Sound Target Detection Under Noisy Environment Using Brain-Computer Interface

Ruidong Wang, Ying Liu, Jianting Shi, Bolin Peng, Weijie Fei, and Luzheng Bi

***Abstract*— As an important means of environmental reconnaissance and regional security protection, sound target detection (STD) has been widely studied in the field of machine learning for a long time. Considering the shortcomings of the robustness and generalization performance of existing methods based on machine learning, we proposed a target detection method by an auditory brain-computer interface (BCI). We designed the experimental paradigm according to the actual application scenarios of STD, recorded the changes in Electroencephalogram (EEG) signals during the process of detecting target sound, and further extracted the features used to decode EEG signals through the analysis of neural representations, including Event-Related Potential (ERP) and Event-Related Spectral Perturbation (ERSP). Experimental results showed that the proposed method achieved good detection performance under noisy environment. As the first study of BCI applied to STD, this study shows the feasibility of this scheme in BCI and can serve as the foundation for future related applications.**

***Index Terms*— Sound target detection, BCI, auditory ERP, ERSP, SVM.**

I. Introduction

SOUND target detection (STD) refers to detecting the target in the sound stream. Categorically, it belongs to the problem of sound event detection (SED). STD can serve as a part of a public safety surveillance or military reconnaissance system to detect potentially dangerous targets (such as Unmanned Aerial Vehicles (UAVs)), demonstrating vital practical values of STD.

Many researchers have proposed various methods for STD by using signal processing and machine learning. Taking the UAV detection (i.e., judging the presence of UAVs from the sound stream) as an example, Yang et al. [1] used short-time Fourier Transform (STFT) to extract the sound signatures of UAVs during flight and proposed a real-time detection system based on support vector machine. Solis et al. [2] discussed the performance of support vector machine and convolutional neural network classifiers for UAV sound detection. They used Mayer cepstrum coefficients of UAV sound as the classification features, and the results showed that the detection accuracy of the convolutional neural network classifier for UAVs was only 60-69%. In contrast, the classifier based on the support vector machine completed the classification task with an accuracy of 92%.

Although there have been many studies on the STD, the STD in noisy environments is still challenging. In a real scenario, the signal-to-noise ratio (SNR) of the target sound relative to the noise is likely to be at low levels and may change (such as UAV sound frequency changes caused by a sudden increase or decrease of rotor speed), leading to a substantial performance decline of STD based on machine learning. In 2005, Clavel et al. [3] pointed out that the SNR reduction causes a sharp decline in detection performance. On the premise of the same training strategy, the detection accuracy significantly decreases when the SNR decreases from 20 dB to 5 dB. Papadimitriou et al. [4] also found this problem. For the same SED model, 30 dB SNR data were used for training, and then 30 dB SNR and −5 dB SNR data were used for testing. The performance of the model in the low SNR test set was much lower than in the high SNR test set (precision decreased by 36.51%, and recall decreased by 57.58%). Ren et al. [5] demonstrated the influence of SNR on the detection performance in the study of the dangerous SED. In the same noise scene, the detection error rate of the same sound target on the test set reached 37.38% when the SNR was −15 dB and increased by 30.2% compared with the condition of −5 dB. This phenomenon also exists in [6]. Turpault et al. [7] pointed out that the performance of the algorithm decreases greatly when the algorithm tries to recognize the new sound clips that do not appear in the training set for the top acoustic event recognition algorithms in the 2019 DCASE (Challenge on Detection and Classification of Acoustic Scenes and Events) Challenge. They draw this conclusion by using the top 10 algorithms in the acoustic event detection problem in 2021 DCASE Task 4 [8]. By establishing four evaluation datasets, under the same training baseline, the effects of sound event occurrence density, occurrence time, duration, non-target aliasing, and reverberation on the performance of the algorithms were tested. It can be seen from the results that when any one of the above factors changes, there is a great influence on the performance.

One of the main reasons for the above phenomenon is that the algorithm models cannot achieve robustness to

Manuscript received 5 July 2022; revised 1 October 2022; accepted 1 November 2022. Date of publication 4 November 2022; date of current version 31 January 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 51975052. *(Corresponding author: Luzheng Bi.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the Beijing Institute of Technology.

The authors are with the School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China (e-mail: 1981723206@qq.com; jat_shi@163.com; b70ivor@gmai.com; 512236155@qq.com; bhxblz@bit.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2022.3219595

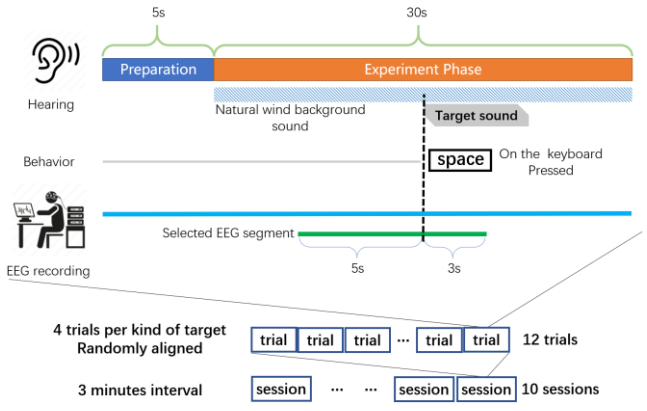

Fig. 1. Details of paradigm design. In each single trial, The subjects were given up to 5s to prepare, followed by a 30s period of natural wind noise with the target sound appearing at random timings. Once the subject perceived the presence of the target, the space bar on the keyboard was pressed as soon as possible. Each session consisted of 12 consecutive trials, and the whole experiment lasted at least 10 sessions.

environmental noise and sound signature change of targets. Brain-computer interfaces (BCIs) have been demonstrated to be capable of decoding human brain activity to visual target recognition to improve the performance of image target detection [9], [10]. In the same spirit as [9] and [10], in this article, we propose an EEG-based decoding method to translate the neural signature of human auditory target recognition to perform the STD.

The contribution of this paper is that it is the first work to develop a BCI to decode EEG signals associated with human auditory target recognition to detect sound targets. This work makes it possible for people to participate in the detection system efficiently when the confidence of the auto-detection algorithm drops and opens a new avenue to the research and development of STD techniques and advance the study of BCIs. The rest of this paper is organized as follows: Section II gives a detailed description of the stimulus and experiment paradigm. Section III describes the EEG data preprocessing pipeline and shows our method details for analyzing auditory ERP and ERSP. Section IV presents the process of feature extraction and the establishment and validation of the decoding model. Section V presents all the results of analysis above. Section VI gives a discussion based on our result and relevant study, and presents the summary and conclusion.

II. Paradigm and Data Acquisition

A. Paradigm Design

In this paper, we took the sound detection of UAVs as a sample. We only considered the variation of the target sound and the environmental noise. In the experimental paradigm of this paper, we used the wind recorded outdoor as the environmental noise and the sound segments emitted by three different kinds of UAVs in random flight as the target sounds to be recognized (to simulate the change of the target itself). We designed the following experimental paradigm. The information about sound materials is shown in Section II.B.

The details of the paradigm design are shown in Fig. 1. During the whole process of the experiment, the subjects sat in a comfortable chair in a relaxed posture within a reasonable range, and wore an EEG collecting cap (NeuroScan SynAmps2, NeuroScan, America) and in-ear headphones, and were told to look directly at the commands on the screen in front of them not to make large movements during the experiment.

In every single trial, there were a preparation phase and an experimental phase. In the preparation phase, the screen first showed a line of words, “press any button when you are ready for the next trial”, and then the subjects were given 5 seconds to prepare. In 5 s, they could actively skip the process by pressing a button when they were ready. Then the experimental phase started.

In the experimental phase, subjects heard a sound clip containing a natural wind background sound and target sounds (via in-ear headphones). The target sound (lasting 5 s) was inserted in this sound clip at random moments. In a single trial, only one of the three kinds of target UAV sounds were inserted. The subjects were asked to judge the presence of the target and press the key “space” on the keyboard with the index finger of their right hands as soon as possible. The reason for using button pressing as a response to hearing the target was to make sure that subjects heard the target sound and calculate their reaction time (RT). Once the background sound was finished, this trial was complete, and a new trial started. The experimental paradigm was implemented by PsychoPy.

For one single session, there were 12 trials. Because we used three kinds of UAV sounds as targets, target sounds appeared in pseudo-random order (four times for each kind). After completing a session, the subjects were given a three-minute break. For the entire experiment, the total duration of the 10 sessions and breaks was no more than two hours.

B. Stimulus Materials

In this paper, the sounds of three types of UAVs were used as the target sounds for auditory target detection. DJI3, DJI Tello, and a UAV powered by a duct fan. We recorded the sounds of three kinds of UAVs in a soundproof chamber (length 15 meters, width 10 meters). For recording, single-channel microphones were placed in the middle of a soundproof chamber where the UAV flew.

The background wind noise was collected manually by the microphone of the mobile phone outdoors. While collecting, the sound collector was placed on the support rack in a fixed position outdoors. When we collected the sound, the outdoor wind was at Beaufort scale 4-5.

Since most STD studies focus on the condition of SNR (in most of the research on sound event detection, 0dB∼-5dB is considered as the low level of SNR. For example, [4], [5], and [6], we calculated the SNR of the target sound relative to the background sound. As the background noise used in this study was the wind recorded in the field, the intensity of sound changed over time. Table I shows the maximum, minimum, and average SNR of the three target sounds used relative to the same segment of background noise.

TABLE I
SNR OF EACH SOUND TARGET

UAV type	Type Index	SNR (dB) Average	Max	Min
DJI 3	1	-6.0192	-0.5339	-8.9589
Duct Fan UAV	2	-11.5521	-6.0851	-14.4806
DJI Tello	3	-14.5139	-9.0286	-17.4536

C. EEG and Behavior Data Acquisition

Continuous EEG data (sampling frequency: 1000 Hz) were recorded from 60 Ag/AgCl sintered electrodes using standardized EEG recording sites (Fp1, Fpz, Fp2, AF7, AF3, AF4, AF8, F7, F5, F3, F1, Fz, F2, F4, F6, F8, FT7, FC5, FC3, FC1, FCz, FC2, FC4, FC6, FT8, T7, C5, C3, C1, Cz, C2, C4, C6, T8, TP7, CP5, CP3, CP1, CPz, CP2, CP4, CP6, TP8, P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO7, PO5, PO3, POz, PO4, PO6, PO8, O1, Oz, O2). Electrodes were mounted on a Neuroscan EEG cap. The electrode sites of this montage were arranged according to the international 10/20-system. Forehead electrode AFz was used as a ground electrode. All electrode impedances were kept below 10 kΩ.

A total of eight students participated in the experiment. Physical examination results showed that these students had no hearing impairment or brain-related diseases, had slept well the day before the experiment, and had not taken alcohol or psychoactive drugs. All students were confirmed to have a basic understanding of drones (having heard or used one) before the experiment began. This study adhered to the principles of the 2013 Declaration of Helsinki.

The moment the target voice appeared was transmitted via parallel communication from the computer to the Neuroscan acquisition device via PsychoPy. We defined the response time of the subjects as the difference between the time the target was onset and the time the subjects responded to the button (recorded by PsychoPy). Since the later parts of this paper involve the analysis of neural representations and the establishment of decoding models, we choose to use the EEG segments time-locked to stimulus onset.

III. DATA ANALYSIS

A. EEG Preprocessing

Generally, EEG data was first baseline-corrected(using the average of the first second for each EEG 8s segment from each channel) to remove the drifting. All trials with amplitude distortion were abandoned, and the remaining trials were FIR bandpass filtered (1-49Hz). Then all the EEGs were re-referenced to the common average (CAR). Finally, independent component analysis (ICA) was applied to remove eye movement and EMG artifacts. We used EEGlab to implement the entire preprocessing above.

In particular, the ICA label algorithm in EEGlab was used to determine the category of each independent component in the process of eye movement artifacts and EMG artifacts removal. Eye movement artifacts correspond to the "eye" category and EMG artifacts correspond to the "Muscle" category. For each decomposed independent component, as long as the "eye" category discriminant confidence or "Muscle" category discriminant confidence was greater than 80%, the component was removed.

On average, 2.5 trials were discarded for each subject(due to amplitude distortion or invalid reaction time) and most of the trials (more than 98%) were finally used for ERP and ERSP analysis.

B. ERP Analysis

Previous research on auditory target detection has shown that the process of perceiving an auditory target causes EEG changes in the central region of the brain. Thus, ERPs from channels Fz, FCz, Cz, CPz were mainly focused on. We segmented preprocessed EEG signals into epochs from -5000 to 3000 ms relative to the onset of the target sound.

C. ERSP

Previous studies have shown that, in addition to ERP, the presence of target sounds affects the spectrum of EEG signals in different brain regions. Most current studies measured this change using ERSP. In this article, ERSP was calculated using "Time-Frequency Analysis" in EEGlab. We applied a three-cycle wavelet with an expansion factor of 0.5 to complete the time-frequency decomposition. Stimulus-locked epochs consisted of 200 time points between –1000 and 2000 ms relative to the target sound onset. Computations were based on frequencies ranging from 1.2 to 20 Hz with a step of 0.1 Hz. This band covers the delta, Theta, alpha, and beta bands of the EEG rhythm.

IV. AUDITORY PERCEPTION DECODING MODEL

A. Dataset Establishment and Feature Extraction

The EEG decoding model is used to detect the sound target, so the decoding model needs to distinguish between the EEG signals corresponding to the subjects' "normal state" and "perceived sound target". In this paper, two kinds of EEG features were used in the decoding model, namely, EEG amplitude and time-frequency power spectrum.

For each trial, the EEG signals in the [−3s,−2s] interval were taken as non-target samples and EEG signals in the interval [0s,1s] as target samples (RT data showed that basically all subjects detected targets in this interval, see section V.A). We down-sampled the sampling frequency of the 60-channel EEG 1s segment to 100 Hz.

For all selected samples, we extracted the time-frequency features using the short-time Fourier Transform (STFT). The Time-Frequency(TF) information of each channel was a matrix with F row and T column, and we rearranged the time-frequency information from 60 channels (with a size of [F,T,60]) into a 1-dimensional vector as the raw time-frequency feature vector of this sample. The selected frequency band was set to 1-12Hz. All the raw feature vectors obtained from target and non-target samples constituted the Time-Frequency data set. In addition, the STFT was implemented through the Spectrogram() function in MATLAB with a window length of 32 and a window shift of 20.

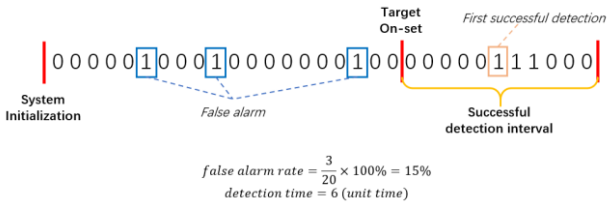

Fig. 2. Definition of performance metrics.

B. Off-Line Training and Validation

In Section IV.A, the raw feature vector of the samples contained all the amplitude or the time-frequency information of the selected frequency band. We used principal component analysis (PCA) to compress the original feature vectors. In this paper, the PCA eigenvalue contribution rate threshold was 90%, and the dimension changes of all subjects' data sets before and after compression will be shown in the results.

We chose Support Vector Machine(SVM) as the algorithm of sample classification and measured the performance of our decoding model through the 5-fold cross-validation method. In the process of SVM training, we used the built-in 'OptimizeHyperparameters' option of Matlab to optimize all the hyperparameters of the SVM model, including 'BoxConstraint', 'KernelFunction' (i.e., the type of KernelFunction), 'KernelScale', 'PolynomialOrder' and 'Standardize'. The optimization algorithm was sequence minimum optimization (SMO), and the number of iterations was 120 (See Matlab's description of fitcsvm () for more information).

C. Pseudo Online Test

To further validate the performance of the proposed sound target detection method, we used pseudo online testing to calculate the detection rate, false alarm rate, and detection time of the method. As shown in Fig. 2, the detection rate was defined as the proportion of the trials detected within 2 s of the appearance of the target in all trials. The false alarm rate was defined as the proportion of false alarm (misjudging non-target as "target") commands output by the BCI in all non-targets. The detection time was defined as the lag time between the earliest and correct detection of the target and the occurrence of the target.

During the training of the SVM model used in the pseudo online test, the process was the same as the offline part, the only difference was that the number of non-target samples had been expanded. We added the EEG signals corresponding to $[-3s,-2s]$ and $[-1s,0s]$ to the non-target samples. That is, each trial generated three non-target samples and one target sample. To avoid the model shifting to one of the two classes in the training process, we adjusted the misclassification cost of non-target and target samples to 1:3 (originally 1:1).

It should be noted that for each subject, 80% trials (randomly divided) were used to train ICA unmixing matrix (for removing eye-moving artifact) in EEG preprocessing, PCA feature compression matrix, and SVM model, and the remaining 20% trials were used for testing. After completing the bandpass filtering and common average re-reference of the test EEG set, we directly used the information from the training set for eye-moving artifact removal and feature extraction of the test set.

After the training of the model and preprocessing of the test EEG set were completed, the pseudo-online test was started. During the test, the EEG window length was 1s, starting from the -2s point (that is, the first EEG segment was $[-2s,-1s]$), and the window shift was 50 ms. On this basis, the detection threshold $Threshold$ was set (2, 3, 4, and 5, respectively). If the model judges "target appearance" for consecutive $Threshold$ times, it is considered that a sound target appeared at this time; otherwise, it is considered that no sound target appeared.

Besides the normal pseudo-online test, we cared about the generalization performance of the detection model trained. To testify the generalization performance, we used the EEG from all subjects to train a generic model. During the process of training, we only used the EEG trials corresponding to 2 targets (namely, tello UAV and duct fan UAV), and all the EEG trials corresponding to DJI3 UAV consisted of the test dataset. The other remaining steps were the same as the normal pseudo-online test. The detection performance (detection rate, false alarm, and detection time) was calculated on the training dataset and test dataset, respectively.

V. Results

A. Reaction Time Analysis

The results of RT statistical analysis are shown in Table II. We saw that the detection rates of all eight subjects were high (all over 97%, with an average of 99.08%; the detection rates of Subjects 2, 4, 5, and 7 were 100%). The average RT of 4 subjects (1, 2, 4, 7) was around 0.7 seconds, whereas the RT of the other four subjects was about 0.5 seconds (3, 5, 6, 8). For the successful trials, the standard deviation of RT was around 20-196 ms. Thus, the neural signature corresponding to the perceived sound target was located within 1s of the target's appearance in the vast majority of successful trials.

B. Neural Signature Results

1) ERP Results: In summary, after averaging the data from all subjects, we observed significant ERP in the central area. Fz, FCz, Cz, and Pz were selected as the representatives of all channels to briefly display the results.

For the early stages of ERP, as shown in Fig. 3. ERP results showed that there was a significant negative offset at around 130 ms (with an amplitude of around 0.8μV). We showed only four channels near the central area. In fact, this offset was widespread in those channels near the fronto-central region of the scalp and showed an obvious left-right symmetry. The topologies of 130 ms ERP are also demonstrated in Fig. 4.

Besides the early stage after the appearance of the stimulus, a significant positive shift was observed near the frontal parietal and parietal lobes at 300 ms after the appearance of the stimulus with a magnitude of about 2.5 μV (Fz channel). The amplitude of this positive offset decreased progressively from the frontal to parietal lobes, as shown in Fig. 3 and Fig. 4 (320ms). In addition, at the subsequent 570 ms, there

TABLE II
RT STATISTICS OF EACH SUBJECT

	Sub.1	Sub.2	Sub.3	Sub.4	Sub.5	Sub.6	Sub.7	Sub.8	average
SUCCESS RATE	0.9924	1	0.9924	1	1	0.9750	1	0.9666	0.9908
$\overline{RT}$	0.7359	0.6305	0.4735	0.7995	0.4765	0.5145	0.7641	0.5628	0.6196
SD	0.1960	0.1574	0.1171	0.1570	0.0980	0.1685	0.1479	0.0219	0.1329
$\mathbf{P(RT < 1)}$	0.9242	0.9722	0.9924	0.89394	1	0.9666	0.9015	0.95	0.9501

In the table, SUCCESS RATE represents the proportion of trials with RT<2 in all trials; $\overline{RT}$ and SD represent the average RT and standard deviation of correct RT, respectively; $\mathbf{P(RT < 1)}$ represents the proportion of trials with RT<1 in all detection-success trials

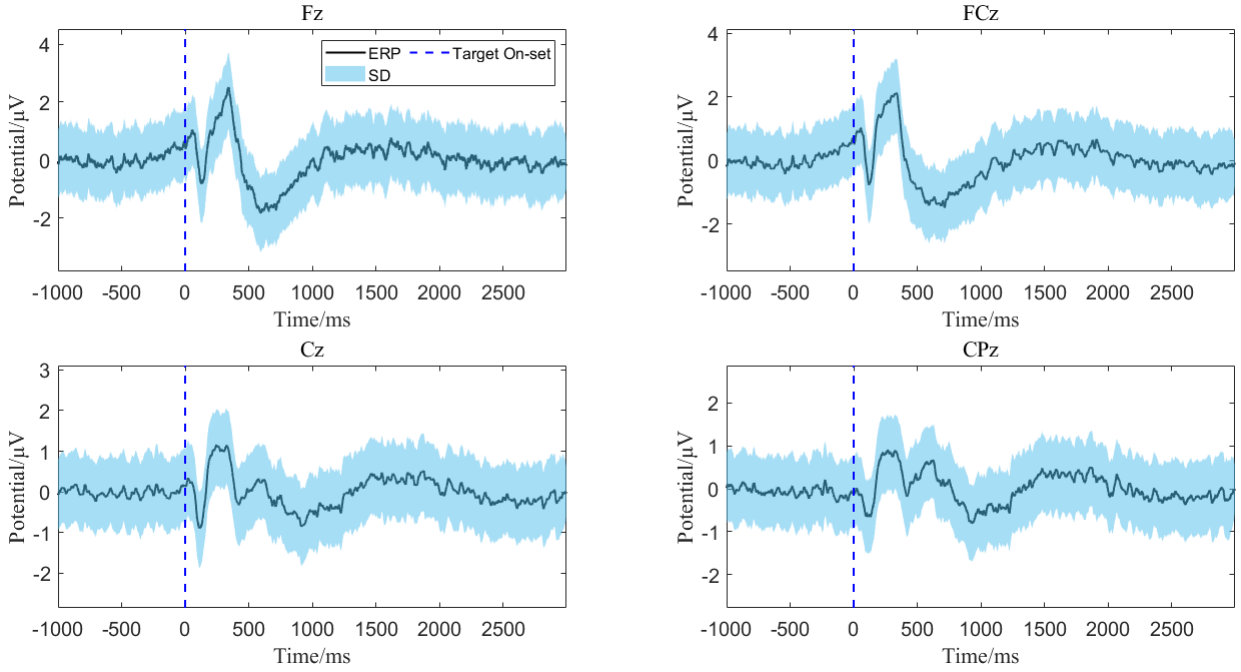

Fig. 3. ERP from Fz,FCz,Cz, and CPz.

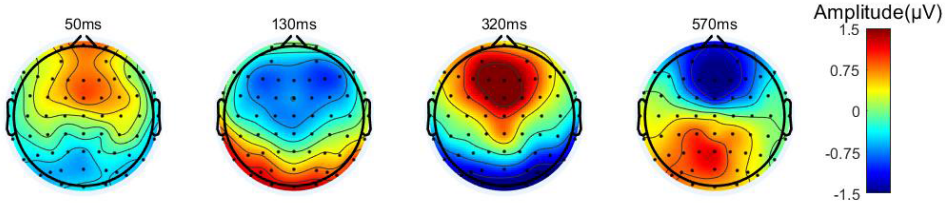

Fig. 4. ERP topology.

was a weak positive shift in the parietal lobe region, centered on the Pz channel, with a magnitude of about 1.25 μV.

In the experimental paradigm, we used three different target sounds. In Fig. 5, we gave the ERP waveforms corresponding to three different targets. It can be seen that for the early component of ERP, the N100 component induced by three target sounds had the nearly same amplitude. For the late component of ERP, P3 induced by target sound 3 reached the highest amplitude, while the P3 component of target sound 1 and target sound 2 were only slightly different. On the whole, the waveforms of ERP induced by the three target sounds were the same.

2) ERSP Results: Fig. 6 shows the time-frequency information changes of EEG signals in six main channels located in the central region before and after the appearance of the target sound. After the appearance of the target sound, the energy changes of the six channels of EEG signals were significant at 1-6Hz (covered Delta rhythm) and 8-14Hz (Alpha rhythm). Specifically, the energy of delta rhythm increased significantly after the presence of the target, lasting 800 to 1000 ms and

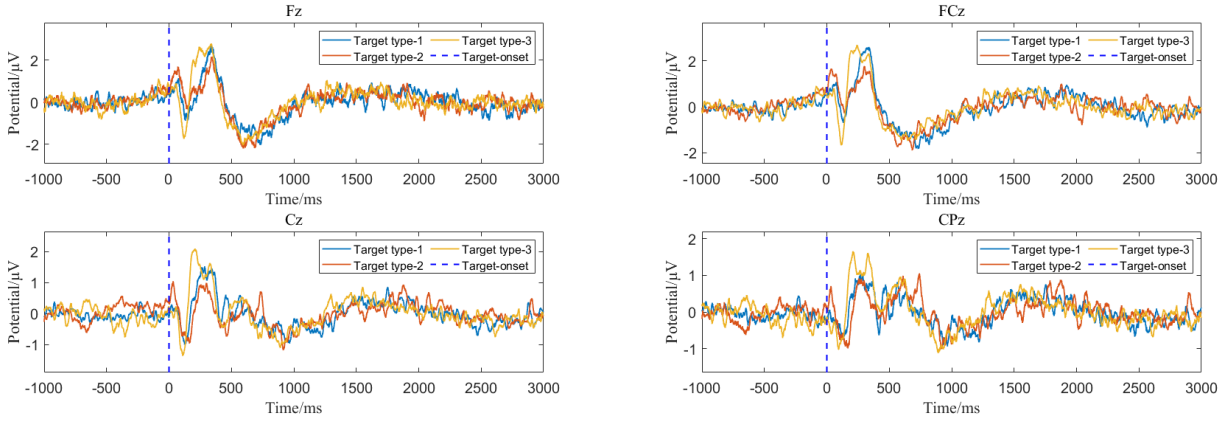

Fig. 5. ERP similarity between different kinds of targets.

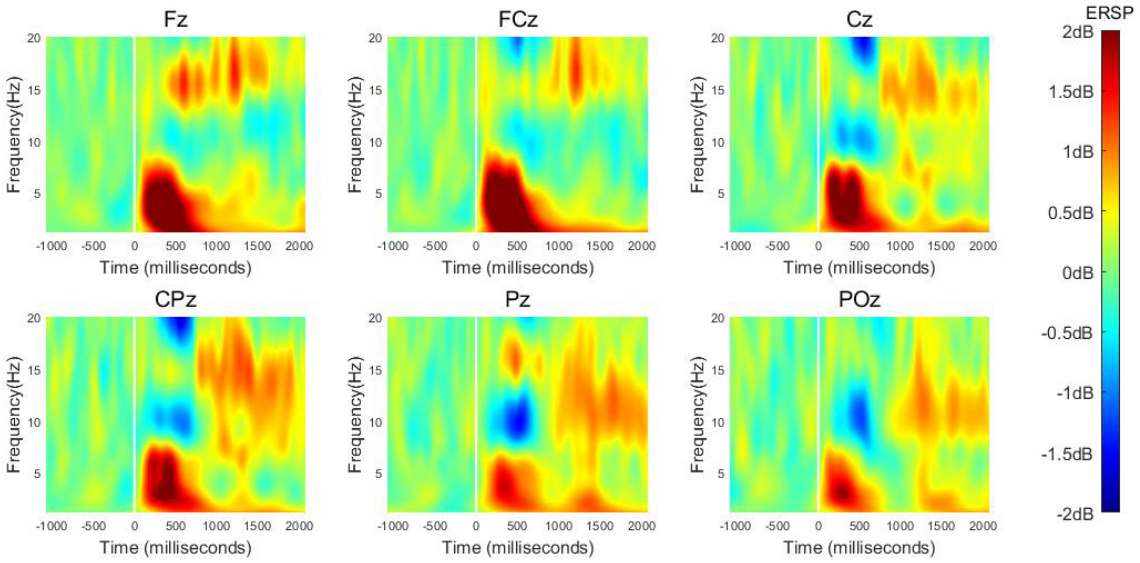

Fig. 6. ERSP of EEG signals from central area channels.

peaking around 300 ms. The energy of alpha rhythm decreased significantly after the presence of the target, and the duration varied from channel to channel. CPz and Pz channels were the most significant in terms of the magnitude of the decline.

Similar to ERP, we also examined the ERSP of the three target sounds in the P1, Pz and P2, as shown in Fig.7(From top to bottom, the first row corresponds to P1, the second row corresponds to Pz, and the third row corresponds to P2). For the delta band, the presence of three different target sounds all caused the energy of this band to rise (regardless of which channel). The energy changes in the alpha band caused by the three targets are slightly different. The energy drop in the alpha band induced by target type1 and 3 have similar intensity in all three channels. That induced by target type2 was slightly weaker and not obvious in P1 and Pz channels. In general, ERSP induced by the three target sounds showed the same change trend, with only a difference in amplitude.

C. Decoding and Detection Performance

The performance of the decoding model is shown in Fig. 8 and Table III, including offline classification accuracy and pseudo-online detection rate, false alarm rate, and detection time. For the offline classification test, the average accuracy of 8 subjects reached 81%, and 6 of them were over 80%, with a small variation range. The highest accuracy came from the decoding model for Subject 3, and the average accuracy was 85% in the 5-fold cross-validation test.

In the pseudo-online test, different detection thresholds were set. When the threshold was 2, the average detection rate of the 8 subjects was 84%, the false alarm rate was 6%, and the average detection time was 0.817 seconds. With the increasing detection threshold (from 2 to 5), the average detection rate decreased to 69%, the false alarm rate decreased to 2%, and the detection time increased to 0.981s.The best detection performance came from Subject 8. When the detection threshold was 2 and 3, the detection rate was 100%, and the false alarm rate was 3%-4%. When the detection threshold increased to 5, the detection rate remained at 96%, and the false alarm rate decreased to 1%.

To further show the performance of the proposed method, we chose a sound target detection algorithm in DCASE 2020 (Detection and Classification of Acoustic Scenes and Events) Task 2 as our benchmark[19]. This algorithm can be briefly summarized as a detection algorithm based on

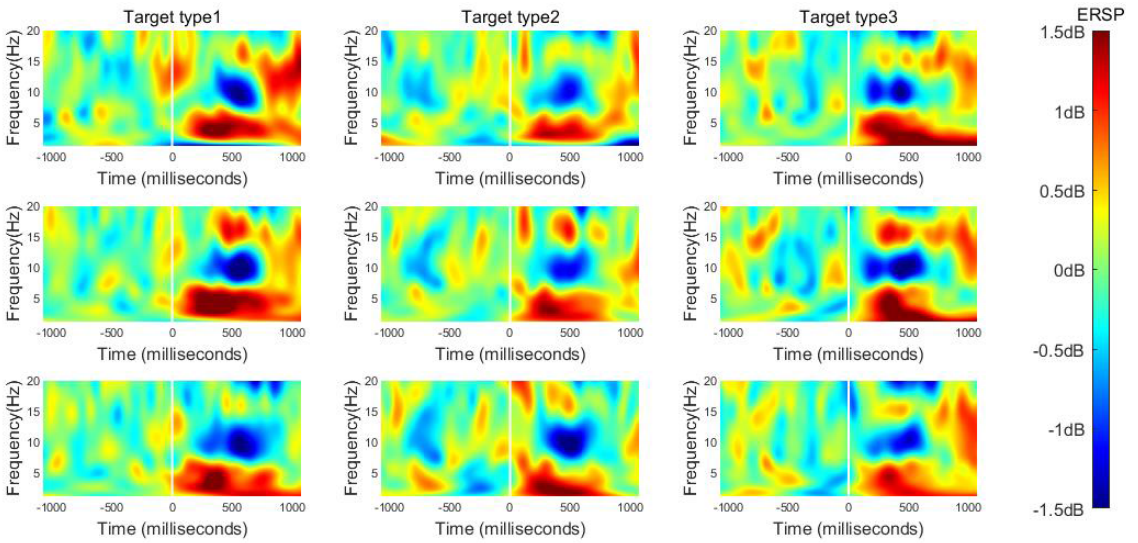

Fig. 7. ERSP difference and similarity between different kinds of targets.

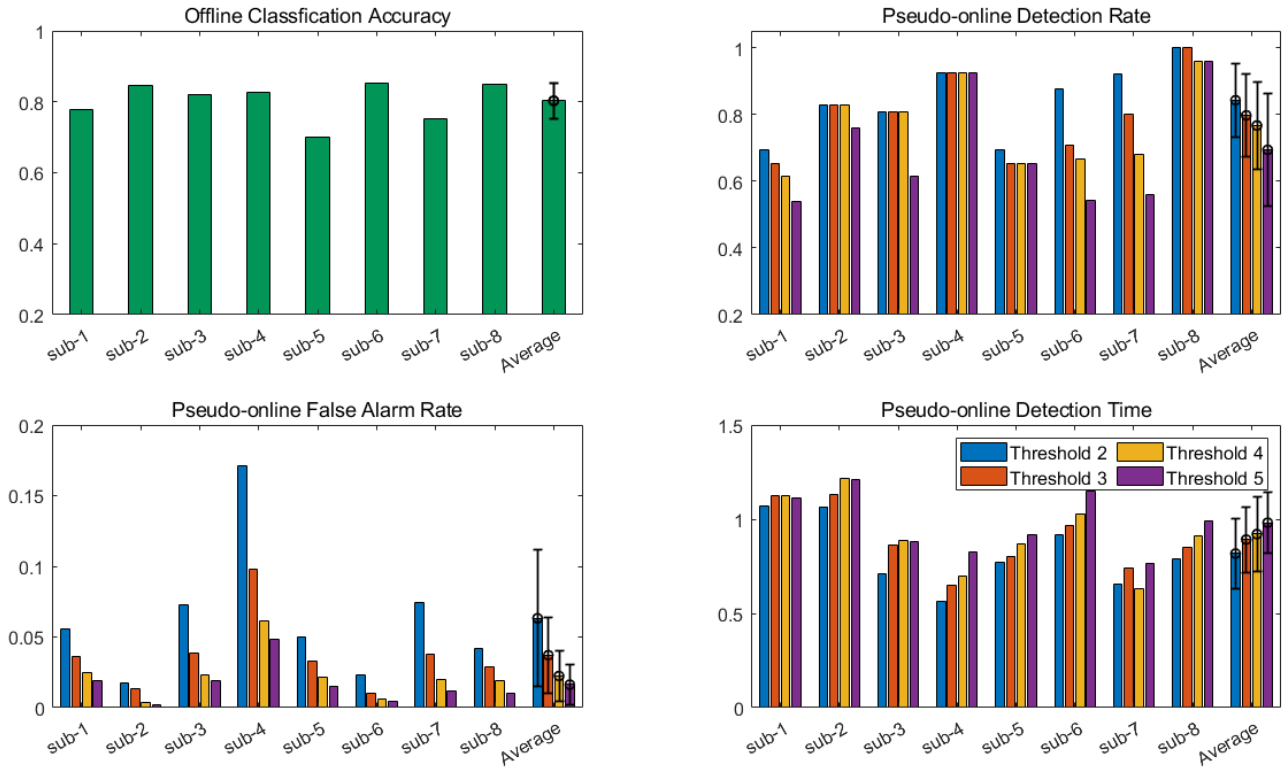

Fig. 8. Offline decoding and pseudo-online detection performance for each subject.

ResNet network, which extracts the logarithmic MEL frequency feature (128*128) of sound signals, and outputs the two-dimensional detection results through operations, such as convolution and pooling. The training dataset consists of 306 targets and 306 non-targets, including target sounds, such as Tello UAV and Duct Fan UAV. The test dataset consisted of 114 target samples and 114 non-target samples, and only contained DJI3 target sound. The batch size was set to 32 in the training process, and the Adams parameter optimizer was used. The training lasted for 500 epochs. When the training was completed, the loss had converged.

We defined the detection performance of the two methods in the face of a new target (i.e., the test set) as the generalization performance. We calculated the detection performance of the two methods on the training set and the test set, respectively, as shown in Table IV. Both methods achieved high detection rate on the raining set(97.50% vs 99.85%), whereas the false alarm rate of the BCI method was higher. The biggest

TABLE III
PSEUDO-ONLINE DETECTION PERFORMANCE OF DECODING MODEL ACROSS SUBJECTS

Subject	Detection Rate (different threshold)				False Alarm Rate				Detection Time(s)			
	2	3	4	5	2	3	4	5	2	3	4	5
1	69%	65%	62%	54%	6%	4%	3%	2%	1.071	1.125	1.123	1.114
2	83%	83%	83%	76%	2%	1%	0%	0%	1.066	1.133	1.216	1.208
3	81%	81%	81%	62%	7%	4%	2%	2%	0.710	0.865	0.886	0.882
4	92%	92%	92%	92%	17%	10%	6%	5%	0.564	0.652	0.700	0.825
5	69%	65%	65%	65%	5%	3%	2%	2%	0.768	0.804	0.866	0.916
6	88%	71%	67%	54%	2%	1%	1%	0%	0.917	0.969	1.029	1.148
7	92%	80%	68%	56%	7%	4%	2%	1%	0.653	0.740	0.634	0.767
8	100%	100%	96%	96%	4%	3%	2%	1%	0.789	0.850	0.914	0.993
Average	84%	80%	77%	69%	6%	4%	2%	2%	0.817	0.892	0.921	0.981

TABLE IV
PERFORMANCE COMPARISON ON DIFFERENT DATASET

Method	Training Set			Test Set		
	Detection Rate	False Alarm Rate	Detection Time(s)	Detection Rate	False Alarm Rate	Detection Time(s)
Benchmark (Network from DCASE)	97.50%	0%	0.700	43.00%	0%	1.57
BCI method (Proposed)	99.85%	9.75%	0.445	84.84%	3.28%	0.7758

During the generalization performance test, the detection thresholds of Benchmark method and BCI method (proposed) were all set to 3.

difference between the two performance was in detection rate on the test set (43.00% vs 84.84%). It can be seen from the results that the detection rate of the Benchmark method decreased significantly (97.50% vs 43.00%) in the face of targets that do not appear in the training set. However, the detection rate of the BCI method only showed a relatively smaller decrease (99.85% vs 84.84%). In addition, the new target resulted in an increase in the average detection time for both methods. However, there was a smaller increase for the BCI method (0.33s vs 0.87s) than the benchmark.

VI. DISCUSSION AND CONCLUSION

In this paper, an STD method based on a BCI was proposed to solve the problem of STD in the natural sound field with a low SNR. We designed the experimental paradigm according to the real STD scenario and analyzed neural representations, including ERP and ERSP of the EEG signals induced by the presence of target sound. We extracted two different recognition features according to the observed neural representation. Furthermore, we established an SVM-based EEG decoding model to distinguish target and non-target, showing the feasibility of using a BCI to detect a real sound target in low SNR conditions. This work can lay a foundation of the research and development of EEG-based STD.

From the perspective of neural signature, the early components of ERP showed clear N100 (N1) components, indicating that the target sound entered the auditory perception pathway. On the other hand, the late component of ERP showed a clear P3 component, indicating that the target sound caused the cognitive activities of the subjects. The ERP waveform obtained by us was similar to the results obtained by Gabriela et al. [17] and Sujoy et al. [18]. In essence, these ERP results reflect the cognitive activity of a few "deviant" stimuli from the perspective of EEG signals.

In addition, ERSP results showed that the presence of targets caused a significant decrease in alpha rhythm energy. The results of neurological studies showed that the change of alpha rhythm energy was related to the change of subjects' attention, such as selective attention [13], [14], [15], target detection, and localization [11], [12]. Because the target sound in the experimental paradigm was not as short as other literature experiments of tens to hundreds of milliseconds. The reduced alpha energy may reflect the attention of subjects to the target sound in a period after the appearance of P3, which was consistent with [16] and other previous literature on alpha ERD.

We selected the time-frequency feature and built decoding models. The results showed that our EEG decoding model achieved reliable off-line decoding performance (the average accuracy of the optimal model for a single subject is 81%). The pseudo-online test results showed that our decoding model had a good detection rate, acceptable false alarm rate, and fast response time. On the whole, our results showed that in the real environment, human perception of target sound under low SNR conditions can be captured through EEG Decoding, which can be used to perform sound target detection.

In addition to its reliable detection performance at the low SNR, the proposed method has another advantage, namely, its robustness to different sound targets. As can be seen from the generalization test result, the performance of the benchmark method dropped severely when the unseen sound target appeared, whereas the proposed method still had good

detection performance. Because the method proposed in this paper does not rely on the sound data itself, but on the prior knowledge of the sound environment and the ability to infer from human perception and recognition, we believe that as long as the target sound belongs to the category of UAVs, the neural signature corresponding to the new sound target should be close to the same as the three presented in this paper, combined with our feature extraction process, this characteristic makes the scheme proposed in this paper have a certain robust performance for STD. In the practical use, this characteristic makes BCI developers need to worry less about the sound materials used in the paradigm, thus reducing the cost of training, and developers also can take this method as a complementary or collaborative approach to perform a more robust target detection.

There are still some limitations in this paper, which need to be further improved in our future work. First, in the aspect of decoding EEG signals, we used SVM to make a basic attempt, and the decoding model in this paper has some room for improvement from the perspectives of channel selection, feature extraction, and classifier algorithm. Second, in this paper, the three sound targets all belong to the category of low SNR. Thus, the relationship between decoding performance and SNR of sound targets were not investigated. Such relationship should be explored. In addition, online verification of the whole acoustic target detection system is an important issue. In the final detection system, this method can co-work with an automated detection algorithm (for example in this paper, the benchmark method).

REFERENCES

[1] B. Yang, E. T. Matson, A. H. Smith, and J. E. Dietz, and J. C. Gallagher, "UAV detection system with multiple acoustic nodes using machine learning models," in *3rd IEEE Int. Conf. Robotic Comput. (IRC)*, Naples, Italy, Feb. 2019, pp. 493–498, doi: 10.1109/IRC.2019.00103.

[2] E. R. Solis, D. V. Shashev, and S. V. Shidlovskiy, "Implementation of audio recognition system for unmanned aerial vehicles," in *Proc. Int. Siberian Conf. Control Commun. (SIBCON)*, Kazan, Russia, May 2021, pp. 1–8, doi: 10.1109/SIBCON50419.2021.9438906.

[3] C. Clavel, T. Ehrette, and G. Richard, "Events detection for an audio-based surveillance system," in *Proc. IEEE Int. Conf. Multimedia Expo*, Amsterdam, The Netherlands, Jul. 2005, pp. 1306–1309.

[4] I. Papadimitriou, A. Vafeiadis, A. Lalas, K. Votis, and D. Tzovaras, "Audio-based event detection at different SNR settings using two-dimensional spectrogram magnitude representations," *Electronics*, vol. 9, no. 10, p. 1593, Sep. 2020, doi: 10.3390/electronics9101593.

[5] X. Ren, Z. Feng, H. Lu, and Q. Zhou, "Learning target template for acoustic event detection from low-SNR training data," *IEEE Access*, vol. 9, pp. 84490–84500, 2021, doi: 10.1109/access.2021.3087713.

[6] L. Shi, I. Ahmad, Y. He, and K. H. Chang, "Hidden Markov model based drone sound recognition using MFCC technique in practical noisy environments," *J. Commun. Netw.*, vol. 20, no. 5, pp. 509–518, Oct. 2018, doi: 10.1109/jcn.2018.000075.

[7] N. Turpault, R. Serizel, J. Salamon, and A. P. Shah, "Sound event detection in domestic environments with weakly labeled data and soundscape synthesis," in *Proc. Detection Classification Acoustic Scenes Events Workshop (DCASE)*, 2019, pp. 1–5.

[8] N. Turpault et al., "Sound event detection and separation: A benchmark on desed synthetic soundscapes," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 840–844, doi: 10.1109/ICASSP39728.2021.9414789.

[9] C. R. Brydges and F. Barceló, "Functional dissociation of latency-variable, stimulus- and response-locked target P3 sub-components in task-switching," *Frontiers Hum. Neurosci.*, vol. 12, p. 60, Feb. 2018, doi: 10.3389/fnhum.2018.00060.

[10] A. D. Gerson, L. C. Parra, and P. Sajda, "Cortically coupled computer vision for rapid image search," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 14, no. 2, pp. 174–179, Jun. 2006, doi: 10.1109/tnsre.2006.875550.

[11] L.-I. Klatt, S. Getzmann, E. Wascher, and D. Schneider, "The contribution of selective spatial attention to sound detection and sound localization: Evidence from event-related potentials and lateralized alpha oscillations," *Biol. Psychol.*, vol. 138, pp. 133–145, Oct. 2018, doi: 10.1016/j.biopsycho.2018.08.019.

[12] J. Zimmermann, B. Ross, M. Moscovitch, and C. Alain, "Neural dynamics supporting auditory long-term memory effects on target detection," *NeuroImage*, vol. 218, Sep. 2020, Art. no. 116979, doi: 10.1016/j.neuroimage.2020.116979.

[13] A. Strauss, M. Wostmann, and J. Obleser, "Cortical alpha oscillations as a tool for auditory selective inhibition," *Frontiers Hum. Neurosci.*, vol. 8, p. 350, May 2014, doi: 10.3389/fnhum.2014.00350.

[14] J. J. Foxe and A. C. Snyder, "The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention," *Frontiers Psychol.*, vol. 2, p. 154, Jul. 2011, doi: 10.3389/fpsyg.2011.00154.

[15] A. Ikkai, S. Dandekar, and C. E. Curtis, "Lateralization in alpha-band oscillations predicts the locus and spatial distribution of attention," *PLoS ONE*, vol. 11, no. 5, May 2016, Art. no. e0154796, doi: 10.1371/journal.pone.0154796.

[16] E. G. Blundon and L. M. Ward, "Search asymmetry in a serial auditory task: Neural source analyses of EEG implicate attention strategies," *Neuropsychologia*, vol. 134, Nov. 2019, Art. no. 107204, doi: 10.1016/j.neuropsychologia.2019.107204.

[17] G. M. Pawlowski et al., "Brain vital signs: Expanding from the auditory to visual modality," *Frontiers Neurosci.*, vol. 12, p. 968, Jan. 2019, doi: 10.3389/fnins.2018.00968.

[18] S. G. Hajra et al., "Developing brain vital signs: Initial framework for monitoring brain function changes over time," *Frontiers Neurosci.*, vol. 10, p. 211, May 2016, doi: 10.3389/fnins.2016.00211.

[19] P. Primus, "Reframing unsupervised machine condition monitoring as a supervised classification task with outlier-exposed classifiers," DCASE 2020 Challenge, Inst. Comput. Perception, Johannes Kepler Univ. Austria, Linz, Austria, Tech. Rep., Jul. 2020. [Online]. Available: https://dcase.community/documents/challenge2020/technical_reports/DCASE2020_Primus_36_t2.pdf