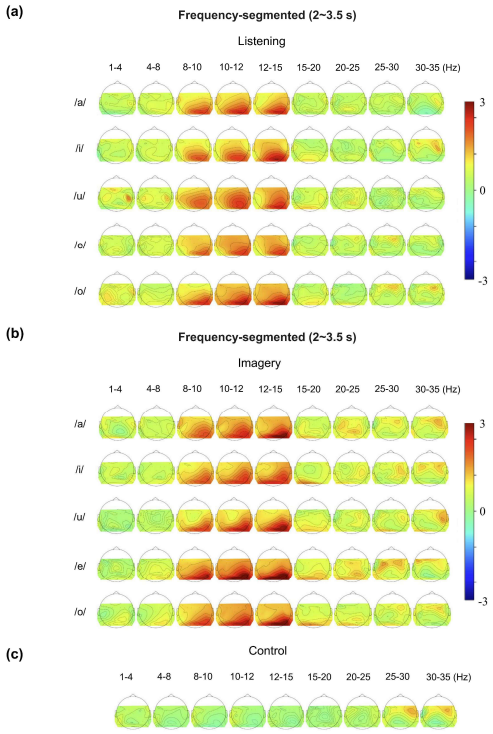

Fig. 5. Frequency-segmented topographical maps for three conditions: (a) Listening conditions, (b) Imagery conditions, (c) Control condition.

right-hemispheric homolog (CPP6h). Following the first auditory stimulus at 0 s, power gradually intensified and peaked around 2 s. The second stimulus further enhanced activity, reaching a maximum between 2.5–3.0 s. Thereafter, task-related effects diminished, and power returned to baseline from 3.5 s onward.

In the *Imagery* condition (Fig. 4(b)), unlike in the *Listening* condition, no significant power changes were observed immediately after the first stimulus, and the topographical maps reflected baseline activity. This pattern may correspond to the preparatory phase for imagery (see Discussion). Starting at 2 s, when participants were required to initiate imagery, clear power increases emerged in the bilateral temporal regions, including Wernicke's area (CPP5h) and its right-hemispheric homolog (CPP6h). These increases intensified during the task interval, peaked between 2.5–3.0 s, and then diminished, returning to baseline from 3.5 s onward. Notably, during the imagery period (2.0–3.5 s), the spatiotemporal patterns of power increase closely resembled those observed in the *Listening* condition.

In the *Control* condition (Fig. 4(c)), by contrast, topographical maps remained near baseline throughout the entire task period.

Based on the time-segmented topographical maps, the 2.0–3.5 s interval was selected as the task-related window for subsequent analyses.

2) Frequency-Segmented Topographical Map: Within the 2.0–3.5 s window, the data were further divided into eight frequency ranges (see Methods) to characterize task-related frequency components in the *Listening* and *Imagery* conditions.

In the *Listening* condition (Fig. 5(a)), pronounced power increases were observed in the low-alpha (8–10 Hz), high-alpha (10–12 Hz), and beta-I (12–15 Hz) bands, with weaker yet consistent activations in the beta-II band (15–20 Hz). These changes were predominantly localized in the temporal regions, particularly around electrode CPP5h (Wernicke's area) and CPP6h, with more potent effects in the right hemisphere. Power modulations above 20 Hz were also present but considerably weaker than those in the alpha and beta ranges.

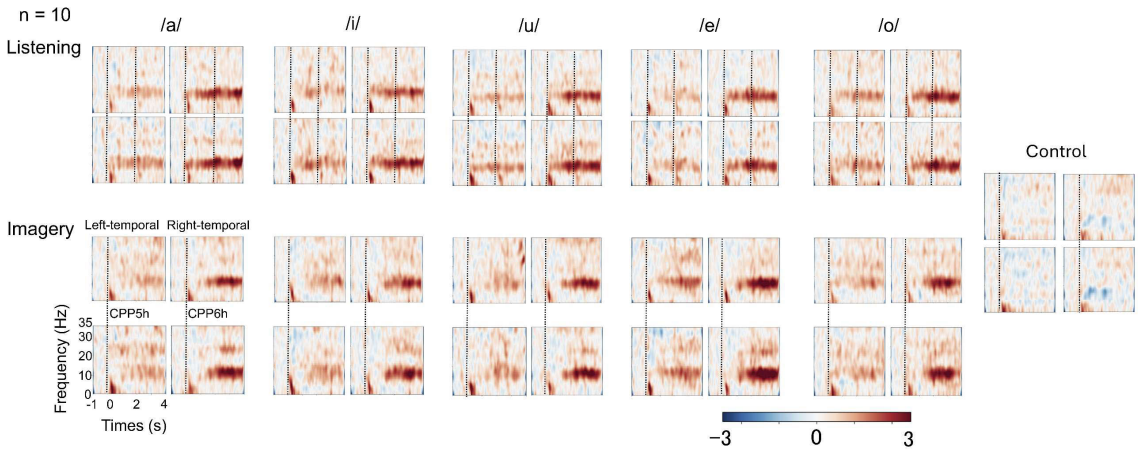

Fig. 6. ERSP results for five vowels under three experimental conditions.

Minimal changes were observed in the theta (4–8 Hz) and Gamma (30–35 Hz) bands. Although vowel-specific variations were present, the overall spectral patterns were largely consistent across vowels.

In the *Imagery* condition (Fig. 5(b)), spectral patterns similar to those in the *Listening* condition were observed. The alpha (8–12 Hz) and beta-I (12–15 Hz) ranges again constituted the primary contributors, with the most pronounced increases occurring in the high-alpha and beta-I bands. As in the *Listening* task, these modulations were localized to bilateral temporal regions, particularly around electrode CPP5h (Wernicke's area) and CPP6h, with a right-hemispheric predominance.

By contrast, in the *Control* condition (Fig. 5(c)), none of the examined frequency bands showed significant power changes.

Taken together, these findings suggest that the alpha and beta-I frequency bands are the dominant contributors to task-related power modulations. The regions of interest (ROIs) for subsequent analyses were defined as the left temporal area, Wernicke's area (CPP5h), the right temporal area, and CPP6h, the right-hemispheric homolog of Wernicke's area.

C. ERSP Analysis

ERSPs were computed within the predefined ROIs as shown in Fig. 6. Electrodes in the left and right temporal regions were averaged to represent the ERSP of each corresponding region. Power increases, hereafter termed event-related synchronization (ERS), are shown in red, whereas power decreases, referred to as event-related desynchronization (ERD), are shown in blue.

In the *Listening* condition, two transient low-frequency ERS responses were observed following each sound onset across all five vowels, consistent with the ERP results. In addition, all vowels exhibited sustained ERS predominantly within the alpha and beta-I frequency ranges (8–15 Hz) throughout the task period, with more vigorous activity in the right hemisphere.

In the *Imagery* condition, transient ERS responses time-locked to the first auditory stimulus were observed. In contrast to the *Listening* condition, no sustained ERS was detected immediately after the first stimulus in the alpha and beta-I ranges. Instead, robust ERS emerged during the imagery interval (~2 s), concentrated in the 10–15 Hz range, showing a clear right-hemispheric predominance. Although shorter in duration than in *Listening*, the sustained ERS patterns observed during the task period (2.0–3.5 s) closely resembled those of the Listening condition.

In the *Control* condition, only transient ERS responses time-locked to the first auditory stimulus were observed, while no sustained ERS was detected throughout the task period. This absence of task-related ERS suggests that the oscillatory activity observed in the *Listening* and *Imagery* conditions reflects task engagement rather than nonspecific auditory responses.

D. Statistical Analysis

Within the task window (2.0–3.5 s), ERSP values were extracted from eight predefined frequency bands and compared across the *Listening*, *Imagery*, and *Control* conditions. Analyses were performed both within the cortical regions defined in Fig. 2 and at the channel level. A one-way repeated-measures ANOVA was applied to evaluate the overall effects. Where significant main effects were identified, post hoc pairwise comparisons were performed to detect condition-specific differences. Multiple comparisons over regions or channels were corrected using the Benjamini–Hochberg (BH) false discovery rate (FDR) method.

1) Brain-Region Analysis: Fig. 7 presents the results for all eight brain regions within the beta-I band (12–15 Hz). Light color denotes *Imagery*, dark color denotes *Listening*, and black denotes *Control*. The y-axis represents the power change relative to baseline, and the x-axis corresponds to the eight brain regions.

When comparing *Imagery* with *Listening*, no statistically significant differences were found across vowels, suggesting that both tasks elicited similar power changes in the beta-I band during the task period. In contrast, significant differences were found when either *Imagery* or *Listening* was compared with the *Control* condition, particularly in the right temporal region and at CPP6h. In the *Control* condition, which involved minimal cognitive engagement, power changes remained near zero across all regions. By comparison, both *Listening* and *Imagery* produced marked power increases relative to *Control*,

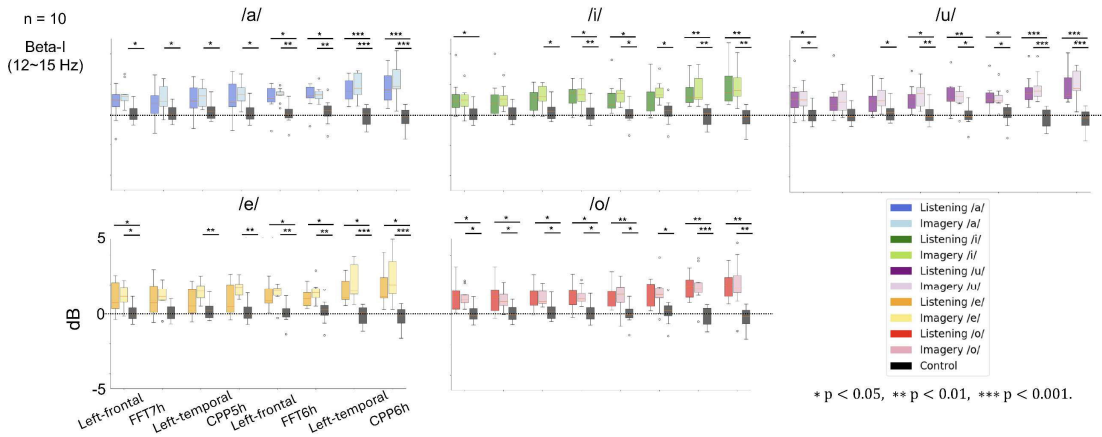

Fig. 7. Statistical analysis in the beta-I frequency band across eight brain regions for five vowels under three conditions.

indicating distinct neural activations within the beta-I band. Similar patterns were observed in the low-alpha and high-alpha bands, whereas effects were less pronounced in other frequency bands. Complete statistics, including p-values for each frequency band, are provided in the Supplementary.

2) Channel-Level Analysis: Channel-level differences were visualized using topographical p-value maps for both time-segmented and frequency-segmented analyses. For the time-segmented analysis, frequency bands were restricted to the alpha and beta-I ranges (8–15 Hz). In the frequency-segmented analysis, the time window was fixed at the task-related interval (2.0–3.5 s). For clarity, the color bar was restricted to 0–0.05, with darker blue indicating greater statistical significance and white denoting non-significance.

Fig. 8 illustrates the time-segmented results. When comparing *Listening* with *Control* (Fig. 8(a)), significant differences were observed from 0.5 s until the end of the task, most prominently in the temporal regions and at electrode CPP5h and CPP6h, with more potent effects in the right hemisphere. This pattern aligns with the two auditory onsets in the *Listening* condition: the first significant interval (~0.5–2.0 s) corresponded to the initial auditory presentation, whereas the second significant interval (~2-3.5 s) reflected the second auditory stimulus. By contrast, no significant differences appeared after the first stimulus when comparing *Imagery* with *Control* (Fig. 8(b)). As imagery was initiated around 2 s, significant differences emerged during the imagery period (2.0–3.5 s), with some vowels showing earlier onsets (~1.5 s), possibly reflecting early initiation of imagery in certain participants (see Discussion). Detailed channel-wise p-values are provided in the Supplementary.

Comparison between *Listening* and *Imagery* (Fig. 8(c)) revealed no significant differences before 2 s, despite the expectation that the two conditions would diverge after the first auditory stimulus. Uncorrected p-values indicated apparent differences around the right temporal region and CPP6h. However, after FDR correction, most values exceeded 0.05, indicating only marginal significance. Consequently, no reliable effects were observed in the topographical maps. Complete p-values are provided in the Supplementary.

Fig. 9 shows the results of the frequency-segmented analysis within the task-related window (2.0–3.5 s).

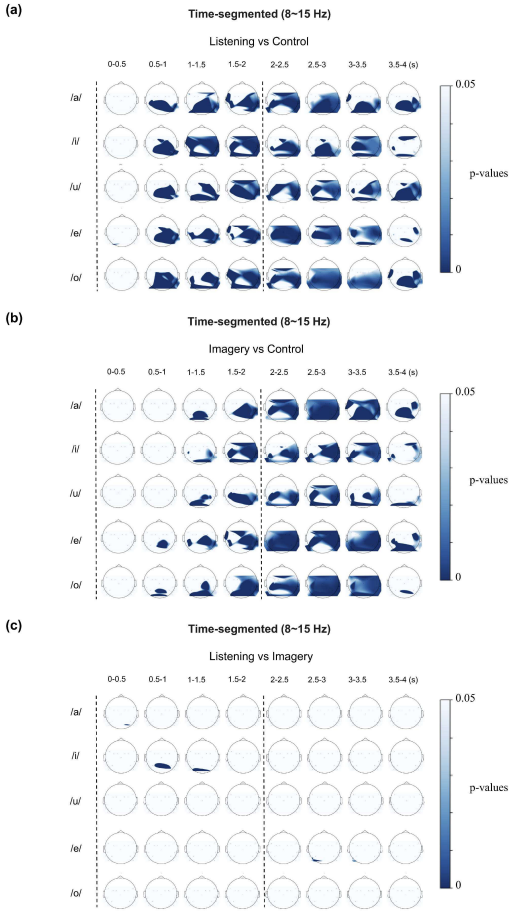

Fig. 8. Time-segmented p-value topographical maps showing results in the frequency range from Alpha to Beta-I for three comparison groups: (a) listening vs. imagery; (b) listening vs. control; (c) imagery vs. control.

When each task was compared with the *Control* condition, significant differences emerged in the low-alpha, high-alpha, and beta-I ranges, particularly in the temporal regions and at electrode CPP5h and CPP6h, with stronger effects in the right hemisphere. Vowel-specific effects included significant differences in beta-II (15–20 Hz) for **/u/** and **/o/** (*Listening*), and for **/u/**, **/e/**, **/o/** (*Imagery*), with **/e/** additionally showing significance in the beta-III band (20–25 Hz). In contrast, no significant

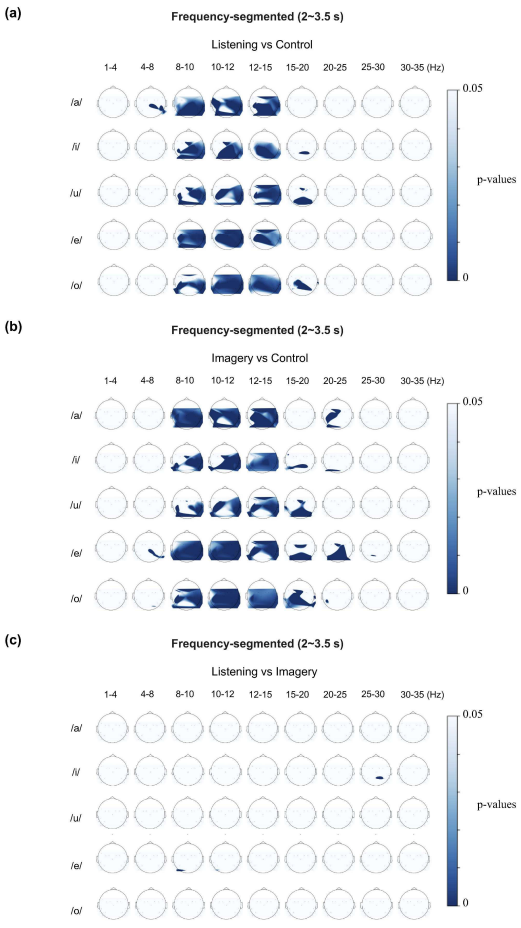

Fig. 9. Frequency-segmented p-value topographical maps showing results in the task-related window (2-3.5 s) for three comparison groups: (a) listening vs. imagery; (b) listening vs. control; (c) imagery vs. control.

differences were found between *Listening* and *Imagery* (Fig. 9(c)). Complete channel-wise statistics are provided in the Supplementary.

Taken together, these findings suggest that *Listening* and *Auditory Imagery* share highly similar neural characteristics, as reflected by the absence of significant differences during the task period. Both tasks elicited significantly greater power increases than the *Control* condition, especially within the alpha and beta-I bands. These effects were most prominent in the temporal regions and at electrodes CPP5h and CPP6h, with more vigorous activity in the right hemisphere.

E. Classification

All classifications were performed in a subject-dependent manner, with one model trained per participant using five-fold cross-validation. Performance was evaluated using macro-averaging across classes with balanced class weights, and results are reported as mean ± SEM across subjects. Detailed subject-wise results, standard deviations, p-values, Cohen's d values, and sample numbers are provided in the Supplementary.

1) 3-Class Classification: A three-class classification task was first conducted, following the design of a previous study [12], using /a/, /i/, and Control as classes. Four baseline

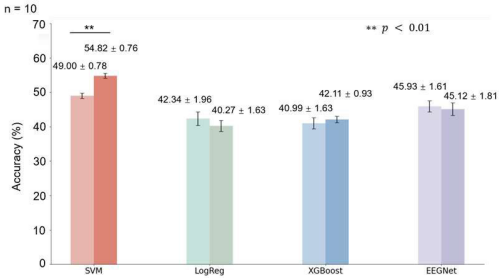

Fig. 10. Three-class classification results of four baseline models using two methods (lighter bars: $I \rightarrow I$ and darker bars: $L \rightarrow I$). Colors indicate classifiers: SVM (red), Logistic Regression (green), XGBoost (blue), and EEGNet (purple).

models were evaluated: Filter Bank Common Spatial Pattern (FBCSP) + Support Vector Machine (SVM), FBCSP + Logistic Regression, Wavelet Scattering Transform (WST) + Kernel Principal Component Analysis (KPCA) + eXtreme Gradient Boosting (XGBoost) [38], and EEGNet. The frequency bands were defined as low-alpha, high-alpha, beta-I, beta-II, beta-III, and beta-IV, while theta and gamma were excluded due to minimal task-related activations in these ranges. To minimize the influence of the first auditory stimulus, only the task-related window (2–3.5 s) was used as input.

Two training–testing paradigms were compared: (i) the conventional approach, training and testing on imagery data (*Imagery*→*Imagery*, $I \rightarrow I$), and (ii) the proposed approach, training on listening and testing on imagery (*Listening*→*Imagery*, $L \rightarrow I$). In $I \rightarrow I$, imagery data were split into five folds (80% training, 20% testing). In $L \rightarrow I$, the same imagery test folds were used, while listening data were also split into five folds, with fold-specific 80% listening data used for training.

Fig. 10 presents the classification accuracy for the two methods ($I \rightarrow I$ and $L \rightarrow I$). All classifiers achieved performance above chance level in both settings. Among them, SVM yielded the highest performance across accuracy, precision, and recall, with accuracies of 49.00 ± 0.78% for $I \rightarrow I$ and 54.82 ± 0.76% for $L \rightarrow I$. Paired t-tests were conducted to compare the two settings within each model, with FDR correction across the four classifiers. No statistically significant differences were observed between $I \rightarrow I$ and $L \rightarrow I$ for any classifier, except for SVM, which showed significant differences in accuracy ($p = 0.0047$, Cohen's $d = 2.01$), precision ($p = 0.020$, Cohen's $d = 1.1$), and recall ($p = 0.00027$, Cohen's $d = 2.75$), with higher performance in $L \rightarrow I$ than in $I \rightarrow I$.

At the individual level, SVM consistently ranked highest. One-sided permutation tests (n = 1000) confirmed that accuracies were statistically significant after FDR correction across all participants ($p < 0.05$).

Overall, models trained solely on listening data achieved accuracies comparable to those reported in a prior study that combined fMRI with EEG [12] (≈ 58%). These results support comparable performance between $L \rightarrow I$ and $I \rightarrow I$.

2) Binary Classification: Next, a binary classification task was conducted on all vowel pairs to examine whether the substitution effect varied across different vowels. The results are shown in Fig. 11. SVM was employed as the classifier,

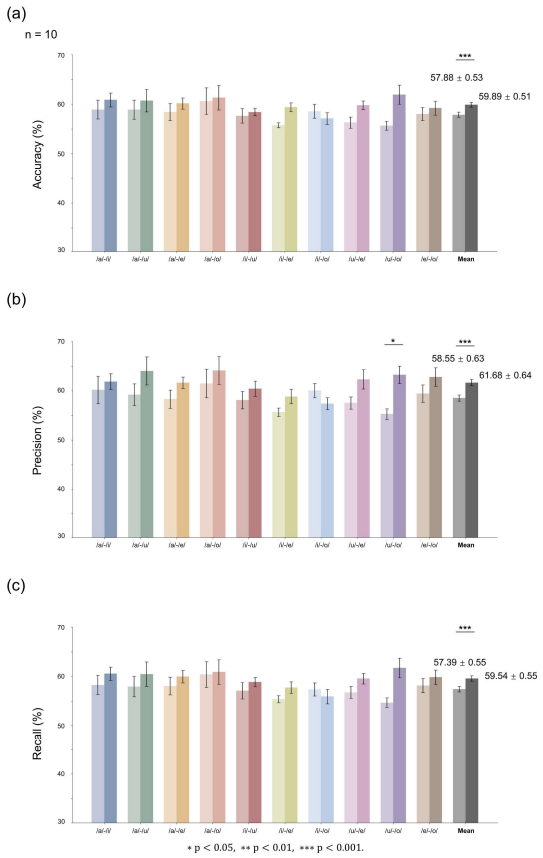

Fig. 11. Binary classification results for all vowel pairs (five vowels, ten pairs): (a) accuracy, (b) precision, and (c) recall. Colors indicate vowel pairs, with black bars representing mean performance; lighter and darker bars denote $I\rightarrow I$ and $L\rightarrow I$, respectively.

with parameters optimized individually using a grid search:

$$C \in \{0.1, 1, 10, 100\}, kernal \in \{linear, rbf, poly\},$$
$$\gamma = scale$$

For each outer fold, hyperparameter optimization was performed using an inner 5-fold cross-validation within the training data only. The held-out test fold was not used during model selection or tuning. Paired t-tests were performed for each vowel pair with FDR correction for all pairs. A separate paired t-test was conducted on the mean values.

For most vowel pairs, no significant differences were observed between the two methods. However, performance varied across specific combinations. Significant differences in accuracy, precision, and recall were found for **/a/–/u/**, **/i/–/e/**, and **/u/–/o/**. Additional effects were observed for **/u/–/e/** in precision and **/i/–/u/** in recall, with the $L\rightarrow I$ outperforming $I\rightarrow I$. At the group level, $L\rightarrow I$ significantly outperformed $I\rightarrow I$ across all metrics: accuracy (59.89 $\pm$ 0.51% vs. 57.88 $\pm$ 0.53%), precision (61.68 $\pm$ 0.64% vs. 58.55 $\pm$ 0.63%), and recall (59.54 $\pm$ 0.55% vs. 57.39 $\pm$ 0.55%). These results suggest that listening data may enhance the classification of imagery signals, although the degree of improvement varies across vowels. Notably, for **/i/–/o/**, $I\rightarrow I$ yielded better performance than $L\rightarrow I$, contrary to the general trend.

To further investigate the effects of $L\rightarrow I$ transfer, improvements were quantified for each vowel pair (Fig. 12). The

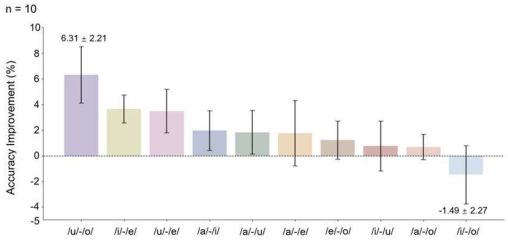

Fig. 12. Improvements in classification results for different vowel pairs, from $I\rightarrow I$ to $L\rightarrow I$.

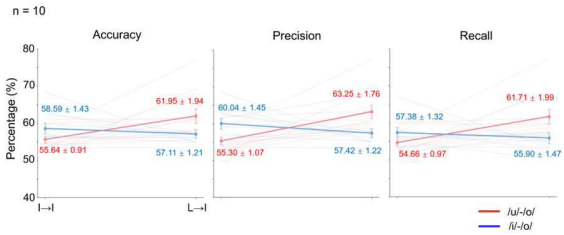

Fig. 13. Individual performance for vowel pairs with the highest improvement (/u/–/o/, red) and the greatest decrease (/i/–/o/, blue) across accuracy, precision, and recall.

color scheme is consistent with Fig. 11. The **/u/–/o/** pair showed the most significant improvement, with gains of 6.31 $\pm$ 2.21% in accuracy, 7.94 $\pm$ 1.94% in precision, and 7.05 $\pm$ 2.14% in recall. These results indicate that, in the binary classification, training on listening data can yield improvements exceeding 5% for specific vowel pairs, representing a substantial enhancement. In contrast, for **/i/–/o/**, listening-based training reduced performance (-1.49 ± 2.27% in accuracy, -2.63 ± 1.97% in precision, and -1.47 ± 2.42% in recall). These findings suggest that the benefit of $L\rightarrow I$ transfer may depend on vowel-specific acoustic features.

Finally, individual performances were examined for two vowel pairs:**/u/–/o/**, which showed the largest improvement, and **/i/–/o/**, which showed a decrease. The results are illustrated in Fig. 13. For **/u/–/o/**, all participants exhibited improvements when trained on listening data, suggesting that listening enhanced model performance consistently across individuals. In contrast, for **/i/–/o/**, most participants showed decreased performance under the same condition. These patterns may be related to vowel-specific phonetic characteristics in the auditory domain (see Discussion).

Taken together, the binary classification results suggest that *Listening* can serve as an effective surrogate for *Imagery* in model training and, in some cases, may even enhance classification performance. However, this effect appears to depend on vowel-specific characteristics.

3) 5-Class Classification: We further extended the analysis to five-class vowel classification as an exploratory analysis. Due to the increased complexity of this task, SVM did not achieve performance above the chance level. Therefore, we employed EEGNet and introduced a data augmentation step before model input. Specifically, the train–test split was first performed at the trial level, after which a 500-sample (1 s) sliding window with a stride of 125 samples (0.25 s) was applied separately within the training and test sets. In addition to $I\rightarrow I$ and $L\rightarrow I$, we incorporated a variational autoencoder (VAE) for signal reconstruction, following the approach of

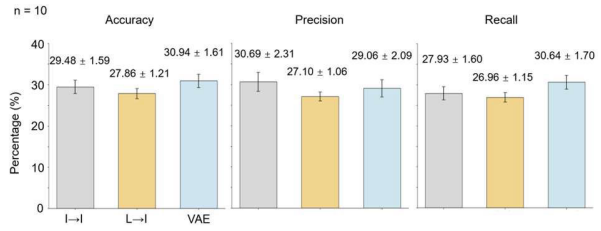

Fig. 14. Five-class classification results of three methods (*I*→*I*, *L*→*I*, and VAE) across accuracy, precision, and recall. Gray, yellow, and blue bars denote *I*→*I*, *L*→*I*, and VAE, respectively.

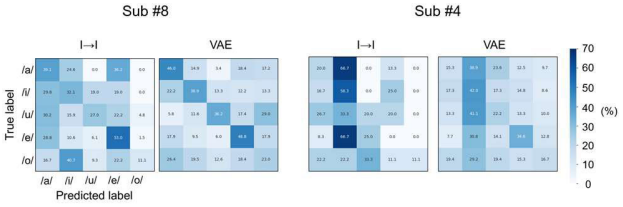

Fig. 15. Confusion matrices for the best-performing and worst-performing subjects using three methods (*I*→*I*, *L*→*I*, and VAE).

a previous study [39]. Whereas the prior work focused on reconstructing motor execution signals from motor imagery, we adapted the method to reconstruct listening signals from auditory imagery. In this framework, the training set consisted of both reconstructed and imagery signals. To limit sample-size imbalance, the VAE stream used a 500-sample window length and a 250-sample (0.5 s) stride. Five-fold cross-validation was applied, with identical test sets across methods. Detailed information on model architectures and training parameters is provided in the Supplementary.

The results of the five-class classification are shown in Fig. 14. All methods performed above chance level, with the VAE model achieving the highest accuracy (30.94 ± 1.61%). However, no significant differences were observed in any pairwise comparisons. McNemar's tests for the VAE model indicated $p < 0.05$ for all participants except Subject 4 ($p = 0.298$), Subject 5 ($p = 0.085$) and Subject 10 ($p = 0.053$). P-values were FDR-corrected across all participants. Confusion matrices comparing *I*→*I* and VAE models for the best (Subject 8) and worst (Subject 4) subjects further illustrate this variability (Fig. 15).

Overall, the five-class results should be interpreted as exploratory rather than as primary evidence supporting the proposed paradigm. Given the small sample size and the complexity of five-class decoding, the present dataset may be insufficient to train models capable of achieving high and consistent classification accuracy. Such variability and five-class complexity highlight both individual differences and the need for more robust, generalizable models for auditory imagery decoding (see Discussion).

IV. Discussion

A. Mechanistic Interpretation of L→I Improvement

In the 3-class classification using SVM, *L*→*I* significantly outperformed *I*→*I* (54.82% vs. 49.00%, p = 0.0047, Cohen's d = 2.01). This result indicates that, under certain conditions, *L*→*I* transfer is not merely comparable to conventional

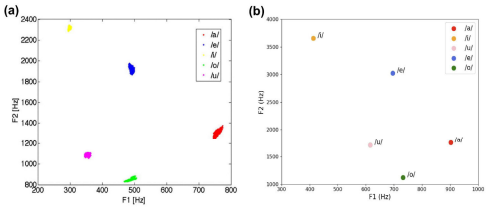

Fig. 16. Auditory formants F1 and F2 for the five vowels: (a) results from a previous study [40]; (b) our stimuli.

imagery-based training but may improve classification performance. This finding is theoretically important because it suggests that models trained on a different but related cognitive state can capture neural features useful for decoding auditory imagery.

A possible explanation is that listening trials provide more reliable training data than imagery trials. Although listening and auditory imagery may share partially overlapping neural representations, listening is externally driven and may therefore produce responses with a higher signal-to-noise ratio, greater temporal consistency, and lower intra-subject variability. In contrast, auditory imagery is internally generated and may vary across trials depending on imagery vividness, attention, and strategy. Thus, training on listening data may allow the classifier to learn more robust related features that can be transferred to imagery decoding.

The binary classification results further suggest that this benefit may depend on vowel-specific acoustic characteristics. The largest improvement was observed for the /u/–/o/ pair, which represents acoustically similar vowels, whereas the smallest improvement was observed for the /i/–/o/ pair, which has more distinct formant separation in the F1–F2 space shown in Fig. 16. This pattern suggests that listening-based training may be particularly helpful when imagery-based features are weak and vowel categories are acoustically close, while leaving less room for improvement when vowels are already acoustically distinct.

Future studies should directly examine this hypothesis by quantifying signal-to-noise ratio, trial-to-trial variability, and feature-space similarity between listening and imagery conditions, and by systematically comparing acoustic distances with classification outcomes.

B. Active and Reactive BCI Systems

This study explored the use of listening signals as a surrogate for auditory imagery, effectively combining a reactive approach with an active paradigm. Reactive BCIs, such as those based on P300 or steady-state visual evoked potentials (SSVEP), are often considered easier to use because they require less training or calibration [16], [41]. However, their reliance on external stimuli for every output limits flexibility and independence, reducing practicality in daily communication [1]. By contrast, active BCIs require more intensive pre-training but ultimately provide greater autonomy. This trade-off between usability and independence remains a central challenge for translating BCI research into real-world applications [18]. Our findings suggest that listening signals may serve as a reactive proxy to enhance the decoding of

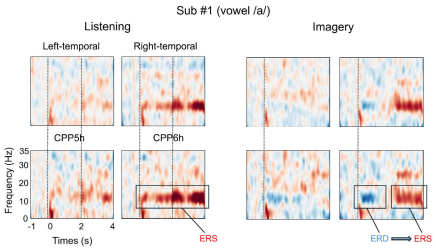

Fig. 17. Power suppression in Subject 1 before the onset of imagery, shifting to ERS during the imagery task.

active auditory imagery, offering a promising direction for future hybrid BCI applications.

C. Power Suppression

In both the time-segmented topographical and ERSP analyses, imagery showed little power modulation during 0–2.0 s, which we interpret as preparatory power suppression. Prior studies reported suppression of auditory cortical activity preceding spontaneous speech [42] and an α ERD that transitions to α ERS extending into low-β during speech production [43]; similar suppression has also been observed in inner speech [24]. Our data exhibit a comparable α ERD→ERS transition at imagery onset.

Fig. 17 illustrates the ERSP results for Subject 1 on the T7 channel. During *Listening*, ERS was observed across the entire task, reflecting auditory perceptual features. By contrast, an ERD was evident during Imagery before 2 s, which then transitioned to ERS as imagery began. At the group-level, the *Imagery* condition did not exhibit pronounced ERS immediately after the first external stimulus, likely reflecting the overlap of two opposing processes: ERS driven by the auditory perception and ERD driven by power suppression. Although prior studies investigated overt speech [42], [43] or inner speech [24], the similarity of the observed pattern suggests that power suppression may represent a fundamental mechanism of auditory processing, extending beyond motor-related tasks to auditory imagery.

D. Hemispherical Dominance

Hemispheric lateralization plays a role in auditory and speech imagery, but—unlike hand dominance, which is easily assessed—auditory or speech dominance is more challenging to determine. Prior studies have reported individual variability in lateralization within language networks [44], [45], and candidate determinants such as sex and handedness remain inconclusive [46].

In our individual-level analyses, hemispheric differences were observed but were less pronounced than those typically reported in motor imagery dominance. No participants displayed opposite lateralization; all exhibited bilateral engagement, with most showing relatively stronger activation in the right hemisphere. This pattern is consistent with a previous report [47], which suggests that the left hemisphere is more involved in auditory perception and temporal processing, whereas the right hemisphere is more engaged in timbre and spectral processing.

Our task design may have accentuated this effect. To minimize semantic influences, we used Japanese vowels rather than full sentences, thereby focusing on acoustic features. In addition, male participants listened to or imagined a female voice, which likely emphasized timbre-related processing. These factors may account for the stronger right-hemispheric activations observed in our study.

E. Future Directions

As proof of concept, several avenues remain open for future research. First, future studies may explore larger datasets and more advanced architectures, such as contrastive learning [48], [49] or gradient reversal layers (GRL) [50], to enhance feature learning. Second, extending beyond the within-subject design, cross-subject validation would help improve generalizability and address inter-individual variability, which remains a central challenge in BCI research. Third, refining experimental protocols to achieve more precise alignment of imagery onset could improve temporal precision, especially for complex models. Finally, expanding investigations beyond vowels to include consonants would provide a more comprehensive understanding of speech-related auditory imagery.

V. Limitations

This study has several limitations. First, the sample size was small (N = 10), and all participants were healthy young male individuals from a single institution and within a narrow age range. In addition, the study employed a within-subject design. These factors limit the generalizability of the findings, particularly to clinical populations such as individuals with locked-in syndrome (LIS).

Second, all participants were instructed to imagine a female voice. While this design helps standardize the imagery target and reduce variability, it may limit generalizability across sexes, voice types (e.g., male voices), and self-voice imagery.

Third, for the purpose of neuroscientific characterization, topographical maps were computed using the entire dataset, which is commonly used for visualizing overall neural response patterns. However, the classification time window (2.0–3.5 s) was subsequently selected based on these time-segmented topographical maps derived from the same dataset. This procedure introduces a methodological limitation, as it may lead to data leakage or optimistic bias due to the use of the same dataset for both feature selection and model evaluation.

Finally, the dataset size and model capacity may have been insufficient for more complex tasks such as five-class classification, which showed substantial inter-subject variability and only modest performance above chance level. This may be due to the increased task complexity, limited dataset size, and potential temporal misalignment of imagery onset across trials, which could affect feature learning, particularly in short analysis windows.

VI. Conclusion

We presented a hybrid training paradigm for EEG speech–imagery BCIs, training on passive listening and testing on auditory imagery. Unlike prior studies [5], [6], [7], [8], [9], [10], [11], [12], [13], [14] that relied solely on imagery

data, our approach reduces reliance on imagery collection. To our knowledge, this is the first EEG study to systematically validate Listening→Imagery decoding across multiple vowels, combining neural characterization with classifier evaluation. Our results indicate that (i) *Listening* and *Auditory Imagery* share similar activation patterns, suggesting overlapping neural representations; (ii) the $L \rightarrow I$ achieves comparable or better classification performance than the conventional $I \rightarrow I$, suggesting that listening may serve as a surrogate for auditory imagery in EEG-based BCI training; and (iii) the improvement observed in the binary classification setting varies across vowels, suggesting that decoding performance may depend on vowel-specific auditory features.

These contributions provide proof-of-concept evidence that listening-based training may serve as a practical alternative to imagery-only calibration within the tested setting, reducing the pre-training burden that hinders active BCI adoption. However, as a first step, this study was conducted with a small sample of healthy young male participants from a single institution and within a narrow age range. Therefore, future work should validate this approach in larger, more diverse cohorts, including target clinical populations, and further explore stronger models, improved temporal alignment, and cross-subject generalization to enhance its applicability in clinical and assistive settings.

REFERENCES

[1] M. P. Branco et al., "Brain–computer interfaces for communication: Preferences of individuals with locked-in syndrome," *Neurorehabil. Neural Repair*, vol. 35, no. 3, pp. 267–279, Mar. 2021, doi: 10.1177/1545968321989331.

[2] R. Abiri, S. Borhani, E. W. Sellers, Y. Jiang, and X. Zhao, "A comprehensive review of EEG-based brain–computer interface paradigms," *J. Neural Eng.*, vol. 16, no. 1, Jan. 2019, Art. no. 011001, doi: 10.1088/1741-2552/aaf12e.

[3] M. L. Martini, E. K. Oermann, N. L. Opie, F. Panov, T. Oxley, and K. Yaeger, "Sensor modalities for brain–computer interface technology: A comprehensive literature review," *Neurosurgery*, vol. 86, no. 2, pp. E108–E117, Feb. 2020, doi: 10.1093/neuros/nyz286.

[4] M. Soufineyestani, D. Dowling, and A. Khan, "Electroencephalography (EEG) technology applications and available devices," *Appl. Sci.*, vol. 10, no. 21, p. 7453, Oct. 2020, doi: 10.3390/app10217453.

[5] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, "Online EEG classification of covert speech for brain–computer interfacing," *Int. J. Neural Syst.*, vol. 27, no. 8, Dec. 2017, Art. no. 1750033, doi: 10.1142/s0129065717500332.

[6] S. Deng, R. Srinivasan, T. Lappas, and M. D'Zmura, "EEG classification of imagined syllable rhythm using Hilbert spectrum methods," *J. Neural Eng.*, vol. 7, no. 4, Aug. 2010, Art. no. 046006, doi: 10.1088/1741-2560/7/4/046006.

[7] L. Wang, X. Zhang, X. Zhong, and Y. Zhang, "Analysis and classification of speech imagery EEG for BCI," *Biomed. Signal Process. Control*, vol. 8, no. 6, pp. 901–908, Nov. 2013, doi: 10.1016/j.bspc.2013.07.011.

[8] S. Alzahrani, H. Banjar, and R. Mirza, "Systematic review of EEG-based imagined speech classification methods," *Sensors*, vol. 24, no. 24, p. 8168, Dec. 2024, doi: 10.3390/s24248168.

[9] M. M. Abdulghani, W. L. Walters, and K. H. Abed, "Imagined speech classification using EEG and deep learning," *Bioengineering*, vol. 10, no. 6, p. 649, May 2023, doi: 10.3390/bioengineering10060649.

[10] M. M. Rahman, F. Khatun, S. I. Sami, and A. Uzzaman, "The evolving roles and impacts of 5G enabled technologies in healthcare: The world epidemic COVID-19 issues," *Array*, vol. 14, Jul. 2022, Art. no. 100178, doi: 10.1016/j.array.2022.100178.

[11] H.-J. Park and B. Lee, "Multiclass classification of imagined speech EEG using noise-assisted multivariate empirical mode decomposition and multireceptive field convolutional neural network," *Frontiers Hum. Neurosci.*, vol. 17, Aug. 2023, Art. no. 1186594, doi: 10.3389/fnhum.2023.1186594.

[12] N. Yoshimura et al., "Decoding of covert vowel articulation using electroencephalography cortical currents," *Frontiers Neurosci.*, vol. 10, p. 175, May 2016, doi: 10.3389/fnins.2016.00175.

[13] W. Akashi, H. Kambara, Y. Ogata, Y. Koike, L. Minati, and N. Yoshimura, "Vowel sound synthesis from electroencephalography during listening and recalling," *Adv. Intell. Syst.*, vol. 3, no. 2, Feb. 2021, Art. no. 2000164, doi: 10.1002/aisy.202000164.

[14] M. Matsumoto and J. Hori, "Classification of silent speech using support vector machine and relevance vector machine," *Appl. Soft Comput.*, vol. 20, pp. 95–102, Jul. 2014, doi: 10.1016/j.asoc.2013.10.023.

[15] J. Mladenović, J. Mattout, and F. Lotte, "A generic framework for adaptive EEG-based BCI training and operation," in *Brain–Computer Interfaces Handbook*, 1st ed. Boca Raton, FL, USA: CRC Press, 2018.

[16] J. Xie, G. Xu, J. Wang, M. Li, C. Han, and Y. Jia, "Effects of mental load and fatigue on steady-state evoked potential based brain computer interface tasks: A comparison of periodic flickering and motion-reversal based visual attention," *PLoS ONE*, vol. 11, no. 9, Sep. 2016, Art. no. e0163426, doi: 10.1371/journal.pone.0163426.

[17] T. Dickhaus, C. Sannelli, K.-R. Müller, G. Curio, and B. Blankertz, "Predicting BCI performance to study BCI illiteracy," *BMC Neurosci.*, vol. 10, no. S1, p. 84, Jul. 2009, doi: 10.1186/1471-2202-10-s1-p84.

[18] S. Saha et al., "Progress in brain–computer interface: Challenges and opportunities," *Frontiers Syst. Neurosci.*, vol. 15, Apr. 2021, Art. no. 578875, doi: 10.3389/fnsys.2021.578875.

[19] M. Stephane, M. Dzemidzic, and G. Yoon, "Keeping the inner voice inside the head, a pilot fMRI study," *Brain Behav.*, vol. 11, no. 4, p. 2042, Apr. 2021, doi: 10.1002/brb3.2042.

[20] C. A. Kell, M. Darquea, M. Behrens, L. Cordani, C. Keller, and S. Fuchs, "Phonetic detail and lateralization of reading-related inner speech and of auditory and somatosensory feedback processing during overt reading," *Hum. Brain Mapping*, vol. 38, no. 1, pp. 493–508, Jan. 2017, doi: 10.1002/hbm.23398.

[21] H. Lee Masson and L. Isik, "Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing," *NeuroImage*, vol. 245, Dec. 2021, Art. no. 118741, doi: 10.1016/j.neuroimage.2021.118741.

[22] S. S. Shergill, E. T. Bullmore, M. J. Brammer, S. C. R. Williams, R. M. Murray, and P. K. McGuire, "A functional study of auditory verbal imagery," *Psychol. Med.*, vol. 31, no. 2, pp. 241–253, Apr. 2001, doi: 10.1017/s003329170100335x.

[23] X. Tian, J. M. Zarate, and D. Poeppel, "Mental imagery of speech implicates two mechanisms of perceptual reactivation," *Cortex*, vol. 77, pp. 1–12, Apr. 2016, doi: 10.1016/j.cortex.2016.01.002.

[24] T. J. Whitford, "Speaking-induced suppression of the auditory cortex in humans and its relevance to schizophrenia," *Biol. Psychiatry: Cognit. Neurosci. Neuroimag.*, vol. 4, no. 9, pp. 791–804, Sep. 2019, doi: 10.1016/j.bpsc.2019.05.011.

[25] D. Zapała, P. Iwanowicz, P. Francuz, and P. Augustynowicz, "Handedness effects on motor imagery during kinesthetic and visual-motor conditions," *Sci. Rep.*, vol. 11, no. 1, p. 13112, Jun. 2021, doi: 10.1038/s41598-021-92467-7.

[26] E. Blagovechtchenski, D. Gnedykh, D. Kurmakaeva, N. Mkrtychian, S. Kostromina, and Y. Shtyrov, "Transcranial direct current stimulation (tDCS) of Wernicke's and Broca's areas in studies of language learning and word acquisition," *J. Visualized Exp.*, no. 149, Jul. 2019, Art. no. e59159, doi: 10.3791/59159.

[27] University of Washington.*Lab 8: Auditory, Vestibular, and Visual Systems*. Rehabilitation 551: Human Physiology and Anatomy. Accessed: Jun. 19, 2026. [Online]. Available: https://uw.pressbooks.pub/rehab551/chapter/lab-8-auditory-vestibular-and-visual-systems/

[28] C. Jorgensen and S. Dusan, "Speech interfaces based upon surface electromyography," *Speech Commun.*, vol. 52, no. 4, pp. 354–366, Apr. 2010, doi: 10.1016/j.specom.2009.11.003.

[29] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004, doi: 10.1016/j.jneumeth.2003.10.009.

[30] S. J. Luck, *An Introduction to the Event-Related Potential Technique*, 2nd ed. Cambridge, MA, USA: MIT Press, 2014.

[31] S. H. Patel and P. N. Azzam, "Characterization of N200 and p300: Selected studies of the event-related potential," *Int. J. Med. Sci.*, vol. 2, no. 4, pp. 147–154, 2005, doi: 10.7150/ijms.2.147.

[32] R. Näätänen, "Processing negativity: An evoked-potential reflection," *Psychol. Bull.*, vol. 92, no. 3, pp. 605–640, 1982, doi: 10.1037/0033-2909.92.3.605.

[33] P. T. Michie, A. M. Fox, P. B. Ward, S. V. Catts, and N. McConaghy, "Event-related potential indices of selective attention and cortical lateralization in schizophrenia," *Psychophysiology*, vol. 27, no. 2, pp. 209–227, Mar. 1990, doi: 10.1111/j.1469-8986.1990.tb00372.x.

[34] R. Wennberg and A. M. Lozano, "Restating the importance of bipolar recording in subcortical nuclei," *Clin. Neurophysiol.*, vol. 117, no. 2, pp. 474–475, Feb. 2006, doi: 10.1016/j.clinph.2005.09.020.

[35] V. Mueller, Y. Brehmer, T. von Oertzen, S.-C. Li, and U. Lindenberger, "Electrophysiological correlates of selective attention: A lifespan comparison," *BMC Neurosci.*, vol. 9, no. 1, p. 18, Jan. 2008, doi: 10.1186/1471-2202-9-18.

[36] A. Bartha-Doering, S. Deuster, M. Giordano, M. Trinka, and C. Dehme, "A Signature of passivity? An explorative study of the N3 event-related potential component in passive oddball tasks," *Frontiers Neurosci.*, vol. 13, p. 365, Apr. 2019, doi: 10.3389/fnins.2019.00365.

[37] B. Rozenkrants and J. Polich, "Affective ERP processing in a visual oddball task: Arousal, valence, and gender," *Clin. Neurophysiol.*, vol. 119, no. 10, pp. 2260–2265, Oct. 2008, doi: 10.1016/j.clinph.2008.07.213.

[38] H. Pan, Y. Wang, Z. Li, X. Chu, B. Teng, and H. Gao, "A complete scheme for multi-character classification using EEG signals from speech imagery," *IEEE Trans. Biomed. Eng.*, vol. 71, no. 8, pp. 2454–2462, Aug. 2024, doi: 10.1109/TBME.2024.3376603.

[39] D.-Y. Lee, J.-H. Jeong, B.-H. Lee, and S.-W. Lee, "Motor imagery classification using inter-task transfer learning via a channel-wise variational autoencoder-based convolutional neural network," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 30, pp. 226–237, 2022, doi: 10.1109/TNSRE.2022.3143836.

[40] M. Funabashi, "Invariance in vowel systems," *J. Acoust. Soc. Amer.*, vol. 137, no. 5, pp. 2892–2900, May 2015, doi: 10.1121/1.4919360.

[41] E. M. Hammer, S. Halder, S. C. Kleih, and A. Kübler, "Psychological predictors of visual and auditory P300 brain–computer interface performance," *Frontiers Neurosci.*, vol. 12, p. 307, May 2018, doi: 10.3389/fnins.2018.00307.

[42] A. Flinker, E. F. Chang, H. E. Kirsch, N. M. Barbaro, N. E. Crone, and R. T. Knight, "Single-trial speech suppression of auditory cortex activity in humans," *J. Neurosci.*, vol. 30, no. 49, pp. 16643–16650, Dec. 2010, doi: 10.1523/jneurosci.1809-10.2010.

[43] D. Jenson, A. W. Harkrider, D. Thornton, A. L. Bowers, and T. Saltuklaroglu, "Auditory cortical deactivation during speech production and following speech perception: An EEG investigation of the temporal dynamics of the auditory alpha rhythm," *Frontiers Hum. Neurosci.*, vol. 9, p. 534, Oct. 2015, doi: 10.3389/fnhum.2015.00534.

[44] A. E. Davis and J. A. Wada, "Speech dominance and handedness in the normal human," *Brain Lang.*, vol. 5, no. 1, pp. 42–55, Jan. 1978, doi: 10.1016/0093-934x(78)90006-8.

[45] J. Sidney, *Language Functions and Brain Organization*. St. Catharines, ON, Canada: Brock Univ. Press, 1983.

[46] Y. S. Sininger and A. Bhatara, "Laterality of basic auditory perception," *Laterality: Asymmetries Body, Brain Cognition*, vol. 17, no. 2, pp. 129–149, Mar. 2012, doi: 10.1080/1357650x.2010.541464.

[47] S. Luthra, "The role of the right hemisphere in processing phonetic variability between talkers," *Neurobiol. Lang.*, vol. 2, no. 1, pp. 138–151, Feb. 2021, doi: 10.1162/nol_a_00028.

[48] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, Vienna, Austria, Jul. 2020, pp. 1597–1607.

[49] T. Akama, Z. Zhang, P. Li, K. Hongo, S. Minamikawa, and N. Polouliakh, "Predicting artificial neural network representations to learn recognition model for music identification from brain recordings," *Sci. Rep.*, vol. 15, no. 1, p. 18869, May 2025, doi: 10.1038/s41598-025-02790-6.

[50] K. Avramidis, C. Garoufis, A. Zlatintsi, and P. Maragos, "Enhancing affective representations of music-induced EEG through multimodal supervision and latent domain adaptation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 4588–4592, doi: 10.1109/ICASSP43922.2022.9746643.