

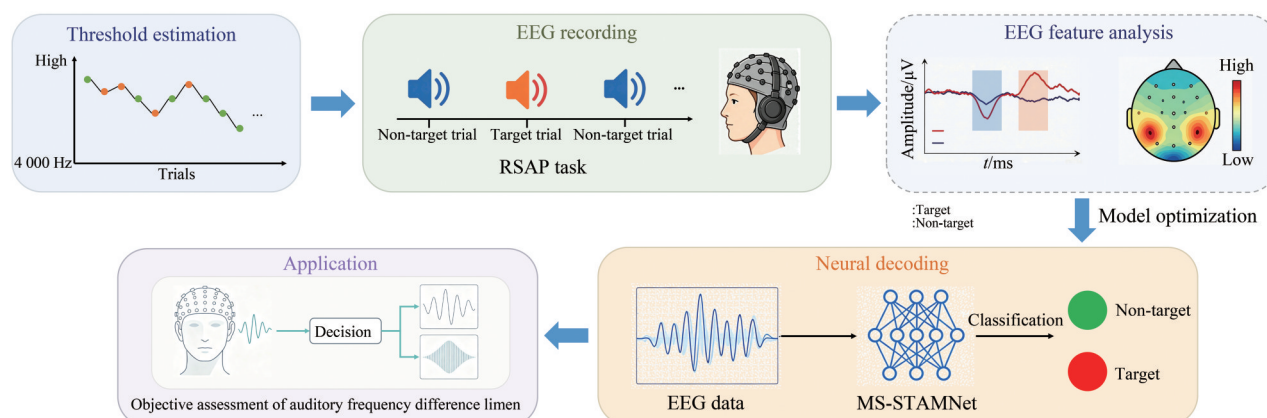


A Personalized Brain-computer Interface Paradigm and Decoding Method for The Objective Evaluation of Auditory Frequency Difference Limen^{△,*}

LI Sheng-Ye, XIAO Xiao-Lin**, YU Shi-Hang, ZHANG Bei-Bei, AN Xing-Wei,
 XU Min-Peng, MING Dong

(Medical School, Tianjin University, Tianjin 300072, China)

Graphical abstract



Abstract Objective The frequency difference limen (FDL) serves as a fundamental metric utilized for effectively quantifying the precise perceptual capabilities of the central auditory system. However, traditional measurement methods rely heavily on the active behavioral responses of subjects and are consequently highly susceptible to the negative influence of confounding subjective factors. Furthermore, existing research paradigms frequently employ uniform stimulus configurations that overlook critical individual perceptual differences. Based on brain-computer interface (BCI) technology, this comprehensive study aims to establish an objective and quantitative evaluation method for auditory frequency discrimination by systematically analyzing and decoding the specific neural responses elicited at the exact threshold state. **Methods** We designed a personalized rapid serial auditory presentation (RSAP) paradigm customized based on each individual's precise FDL. A cohort of eleven healthy participants was recruited to evaluate the paradigm using pure-tone sequences at a baseline frequency of 4 000 Hz. This experimental paradigm simulates a realistic auditory perception environment through the continuous presentation of acoustic stimuli, thereby allowing for an in-depth investigation into the specific neural representations evoked by weak frequency deviations at the threshold state. Given that auditory stimulus-evoked response features exhibit complex and differentiated spatiotemporal distribution patterns across multiple frequency domains, this study further deeply integrates the cross-scale feature interaction module with the dynamic spatiotemporal attention allocation strategy, innovatively proposing the Multi-Scale Spatial-Temporal Dual Attention Network (MS-STAMNet). Specifically, the network constructs parallel processing branches with multiple receptive fields and introduces a dynamic adaptive

[△] Academic papers of the 28th Annual Meeting of the China Association for Science and Technology.

* This work was supported by a grant from The National Natural Science Foundation of China (62106170).

** Corresponding author.

Tel: 86-13752345268, E-mail: xiaoxiao0@tju.edu.cn

Received: May 10, 2026 Accepted: July 6, 2026

weighting strategy to precisely localize core neural activity signals, further deeply integrating multi-scale information through cross-branch feature information interaction to achieve robust single-trial decoding of weak auditory evoked responses. **Results** The comprehensive electrophysiological data analysis demonstrated that subtle auditory frequency deviation stimuli presented at the threshold level successfully elicited pronounced N2 and P3 event-related potential features, reflecting pre-attentive mismatch detection and subsequent cognitive evaluation, which were prominently distributed over the frontal, central, and temporal regions of the scalp. In the complex time-frequency domain, the extracted neural response characteristics exhibited distinct, statistically significant event-related synchronization within both the low-frequency δ and θ frequency bands, which was simultaneously accompanied by a widespread, prominent event-related desynchronization within the higher α band. A comparative analysis of model performance demonstrated that MS-STAMNet achieved an average unweighted average recall (*UAR*) of $(69.67 \pm 6.12)\%$ and area under the curve (*AUC*) of 0.7618 ± 0.07 , significantly outperforming the established baseline models such as EEGNet and PLNet. Furthermore, a distinct dissociation phenomenon was verified between neural decoding and behavioral performance through regression analysis ($R^2=0.016$, $P=0.709$), indicating that this model can effectively capture the implicit features of subtle frequency deviations, even when they fail to trigger explicit conscious responses. Additionally, attention weight visualization analysis further reveals the highly accurate focus of the network on key features concentrated over the bilateral temporal and fronto-parietal regions. **Conclusion** This study systematically and comprehensively uncovers the multi-dimensional spatiotemporal evolutionary patterns of complex neural responses processing subtle acoustic variations under long-sequence threshold auditory stimulation. Concurrently, it verifies the efficacy and robustness of the proposed MS-STAMNet architecture in accurately deciphering weak, single-trial electroencephalogram signals amidst complex background noise. Ultimately, these neurophysiological and algorithmic findings lay a solid theoretical and methodological foundation for the objective and quantitative evaluation of individual auditory cognitive capabilities in clinical applications, transcending the fundamental limitations of traditional behavioral paradigms and providing robust technical support for future auditory research and related clinical assessments.

Key words frequency difference discrimination, brain-computer interface, rapid serial auditory presentation, personalization, deep learning

DOI: 10.3724/j.pibb.2026.0252

CSTR: 32369.14.pibb.20260252

Frequency is an inherent physical property of sound. As the primary determinant of pitch^[1], it serves as a critical cue for sound source localization, target identification, and masking release in complex auditory scenes^[2-4]. The perceptual ability to discriminate frequency differences provides a cognitive foundation for activities such as speech communication, emotional expression, and musical creation^[5-6], and is typically quantified using the frequency difference limen (FDL)^[7-8]. As a crucial parameter reflecting the efficacy of auditory nervous system, the FDL provides specific metrics for exploring the developmental trajectories of the central auditory system and the mechanisms underlying cognitive processing in the brain; furthermore, it serves as a clinical diagnostic criterion for related auditory and cognitive impairments^[9]. Traditional psychophysical measurements frequently employ alternative forced-choice (AFC) tasks, fitting behavioral accuracy into psychometric functions to extract discrimination thresholds at a target probability^[10-11]. To enhance testing efficiency, adaptive procedures, notably the weighted up-down method^[12], have been widely applied. However, these paradigms rely on the active behavioral responses of subjects, making the measurements highly susceptible to subjective factors such as attentional fluctuations, variations in cognitive load, fatigue, and the subject's comprehension of task instructions.

Brain-computer interface (BCI) is a cutting-edge

technology that establishes a direct communication and control pathway between the brain and external devices. By acquiring and decoding electrophysiological signals from the central nervous system, BCIs can objectively interpret user intentions or cognitive states. Analyzing event-related potential (ERP) responses within electroencephalogram (EEG) signals allows for the accurate characterization of the brain's automatic processing and synchronous tracking of frequency differences, thereby providing a novel approach for the objective quantification of individual frequency discrimination capabilities. Gu *et al.*^[13] demonstrated a significant correlation between the amplitude of mismatch negativity (MMN) and discrimination ability. Furthermore, Avsar Aksu *et al.*^[14] verified that MMN amplitude remains unaffected by the subjective state of the participants, suggesting its potential to replace behavioral tests for the objective evaluation of thresholds. Regarding the frequency-following response (FFR) as a neural representation signal at the subcortical level, Xu *et al.*^[15] revealed that the FFR not only accurately reflects an individual's FDL but also detects frequency stimulus differences that are smaller than behavioral thresholds. Van Overwalle *et al.*^[16] constructed an objective evaluation framework for the spontaneous categorical processing, discrimination, and classification of visual morphology by analyzing features such as response lateralization and mean amplitude. Furthermore, deep learning-based EEG

decoding models have significantly advanced the application of BCI technology in objective threshold detection. Alsuradi *et al.*^[17] employed a generalized EEG deep learning decoding model to achieve multimodal tactile EEG decoding based on physical stimulus annotations, providing a robust basis for the Objective detection of tactile perception thresholds. van den Berg *et al.*^[18] modified the EEGNet architecture to enable the real-time detection of perception thresholds for nociceptive stimuli *via* a two-alternative forced-choice (2-AFC) task, circumventing the need for subjective behavioral responses.

However, current BCI-based auditory research predominantly employs the Oddball paradigm with long inter-stimulus intervals. The uniform application of fixed frequency differences in these paradigms ignores individual perceptual variations^[19] and is prone to introducing non-specific variations arising from inconsistent cognitive loads^[20], thereby compromising the stability of neural metrics. Furthermore, acoustic signals rarely occur in isolation within real-world environments, making traditional Oddball and AFC paradigms incongruent with everyday listening scenarios^[21]. Relevant cognitive psychological studies have demonstrated that, compared to discrete stimuli, the brain encounters heightened cognitive challenges when processing rapid, continuous, long-sequence sound streams, which is accompanied by an attenuation in discrimination sensitivity^[22-23]. Moreover, traditional clinical hearing thresholds solely measure sound audibility and fail to reflect auditory neural coding capabilities; consequently, individuals with clinically normal hearing may still exhibit limited auditory performance in real-world environments^[24]. Therefore, employing BCI technology to investigate the specific neural representations of subtle frequency variations under long-sequence continuous stimulation at the threshold state will facilitate the development of an objective FDL evaluation system that better reflects real-world scenarios.

To address the aforementioned issues, we innovatively designed a personalized rapid serial auditory presentation (RSAP) paradigm based on adaptive FDL, aiming to investigate the specific neural response characteristics of the brain when processing long-sequence continuous acoustic streams and subtle frequency deviations. Building upon a systematic analysis of ERP and frequency-domain features, another innovation of this study lies in the development of a Multi-Scale Spatial-Temporal Dual Attention Network (MS-STAMNet) tailored for decoding EEG signals elicited by this paradigm. This

network integrates multi-scale frequency band partitioning, an enhanced spatial-temporal attention mechanism, and a cross-scale feature interaction module, enabling it to dynamically capture and focus on discriminative high-dimensional neural features. Finally, through a single-trial target sound classification task, the efficacy of the proposed model in EEG decoding was robustly validated.

1 Methods

1.1 Participants

This study recruited 11 healthy participants (5 males, 6 females) aged between 22 and 25 years. None of the participants had a history of hearing impairment or neurological disorders. They were required to have adequate sleep on the day preceding the experiment and to abstain from consuming alcohol or psychoactive medications. Prior to the commencement of the experiment, all participants were fully informed of the study's purpose and provided written informed consent for the collection and use of their experimental data. The experimental protocols were formally approved by the Biomedical Ethics Committee of Tianjin University (Approval No: TJUE2026-H-S-033) and were conducted in strict adherence to the ethical principles outlined in the Declaration of Helsinki. All participants completed both the FDL testing and the EEG experiments.

1.2 Paradigm design

The auditory stimuli consisted of pre-generated pure-tone sequences. Each sequence, constituting a single trial, contained 12 pure tones. Each tone was presented for a duration of 100 ms, with an inter-stimulus interval (ISI) of 100 ms between adjacent tones. To prevent clicking artifacts, a 10-ms linear fade-in and fade-out ramp was applied to each tone. The acoustic stimuli were delivered binaurally *via* headphones (Sennheiser HD25) in a quiet environment, with the sound volume maintained at a constant, comfortable level throughout the experiment. In the non-target trials, the frequency of all tones was fixed at a baseline of 4 000 Hz, which represents a standard high-frequency tone routinely employed in clinical audiometry. In the target trials, a single target tone was embedded at a random position within the sequence, excluding the first and last positions. The frequency of this target tone was determined by the individual's FDL. The individual FDL were measured utilizing a dynamic weighted up-down adaptive procedure, targeting a specific expected response probability of 60%. Detailed methodological procedures are described in the

literature^[11-12]. Drawing on prior neuroimaging designs^[11], the 60% target probability was adopted to maintain the acoustic stimuli precisely at the perceptual boundary, thereby ensuring the successful capture of critical neural features elicited at the very margin of auditory discrimination. To preclude task difficulty discrepancies arising from paradigm variations, the threshold measurement was also formulated as a 2-AFC task based on the identical RSAP paradigm.

The experimental paradigm was implemented using the open-source platform MetaBCI^[25]. The specific experimental design is illustrated in Figure 1. During the formal experimental phase, each block consisted of 60 trials, comprising 40 target trials (accounting for 66.6% of the total trials) and 20 non-target trials. Following the presentation of the entire auditory sequence in each trial, a 2 500 ms response window was provided. During this period, participants were instructed to judge the presence of a target tone and respond as quickly as possible by pressing a

designated key on the keyboard. Trials with a reaction time exceeding 2 000 ms were recorded as incorrect responses. Each participant completed a total of 15 blocks, with each block lasting approximately 5 min. Throughout the experiment, participants were seated in a comfortable chair, maintaining a relaxed posture. They were instructed to wear the EEG cap and headphones, fixate on a crosshair displayed on the center of the screen, and minimize gross motor movements to reduce muscular artifacts.

Continuous EEG data were acquired using a Neuroscan SynAmps2 amplifier and a 64-channel Quick-Cap equipped with Ag/AgCl electrodes, which were positioned according to the international 10-20 system standards. During data acquisition, the vertex was utilized as the online reference, and the ground electrode was positioned between the prefrontal FPz and Fz leads. The EEG signals were digitized at a sampling rate of 1 000 Hz and subjected to an online band-pass filter between 0.1 and 200 Hz.

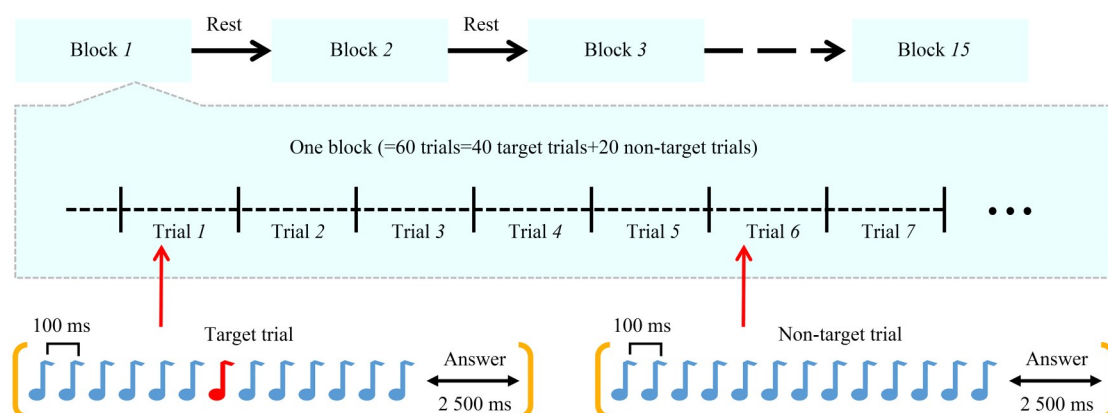


Fig. 1 RSAP paradigm

1.3 EEG data preprocessing

All EEG data were first re-referenced to the average of the M1 and M2 mastoid channels. Subsequently, a finite impulse response (FIR) bandpass filter of 1–45 Hz was applied, and the data were downsampled to 256 Hz. Based on the onset of each audio stimulus within the trials, the continuous EEG data were segmented into epochs ranging from –200 to 1 000 ms, with the pre-stimulus interval of –200 to 0 ms utilized for baseline correction to eliminate signal drift. Finally, independent component analysis (ICA) was performed to remove ocular and electromyographic artifacts.

1.4 EEG feature analysis

Focusing on the cognitive processes involved in

discriminating frequency deviations within long sequences at the FDL, the EEG data were systematically analyzed in both the time and frequency domains. This approach facilitates a comprehensive investigation into how the brain identifies and processes abrupt frequency changes amidst continuous auditory stimulation, thereby revealing the dynamic neural mechanisms and spectral characteristics of brain activity. Previous studies have demonstrated that frequency difference discrimination tasks elicit specific EEG alterations over the central brain regions^[26]. Consequently, our analysis primarily focused on the ERP at the Fz, Cz, CPz, Pz, T7 and T8 electrodes.

In addition to ERP temporal features, auditory

task also significantly modulates the spectral power of EEG signals across different cortical regions^[27]. In this study, the event-related spectral perturbation (ERSP) method was employed to analyze the time-frequency perturbation characteristics of the EEG signals. The ERSP calculations were performed across a frequency range of 1 to 20 Hz with a step size of 0.1 Hz, a range that encompasses the δ , θ , α , and β frequency bands of typical EEG rhythms.

1.5 Decoding algorithm model

EEG signals are characterized by high non-stationarity and complex spatial-temporal dynamics, with the embedded neural information frequently spanning multiple frequency bands. In the RSAP paradigm designed for this study, the brain's continuous processing of background tones and the transient evoked responses elicited by frequency deviants are superimposed, culminating in complex neural representations. Multi-scale network architectures, by constructing parallel and differentiated temporal receptive fields, can effectively capture these cross-band composite temporal representations, thereby demonstrating significant advantages in EEG signal decoding. However, traditional multi-scale networks typically

employ independent branches to process individual frequency bands, leading to high feature redundancy and a deficiency in the ability to dynamically focus on crucial spatial-temporal patterns^[28]. To address these limitations, this paper proposes an MS-STAMNet. By integrating the cross-scale feature interaction mechanism of Octave Convolution^[29] with a dynamic attention allocation strategy, this model adaptively suppresses redundant information during continuous auditory stimulation, accurately identifying and amplifying the specific neural responses evoked by deviant stimuli. Figure 2 illustrates the overall architecture of the proposed MS-STAMNet. Given a single-trial raw EEG input tensor with dimensions (C, E, T)—where C represents the network's input channel dimension for convolution (typically 1 for raw EEG), E denotes the number of physical electrodes, and T is the temporal length—the network sequentially processes the signal through 4 core modules: multi-scale decomposition, spatial attention, temporal attention, and cross-scale feature interaction. The detailed mechanisms of each module are elaborated in the subsequent subsections. The detailed mechanisms of each module are elaborated in the subsequent subsections.

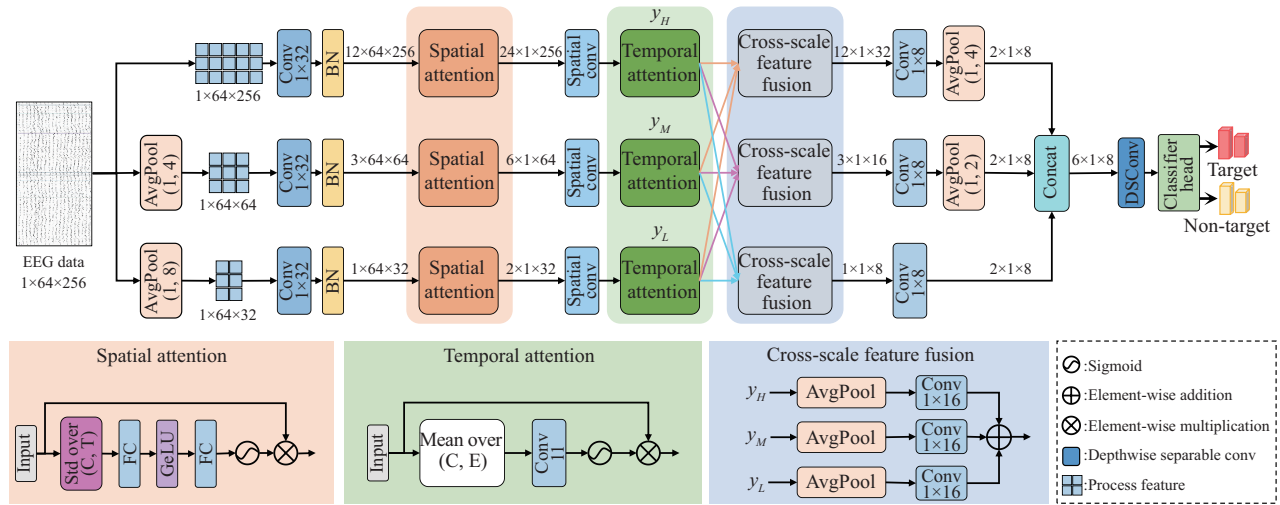


Fig. 2 Overall architecture of the MS-STAMNet model

1.5.1 Multi-scale decomposition module

Serving as the foundational feature extraction stage of the MS-STAMNet, this module aims to parallelly deconstruct neural feature patterns with hierarchical spatial-temporal dynamic characteristics from EEG signals. By introducing pooling operations with differentiated strides, the module maps the raw signals into three independent frequency branches: high, medium, and low. This design effectively

disentangles the feature overlap across different frequency bands, enabling each branch to focus exclusively on the specific representation of neural information within its dedicated frequency band. Consequently, this establishes a robust foundation for the subsequent attention mechanisms to precisely target key features, and for the cross-scale interaction module to achieve global spatial-temporal information fusion.

1.5.2 Spatial attention module

The cortical responses to various stimuli exhibit profound spatial heterogeneity; thus, the accurate identification of task-relevant electrode channels is of paramount importance. Traditional spatial attention mechanisms, which primarily rely on average pooling, frequently struggle to adequately capture complex feature relationships and are prone to smoothing out weak, transient neural activity features^[30]. To mitigate this limitation, this study proposes an energy fluctuation-based spatial attention module. This module utilizes standard deviation pooling to quantify the dynamic energy variations across individual electrode channels, thereby ameliorating the constraints associated with conventional mean-based metrics. Through autonomous learning within a non-linear mapping space, the module dynamically evaluates and recalibrates the spatial importance of the EEG topological nodes. This process adaptively amplifies the neural representations originating from critical brain regions involved in auditory processing and target recognition while concurrently suppressing interference from redundant channels, thus providing a purified representational space for precise subsequent feature extraction.

1.5.3 Temporal attention module

In frequency difference discrimination tasks under continuous rapid stimulation, event-related response features are concentrated within specific time windows. To accurately capture these critical temporal dynamic patterns, this study introduces a one-dimensional (1D) convolution-based temporal attention mechanism. To address the issue that global mapping can disrupt sequence continuity, this module—following spatial feature fusion—employs a 1D local convolution operator with a specific receptive field to deeply explore the contextual dependencies of adjacent time steps, thereby adaptively evaluating the importance of each temporal node. Subsequently, utilizing a temporal attention mask generated *via* non-linear activation, higher weights are dynamically and continuously assigned to the time segments containing core discriminative information. This mechanism not only preserves the continuity of neural waveforms in the temporal dimension but also effectively extracts the temporal evolutionary trajectories of the evoked responses.

1.5.4 Cross-scale feature interaction module

In the RSAP task, the brain's continuous processing of background tones is intrinsically coupled with the response features to target stimuli. Consequently, independent frequency band analysis struggles to effectively capture such complex

correlations. To overcome the limitation of feature isolation inherent in independent branches, the cross-scale feature interaction module establishes information exchange channels across different frequency levels. As illustrated in Figure 2, to ensure that signals with disparate temporal resolutions are aligned prior to fusion, this module implements differentiated average pooling down-sampling strides for interaction paths in specific directions. Through precise temporal alignment and convolutional mapping, the high-, medium-, and low-frequency features undergo deep cross-fusion. This process fosters a synergistic effect between the transient event responses extracted by the high-frequency branch and the long-range auditory stream rhythms represented by the low-frequency branch. Consequently, the network can leverage the global temporal context to effectively calibrate and amplify local critical neural features. Through multi-scale feature aggregation, the model successfully resolves the representational deficiency of a single receptive field, providing a hierarchically distinct and interconnected high-dimensional spatial-temporal composite representation for the final classification decision.

1.5.5 Model evaluation metrics

To comprehensively and objectively evaluate the model's efficacy, several metrics were selected for a thorough assessment. In terms of global performance, the area under the curve (AUC) and the weighted average recall (WAR) characterize the model's overall discriminative robustness and general predictive accuracy, respectively. To address the imbalanced nature of the data distribution, the unweighted average recall (UAR) was utilized. The UAR effectively mitigates majority-class bias and objectively quantifies the model's robustness in recognizing all classes in a balanced manner. Furthermore, the true positive rate (TPR) and false positive rate (FPR) delineate the hit sensitivity for target stimuli and the false-alarm risk associated with non-target backgrounds; combined with the F1-Score, which represents the harmonic mean of precision and recall, these metrics facilitate a comprehensive evaluation of the discriminative capacity for the target class^[31].

1.5.6 Model training settings and implementation

The model was optimized using the Adam optimizer with an initial learning rate of 0.001, and cross-entropy was employed as the loss function. All target and non-target trials, regardless of conscious perception or response accuracy, were used for network training. The dataset was partitioned into training, validation, and testing sets according to a

7 : 2 : 1 ratio. To address the issue of data imbalance, class weights were incorporated into the loss computation. During the training process, a dynamic learning rate adjustment strategy was applied, wherein the learning rate was halved if the validation loss failed to decrease for five consecutive epochs. Furthermore, an early stopping mechanism was implemented to mitigate overfitting and reduce computational time; specifically, training was terminated if the validation loss showed no improvement for 20 consecutive epochs. The batch size for mini-batch gradient descent was set to 64^[32].

All deep learning models in this study were implemented utilizing the PyTorch framework and trained on an NVIDIA GeForce RTX 5060 GPU equipped with CUDA 12.4. All models were trained using identical configurations to ensure fair valid comparisons.

2 Results

2.1 Behavioral results

Behavioral data, specifically response accuracy and reaction time (RT), are the most frequently employed evaluation metrics in psychological research. Table 1 details the individualized target frequencies configured based on each participant’s perceptual threshold, along with their corresponding average response accuracy and reaction time. As indicated in the table, pronounced variations exist among individual thresholds. Through threshold adaptive adjustment, the participants’ response accuracy was maintained at approximately 65%, yielding an overall average accuracy of 66.33% across

all subjects. The average accuracy was slightly higher than the target probability set by the adaptive procedure, which can likely be attributed to perceptual learning or practice effects during the experiment that gradually enhanced participants’ frequency discrimination abilities. Participant reaction times ranged from 270 to 770 ms, with a grand average of 475.67 ms. These behavioral findings demonstrate that the experimental design successfully achieved the criteria for auditory threshold stimulation, thereby establishing a solid foundation for the subsequent analysis of the threshold-level EEG data.

Table 1 Accuracy rate and reaction time of subjects

Subject	Target frequency/Hz	Accuracy/%	RT/ms
S_1	4 015	68.89	278.98
S_2	4 062	66.78	485.37
S_3	4 020	65.33	325.81
S_4	4 058	69.33	618.71
S_5	4 052	65.63	749.33
S_6	4 059	66.11	769.06
S_7	4 028	65.20	413.35
S_8	4 021	63.22	417.37
S_9	4 034	66.78	552.27
S_10	4 014	64.67	316.56
S_11	4 061	67.67	305.61

2.2 ERP results

Following the grand averaging of data across all participants, pronounced ERP features were observed over the central and temporal regions. Figure 3 illustrates the event-related potential waveforms from

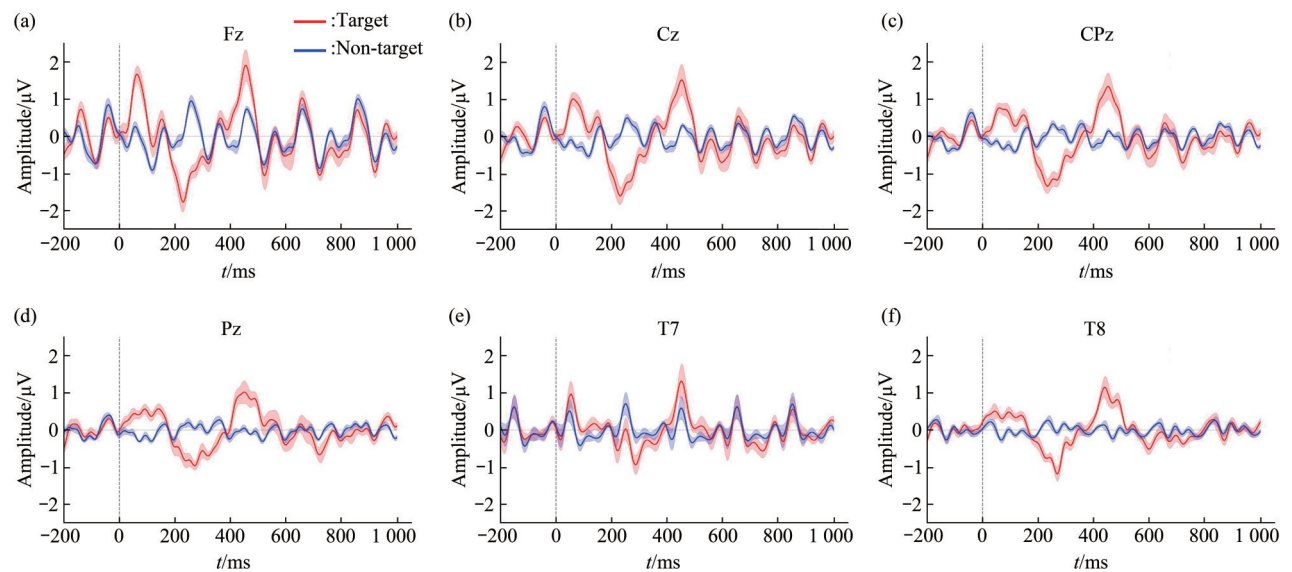


Fig. 3 Comparison of ERP waveforms between target and non-target stimuli
(a-f) show the ERP waveforms at electrodes Fz, Cz, CPz, Pz, T7 and T8 in sequence.

central midline electrodes Fz, Cz, CPz, and Pz, alongside bilateral temporal channels T7 and T8. Target and non-target conditions exhibited distinct temporal dynamics following stimulus onset. Over the central region, target stimuli reliably elicited an N2 component between 200 and 250 ms post-stimulus, followed by a P3 component spanning the 400 to 550 ms window. These task-relevant activations were similarly pronounced over the temporal lobes. Specifically at the T7 and T8 electrodes, target stimuli evoked a marked N2 negative deflection from 200 to 300 ms and a subsequent strong P3 positive peak from 400 to 500 ms. In contrast, non-target stimuli generated no comparable salient responses across any of these evaluated time windows.

Further analysis of the spatial distribution characteristics in Figure 4 reveals that at 200 ms, the negative potential activation elicited by target stimuli was primarily concentrated in the frontal, central, and temporal regions; subsequently, around 450 ms, the target stimuli evoked robust positive potential activation exhibiting a broad scalp distribution, which was most prominent over the frontal and central areas. In contrast, the topographical voltage distribution for non-target stimuli remained relatively flat at these corresponding time points, with no evident regions of strong activation. Therefore, the proposed experimental paradigm successfully elicited the N2 and P3 components, which are highly associated with target recognition.

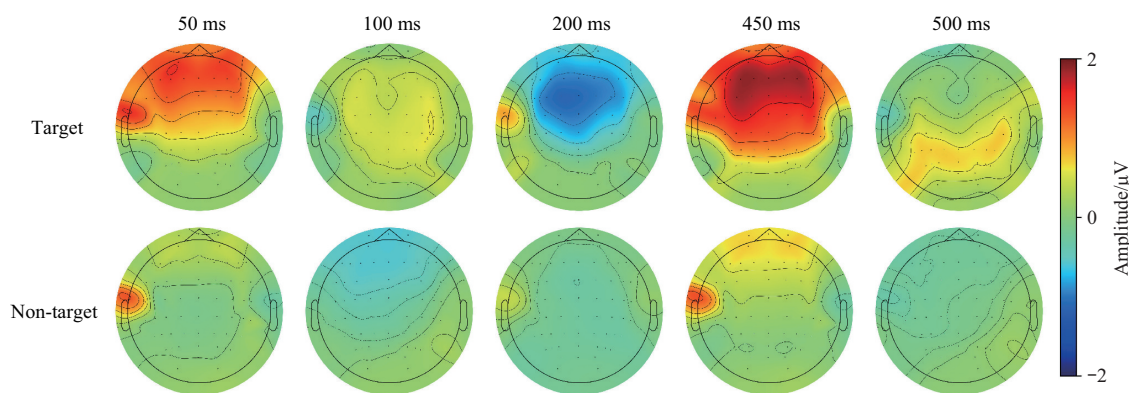


Fig. 4 Time-domain topographic maps for target and non-target stimuli

The maps depict the spatial distribution of ERP amplitudes across five key time points (50, 100, 200, 450, and 500 ms) under both experimental conditions. The color bar denotes the voltage amplitude in microvolts (μV), with warm colors representing positive potentials and cool colors indicating negative potentials.

2.3 ERSF results

To further investigate the neural oscillatory characteristics and their differences between target and non-target stimuli in the time-frequency domain, ERSF features were extracted, and paired-sample *t*-tests with false discovery rate (FDR) correction were performed on the data across the two stimulus conditions.

Figure 5 illustrates the post-stimulus ERSF time-frequency characteristics. The frequency-deviant stimuli induced distinct spectral alterations, characterized by δ and θ event-related synchronization (ERS) that initiated around 200 ms post-stimulus and persisted over the fronto-central region. This was followed by an α band event-related desynchronization (ERD). Notably, the duration of this α attenuation was channel-dependent, peaking in magnitude at the Pz electrode, which may index shifts

in cognitive states.

The topography of the *P*-value distribution derived from the paired-sample *t*-tests further revealed statistically significant differences between the target and non-target stimuli within specific time-frequency windows. Within the 150–600 ms interval following stimulus onset, significant differences were observed across all four electrode sites in the 2–7 Hz frequency range. Integrating these findings with the previously discussed temporal ERP features, the low-frequency power disparities within this time-frequency window correspond well with the neural oscillatory mechanisms underlying cognitive potential components, such as the N2 and P3, which are elicited during target recognition. Furthermore, during the late processing stage (600–1 000 ms post-stimulus), significant difference clusters emerged in the α to low- β bands over the centro-parietal region.

Taken together, these analyses indicate that within this auditory paradigm, the participants' cognitive processing of the target stimuli is characterized not only by the early synchronization of oscillatory power in the δ and θ bands but also by the

late-stage modulation of neural activity in the α band. Moreover, these time-frequency response features exhibited a consistent spatial distribution across the fronto-parietal regions.

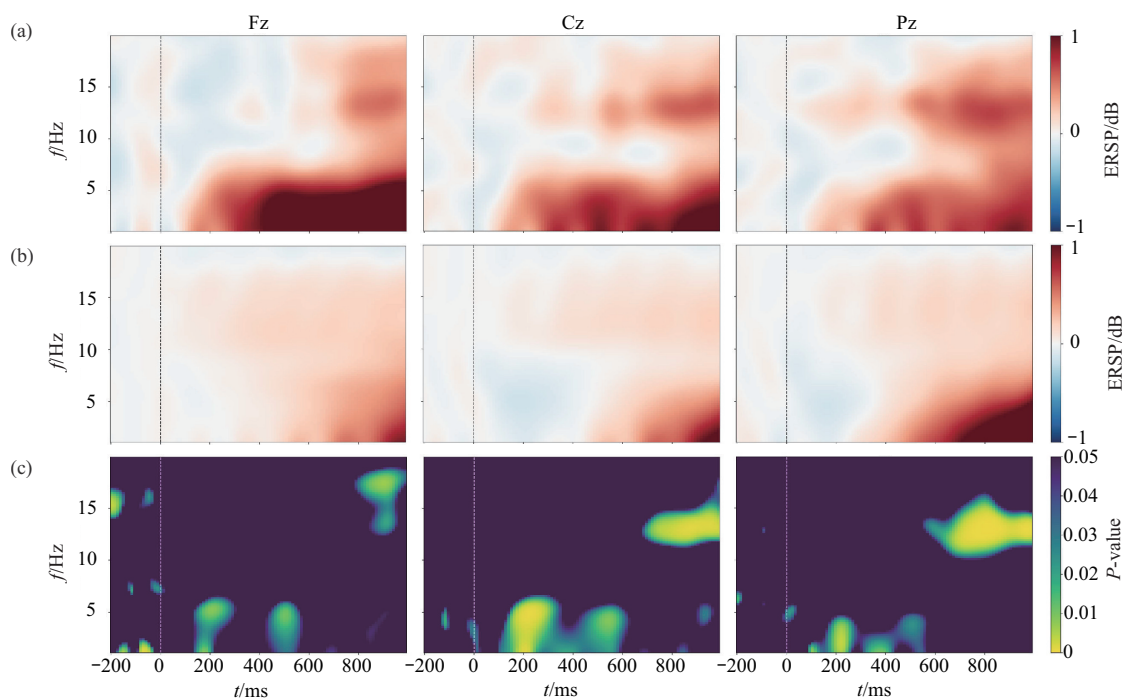


Fig. 5 Comparison of ERSP between target and non-target stimuli

Time-frequency ERSP representations at representative electrodes. (a, b) ERSP dynamics induced by target and non-target stimuli, respectively (red: power enhancement; blue: power attenuation). (c) Statistical significance maps from paired t -tests between the two conditions (yellow colors indicate smaller P -values). Columns from left to right correspond to electrodes Fz, Cz, and Pz.

The topographical maps of ERSP in Figure 6 depict the spatiotemporal evolution of event-related spectral power changes associated with target and non-target stimuli. Under the target stimulus condition, the early-to-middle latency window (254–395 ms) exhibited a significant enhancement in θ -band ERSP over the frontal lobe. Concurrently, ERSP attenuation in the α band was observed across the frontal, temporal, and parietal cortices. During the late processing stage (after 605 ms), the δ and θ bands displayed sustained, high-intensity spectral perturbations with a widespread spatial distribution over the frontal and central cortices. Simultaneously, an increase in α -band ERSP emerged over the parietal lobe.

In contrast, under the non-target stimulus condition, the ERSP across all frequency bands globally remained near the baseline level throughout the entire observation time sequence, with no

discernible spatiotemporal fluctuations. In summary, the two stimulus conditions generated highly contrasting spatiotemporal ERSP differences within the δ , θ , and α bands, thereby providing robust, objective neural representations for decoding target intentions within this experimental paradigm.

2.4 Decoding and detection performance

To comprehensively and objectively evaluate the performance and effectiveness of the MS-STAMNet model, four baseline models—EEGNet^[33], EEG-Inception^[34], PLNet^[32], and MOCNN^[29] were selected for comparative analysis. Paired-sample t -tests were employed for statistical evaluation, as illustrated in Figure 7a.

The results demonstrate that the MS-STAMNet achieved an average AUC of 0.761 ± 0.07 , a UAR of $(69.67 \pm 6.12)\%$, and a TPR of $(69.95 \pm 7.50)\%$, comprehensively and significantly outperforming all baseline models. These findings indicate that the

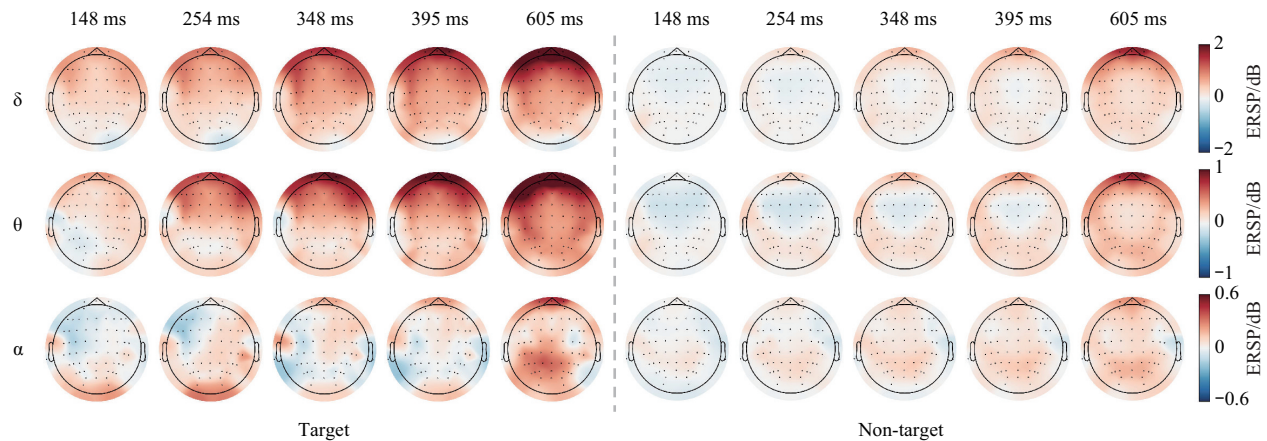


Fig. 6 Spatial distribution of ERSP for target and non-target stimuli

Topographic maps illustrate the ERSP dynamics within the δ , θ , and α frequency bands (rows from top to bottom). For each band, the left and right panels present the neural oscillatory responses elicited by target and non-target stimuli, respectively. A unified color scale is applied across each row to ensure accurate cross-condition comparison, where warm colors indicate spectral synchronization and cool colors represent desynchronization.

proposed network not only possesses a robust capacity for constructing global decision boundaries but also exhibits exceptional robustness when handling highly imbalanced EEG data. It effectively circumvents the trap of majority-class bias, thereby enabling the accurate recognition of frequency-

deviant stimuli at the threshold level. Furthermore, its average *WAR* (69.42 ± 6.42) % and *F1-Score* (0.32 ± 0.07) significantly surpassed those of EEGNet and PLNet, demonstrating superior overall predictive performance.

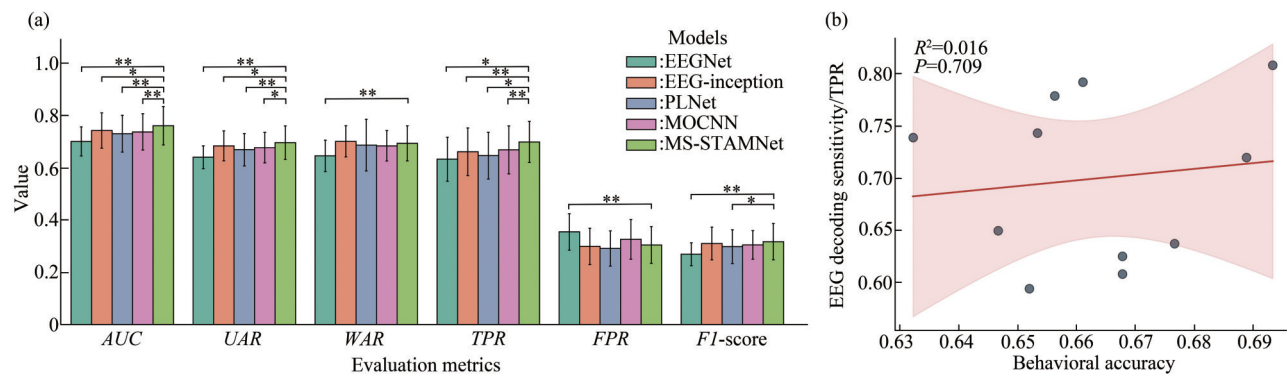


Fig. 7 Model performance evaluation

(a) Comparison between MS-STAMNet and baseline models. Asterisks indicate statistical significance between the models connected by lines (* $P < 0.05$, ** $P < 0.01$). (b) Regression between participants' behavioral accuracy and the model decoding TPR.

To integrate behavioral and neural metrics, we assessed the correlation between individual behavioral accuracy and subject-level decoding performance. TPR was selected as the representative metric, as it directly quantifies the detection of rare, frequency-deviant targets, tightly aligning with the participants' core cognitive task. Regression analysis in Figure 7b revealed a distinct dissociation between behavioral

accuracy and decoding TPR ($R^2=0.016$, $P=0.709$). Furthermore, the average decoding TPR across all subjects notably exceeded their mean behavioral accuracy. This behavioral-neural decoupling suggests that MS-STAMNet effectively captures implicit, pre-attentive neural representations of near-threshold acoustic variations, even when these variations fail to elicit explicit conscious responses. Taken together, the

MS-STAMNet demonstrates remarkable efficacy in the deep extraction and fusion of multi-scale spatial-temporal features when processing complex, high-noise EEG signals.

The distribution maps, generated by averaging the temporal and spatial attention weights of the decoding model across all participants, reveal the model's mechanism for the adaptive extraction and reweighting of specific features within multi-band EEG signals. As illustrated in Figure 8a, the weight topographies across the different branches in the spatial dimension exhibit a high degree of consistency alongside local specificity. The high-weight regions are predominantly concentrated over the bilateral temporal areas and portions of the fronto-parietal lobes, which are integral to auditory processing and attentional networks.

In the temporal dimension, the temporal attention heatmaps in Figure 8b elucidate the temporal dynamics of the weights across various frequency bands: the low-frequency branch displays elevated weights within the 0–300 ms window, primarily

capturing early neural response features; the medium-frequency branch exhibits a relatively concentrated high-weight activation band between 250 and 350 ms, aligning with the identification of the typical P3 event-related potential to accomplish the critical discrimination of auditory targets. Furthermore, the high-frequency branch, which is responsible for capturing local transient oscillations, demonstrates a more complex multiphasic temporal distribution. It not only co-activates with the low and medium frequency branches during the 250–320 ms interval but also presents secondary high-weight windows during the late processing stages (550–600 ms and 750–900 ms), capturing transient features associated with higher-order cognitive processing.

In summary, by incorporating the spatial-temporal attention mechanism, the proposed model effectively integrates the spatial-temporal discriminative information elicited by target stimuli and suppresses redundant features, thereby achieving the highly efficient decoding of EEG responses.

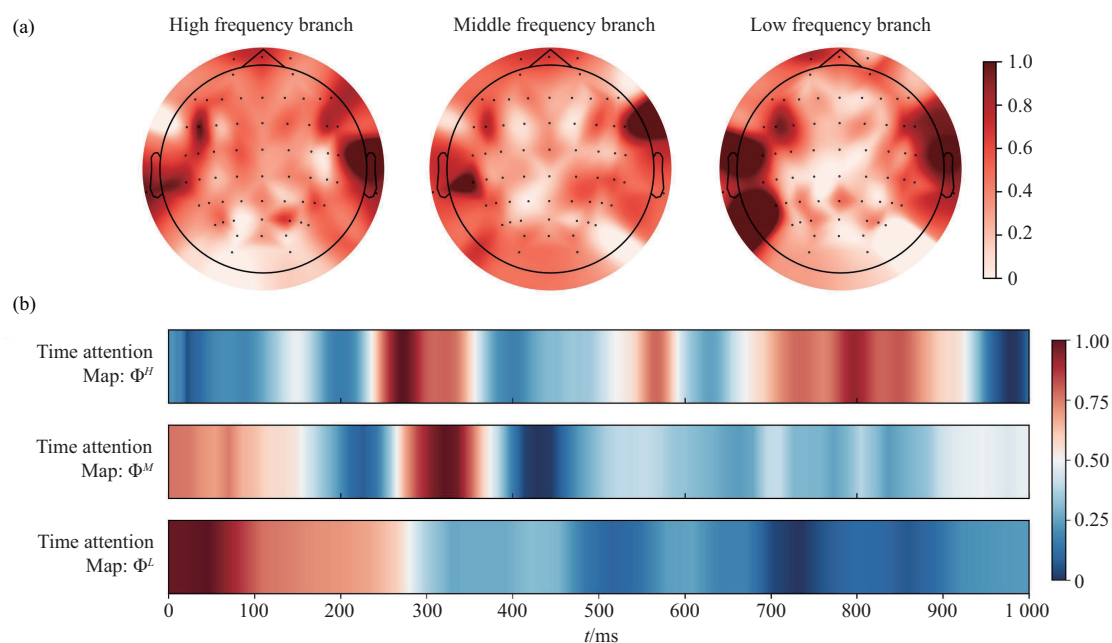


Fig. 8 Spatial and temporal attention weight distribution map

(a) Spatial attention weights across the high, middle, and low frequency branches. (b) Temporal attention weights across the corresponding frequency branches. The color scale represents the weight magnitude (warm colors: high neural importance; cool colors: low importance).

3 Discussion

Traditional psychophysical methods are highly susceptible to subjective behavioral factors, and existing EEG paradigms frequently overlook individual differences. Furthermore, the processing of

continuous acoustic streams imposes increased cognitive resource demands on the brain, accompanied by an attenuation in discrimination efficacy. To address these limitations, this study designed a personalized RSAP paradigm aimed at capturing the specific neural responses to subtle

frequency deviations at the threshold level within long sequences, thereby achieving an objective resolution of individual auditory perceptual boundaries.

At the level of neural representation, through an in-depth analysis of ERP and ERSP, this study systematically reveals the multidimensional response patterns of the brain when processing frequency differences at the threshold level. The results demonstrate that target stimuli elicited pronounced N2 and P3 components over the frontal, central, and temporal regions, which aligns with previous research findings^[35]. Time-frequency analysis further revealed synchronization in the δ and θ bands, alongside desynchronization in the α band. Previous studies have indicated that the enhancement of δ -band activity primarily reflects top-down modulatory signals originating from the frontal lobe, which exert causal regulation on the neural activity of the auditory cortex^[36]. Conversely, the augmentation of θ -band activity in the frontal region corresponds to the working memory demands associated with effortful listening. This reflects the cognitive load imposed on the brain to maintain auditory information and execute internal attentional control during threshold perception tasks^[27]. The desynchronization characteristic of the α band represents the release of cortical inhibition, serving as a marker for the synergistic activation of task-related cortical networks and reflecting top-down attentional control^[37]. Furthermore, the clarity of subjective sound perception has been correlated with the attenuation of α -band activity in the auditory cortex^[38]. Collectively, these findings provide a solid neurobiological foundation for elucidating the brain's mechanisms of feature extraction and intention discrimination regarding subtle auditory targets. They robustly demonstrate that an individual's perception of frequency differences at the threshold level necessitates the modulation of numerous higher-order cognitive systems.

At the level of decoding algorithms, the MS-STAMNet model proposed in this study effectively addresses the issues of feature redundancy and limited receptive fields commonly encountered when processing low signal-to-noise ratio and non-stationary physiological signals by integrating multi-scale decomposition, spatial-temporal attention, and cross-scale interaction modules. In the single-trial decoding of weak auditory targets, its core metrics, such as AUC, significantly outperform those of baseline models. Through dynamic spatial-temporal adaptive weighting, the model precisely localizes the bilateral temporal regions and the frontoparietal lobes

associated with auditory sensation in the spatial dimension. In the temporal dimension, it effectively resists single-trial latency fluctuations by adaptively focusing on intervals with stable time-locking characteristics to extract the evolutionary features of multi-band neural responses. While Meng *et al.*^[39] demonstrated that the decoding performance of EEGNet drops sharply as the disparity between acoustic stimuli diminishes, MS-STAMNet leverages its cross-scale interaction and dynamic focusing mechanisms to effectively avoid feature loss and performance degradation caused by subtle stimuli. By maintaining superior robustness under low SNR conditions, the model achieves an efficient and in-depth characterization of weak threshold-level auditory-evoked features. This provides a robust methodological foundation for the objective and precise quantification of an individual's true auditory perceptual boundaries.

From the perspective of application prospects, the findings of this study offer broad-reaching value. In the field of auditory assessment, EEG-based neural response metrics can serve as biomarkers for evaluating hearing loss, providing objective diagnostic evidence that overcomes the limitations of traditional behavioral tests. Furthermore, considering the critical role of high-frequency acoustic information in speech comprehension and spatial sound localization, this study utilized 4 000 Hz—a frequency particularly susceptible to noise-induced hearing impairment—as the base stimulus frequency^[40], highlighting the translational potential for clinical applications in auditory assistance and rehabilitative interventions. The in-depth analysis of neural response mechanisms at the near-threshold perceptual boundary provided by this study is poised to facilitate the development of closed-loop neurofeedback training systems. By target-enhancing individual electrophysiological representations of subtle stimuli at specific frequency bands, such systems could improve the auditory cognitive capabilities of specialized populations. Moreover, as frequency information serves as a dominant cue for compensatory perception in adverse acoustic environments^[4,41], this research provides a robust reference for brain-computer hybrid intelligence in industrial safety monitoring, sound target recognition in noise-rich environments, and human-machine interaction in smart homes. Ultimately, these advancements may assist operators in achieving more precise perception and decision-making within real-world soundscapes.

Despite the current advancements, several

limitations warrant future investigation. First, the reliance on pure-tone stimuli restricts the model's generalizability to complex, natural acoustic environments. Second, decoding performance for subtle frequency deviations remains suboptimal; future iterations could leverage Transformer architectures to better model long-range EEG temporal dependencies and stabilize decision boundaries. Furthermore, larger and more diverse participant cohorts are required to enhance statistical power and population-level generalizability. Finally, subsequent research should validate the framework's robustness against real-world interference by simulating "cocktail party" scenarios.

4 Conclusion

This study investigated specific neural responses and objective evaluation methods of the brain when processing long-sequence continuous stimuli and frequency deviations at the threshold state by designing a personalized RSAP paradigm based on the FDL and proposing a decoding network, MS-STAMNet, which integrates multi-scale feature interaction with a dynamic spatial-temporal attention mechanism. The research results indicated that target stimuli elicited significant N2 and P3 components in the fronto-central region. In the time-frequency domain, these responses were characterized by event-related synchronization in the δ and θ bands and event-related desynchronization in the α band. Regarding decoding performance, MS-STAMNet achieved an average *UAR* of 69.67% and an *AUC* of 0.761 8, significantly outperforming baseline models and providing robust support for the objective and quantitative evaluation of auditory capabilities. Future research could involve incorporating stimuli from real-world environments, optimizing temporal modeling architectures, and expanding the participant cohort to further refine the evaluation system.

Data Availability The data associated with this paper (DOI: 10.57760/sciencedb.42036, CSTR: 31253.11. sciencedb.42036) are available from the Science Data Bank database (<https://www.scidb.cn/s/ARVbIj>).

References

- [1] Micheyl C, Xiao L, Oxenham A J. Characterizing the dependence of pure-tone frequency difference limens on frequency, duration, and level. *Hear Res*, 2012, **292**(1/2): 1-13
- [2] Deroche M L D, Culling J F, Chatterjee M, *et al.* Roles of the target and masker fundamental frequencies in voice segregation. *J Acoust Soc Am*, 2014, **136**(3): 1225-1236
- [3] Oh Y, Zuwala J C, Salvagno C M, *et al.* The impact of pitch and timbre cues on auditory grouping and stream segregation. *Front Neurosci*, 2021, **15**: 725093
- [4] Kishon-Rabin L, Zaltz Y. The effect of noise on the utilization of fundamental frequency and formants for voice discrimination in children and adults. *Appl Sci*, 2023, **13**(19): 10752
- [5] McClaskey C M. Standard-interval size affects interval-discrimination thresholds for pure-tone melodic pitch intervals. *Hear Res*, 2017, **355**: 64-69
- [6] Zarate J M, Ritson C R, Poeppel D. Pitch-interval discrimination and musical expertise: is the semitone a perceptual boundary?. *J Acoust Soc Am*, 2012, **132**(2): 984-993
- [7] Zhang X, Gong Q. Correlation between the frequency difference limen and an index based on principal component analysis of the frequency-following response of normal hearing listeners. *Hear Res*, 2017, **344**: 255-264
- [8] Vonck B M D, Lammers M J W, Schaake W A A, *et al.* Cortical potentials evoked by tone frequency changes compared to frequency discrimination and speech perception: thresholds in normal-hearing and hearing-impaired subjects. *Hear Res*, 2021, **401**: 108154
- [9] Fung J, Whiteford K L, Mehta A H, *et al.* Pitch perception in school-aged children: pure tones, resolved and unresolved harmonics. *JASA Express Lett*, 2025, **5**(1): 014403
- [10] Grassi M, Felling A, Orlandi N, *et al.* PSYCHOACOUSTICS-WEB: a free online tool for the estimation of auditory thresholds. *Behav Res Meth*, 2024, **56**(7): 7465-7481
- [11] Bianchi F, Hjortkjær J, Santurette S, *et al.* Subcortical and cortical correlates of pitch discrimination: evidence for two levels of neuroplasticity in musicians. *NeuroImage*, 2017, **163**: 398-412
- [12] Kaernbach C. Simple adaptive testing with the weighted up-down method. *Percept Psychophys*, 1991, **49**(3): 227-229
- [13] Gu F, Wong L, Hu A, *et al.* A lateral inhibition mechanism explains the dissociation between mismatch negativity and behavioral pitch discrimination. *Brain Res*, 2019, **1720**: 146308
- [14] Avsar Aksu B, Sahin Ceylan D, Gültekin G. Mismatch negativity as a tool to detect the intensity just noticeable difference. *Arch Acoust*, 2024, **49**(4): 483-489
- [15] Xu Q, Gong Q. Frequency difference beyond behavioral limen reflected by frequency following response of human auditory Brainstem. *Biomed Eng Online*, 2014, **13**: 114
- [16] Van Overwalle J, Van der Donck S, Van de Cruys S, *et al.* Assessing spontaneous categorical processing of visual shapes via frequency-tagging EEG. *J Neurosci*, 2024, **44**(16): e1346232024
- [17] Alsuradi H, Atinafu Y, Eid M. Evaluating haptic experience using EEG and deep learning across multiple modalities: linking stimulus and self-reports. *Front Neurosci*, 2026, **20**: 1666558
- [18] van den Berg B, Vanwinsen L, Jansen N, *et al.* Real-time estimation of perceptual thresholds based on the electroencephalogram using a deep neural network. *J Neurosci*

- Methods, 2022, **374**: 109580
- [19] Chang M, Iizuka H, Naruse Y, *et al.* Unconscious learning of auditory discrimination using mismatch negativity (MMN) neurofeedback. *Sci Rep*, 2014, **4**: 6729
- [20] Pichora-Fuller M K, Kramer S E, Eckert M A, *et al.* Hearing impairment and cognitive energy: the framework for understanding effortful listening (FUEL). *Ear Hear*, 2016, **37**(Suppl 1): 5S-27S
- [21] Lorenzi C, Apoux F, Grinfeder E, *et al.* Human auditory ecology: extending hearing research to the perception of natural soundscapes by humans in rapidly changing environments. *Trends Hear*, 2023, **27**: 23312165231212032
- [22] Akça M, Vuoskoski J K, Laeng B, *et al.* Recognition of brief sounds in rapid serial auditory presentation. *PLoS One*, 2023, **18**(4): e0284396
- [23] Cousineau M, Carcagno S, Demany L, *et al.* What is a melody? On the relationship between pitch and brightness of timbre. *Front Syst Neurosci*, 2013, **7**: 127
- [24] Ruggles D, Bharadwaj H, Shinn-Cunningham B G. Normal hearing is not enough to guarantee robust encoding of suprathreshold features important in everyday communication. *Proc Natl Acad Sci USA*, 2011, **108**(37): 15516-15521
- [25] Mei J, Luo R, Xu L, *et al.* MetaBCI: an open-source platform for brain – computer interfaces. *Comput Biol Med*, 2024, **168**: 107806
- [26] Coy N, Bader M, Schröger E, *et al.* Change detection of auditory tonal patterns defined by absolute versus relative pitch information. A combined behavioural and EEG study. *PLoS One*, 2021, **16**(2): e0247495
- [27] Wisniewski M G, Iyer N, Thompson E R, *et al.* Sustained frontal midline theta enhancements during effortful listening track working memory demands. *Hear Res*, 2018, **358**: 37-41
- [28] Pan J, Dong Z, Zhao P, *et al.* Low-redundancy motor imagery EEG decoding based on dynamic attention and feature reconstruction. *Neurocomputing*, 2026, **661**: 132004
- [29] Jin J, Xu R, Daly I, *et al.* MOCNN: a multiscale deep convolutional neural network for ERP-based brain-computer interfaces. *IEEE Trans Cybern*, 2024, **54**(9): 5565-5576
- [30] Zhang R, Zou R, Zhao Y, *et al.* BA-Net: bridge attention in deep neural networks. *Expert Syst Appl*, 2025, **292**: 128525
- [31] Li Y, Liu W, Feng T, *et al.* Spatio-temporal progressive attention model for EEG classification in rapid serial visual presentation task. *IEEE Trans Biomed Eng*, 2026, **73**(1): 191-207
- [32] Zang B, Lin Y, Liu Z, *et al.* A deep learning method for single-trial EEG classification in RSVP task based on spatiotemporal features of ERPs. *J Neural Eng*, 2021, **18**(4): 0460c8
- [33] Lawhern V J, Solon A J, Waytowich N R, *et al.* EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces. *J Neural Eng*, 2018, **15**(5): 056013
- [34] Santamaria-Vázquez E, Martínez-Cagigal V, Vaquerizo-Villar F, *et al.* EEG-inception: a novel deep convolutional neural network for assistive ERP-based brain-computer interfaces. *IEEE Trans Neural Syst Rehabil Eng*, 2020, **28**(12): 2773-2782
- [35] Li T, Zhu X, Wu X, *et al.* Continuous theta burst stimulation over left and right supramarginal gyri demonstrates their involvement in auditory feedback control of vocal production. *Cereb Cortex*, 2022, **33**(1): 11-22
- [36] Park H, Ince R A A, Schyns P G, *et al.* Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. *Curr Biol*, 2015, **25**(12): 1649-1653
- [37] Kayser J, Tenke C E, Kroppmann C J, *et al.* Auditory event-related potentials and α oscillations in the psychosis prodrome: neuronal generator patterns during a novelty oddball task. *Int J Psychophysiol*, 2014, **91**(2): 104-120
- [38] Weisz N, Hartmann T, Müller N, *et al.* Alpha rhythms in audition: cognitive and clinical perspectives. *Front Psychol*, 2011, **2**: 73
- [39] Meng Q, Tian L, Liu G, *et al.* EEG-based cross-subject passive music pitch perception using deep learning models. *Cogn Neurodyn*, 2025, **19**(1): 6
- [40] Moore B C J. A review of the perceptual effects of hearing loss for frequencies above 3 kHz. *Int J Audiol*, 2016, **55**(12): 707-714
- [41] Winn M B, Chatterjee M, Idsardi W J. Roles of voice onset time and F0 in stop consonant voicing perception: effects of masking noise and low-pass filtering. *J Speech Lang Hear Res*, 2013, **56**(4): 1097-1107

面向听觉频率差异阈值客观评价的 个性化脑机检测范式与解码方法研究^{△,*}

李圣烨 肖晓琳** 于识航 张贝贝 安兴伟 许敏鹏 明 东

(天津大学医学院, 天津 300072)

摘要 **目的** 频率差异阈值 (FDL) 是量化听觉系统感知能力的核心指标, 然而传统测量方法易受主观因素的影响, 且已有研究范式忽略个体感知差异。本研究旨在基于脑机接口 (BCI) 技术, 通过解析阈值状态下的特异性神经响应, 建立听觉频率辨别能力的客观量化评价方法。**方法** 设计基于个体 FDL 的个性化快速序列听觉刺激 (RSAP) 范式, 通过连续刺激模拟真实感知环境, 探究阈值状态下微弱频率突变诱发的神经表征。鉴于听觉刺激诱发的响应特征在多频域下呈现复杂且差异化的时空分布规律, 本文进一步将跨尺度特征交互模块与动态时空注意力分配策略深度融合, 创新性地提出多尺度时空双注意力网络 (MS-STAMNet)。具体而言, 该网络构建多感受野并行处理分支, 并引入动态自适应权重策略精准定位核心神经活动信号, 进而通过跨分支特征信息交互深层融合多尺度信息, 实现对微弱听觉诱发响应的鲁棒单次解码。**结果** 阈值下微弱听觉频率偏差刺激在前额、中央和颞区诱发了 N2 与 P3 事件相关电位特征。时频特征表现为 δ 与 θ 频段的事件相关同步化及 α 频段的去同步化。模型性能对比分析表明, MS-STAMNet 的平均未加权平均召回率 (UAR) 为 (69.67±6.12) %, 曲线下面积 (AUC) 为 0.761 8±0.07, 其性能显著优于 EEGNet、PLNet 等基线模型。神经解码与行为表现的分离现象表明, 本模型能有效捕获微弱频率偏差的内隐特征, 权重可视化分析则进一步揭示了模型对关键特征的准确聚焦。**结论** 本研究系统揭示了长序列听觉阈值刺激下的神经响应演变规律, 验证了 MS-STAMNet 在单次微弱信号解析中的有效性, 为听觉能力的客观量化评价奠定了基础。

关键词 频率差异判别, 脑机接口, 快速序列听觉刺激, 个性化, 深度学习

中图分类号 R741.044, TN911.7

DOI: 10.3724/j.pibb.2026.0252

CSTR: 32369.14.pibb.20260252

[△] 第二十八届中国科协年会学术论文。

* 国家自然科学基金 (62106170) 资助项目。

** 通讯联系人。

Tel: 13752345268, E-mail: xiaoxiao0@tju.edu.cn

收稿日期: 2026-05-10, 接受日期: 2026-07-06