Toward Robust Machine-Learning-Based Discrimination of a 1 Hz Difference in the Auditory Cortex Using fMRI Activation Patterns

Yoshitaka Ooyashiki
Kochi University of Technology
Tosayamada, Kami, Kochi, 782-8502, Japan
e-mail: ooyashikiyoshitaka@gmail.com

Kyoko Shibata
Kochi University of Technology
Tosayamada, Kami, Kochi, 782-8502, Japan
e-mail: shibata.kyoko@kochi-tech.ac.jp

Abstract - **Brain decoding using functional magnetic resonance imaging (fMRI) has attracted considerable attention as a method for investigating how stimulus-related information is represented in patterns of human brain activity and for examining whether subtle differences in sensory input can be reflected in brain activation patterns. In this study, we investigated whether machine-learning-based classification could discriminate between two auditory stimuli differing by 1 Hz using fMRI activation images obtained from the auditory cortex. This study extends our earlier work by examining the reproducibility of the method across multiple participants and by comparing a three-dimensional convolutional neural network (3DCNN) with a support vector machine (SVM). In the individual analysis, the 3DCNN showed consistently high classification accuracy across participants under the present experimental setting, and its performance was higher than that of the SVM. In the group analysis, both the 3DCNN and the SVM showed more variable performance across participants, although several train-test combinations achieved accuracy levels around or above 60%. These results suggest that the proposed framework is particularly effective in participant-wise analysis, whereas cross-participant application remains more challenging under the current conditions. Overall, the findings support the feasibility of fine-grained auditory discrimination from fMRI activation patterns and provide additional motivation for further investigation using more rigorous evaluation procedures, improved model-selection strategies, and more diverse participant cohorts.**

Keywords-fMRI; Brain decoding; Convolutional Neural Network; Support Vector Machine.

I. INTRODUCTION

Brain decoding is a technology that interprets physical and psychological states from measured brain activity, and functional magnetic resonance imaging (fMRI) has been widely used for this purpose in neuroscience. Because fMRI data consist of a large number of voxels, Multivoxel Pattern Analysis (MVPA) is commonly employed for analysis. In recent years, machine learning techniques have attracted considerable attention as part of the MVPA framework and have contributed to advances in brain decoding research. To date, many studies have investigated the use of fMRI to decode visual stimuli. For example, Liang et al. [2] examined emotion decoding from visual cortex images using a Support Vector Machine (SVM) and a Convolutional Neural Network (CNN). Horikawa et al. [3] also investigated the decoding of objects seen in dreams by applying Deep Neural Network (DNN) features to visual cortex data. In contrast, research on fMRI-based auditory cortex decoding has lagged behind that on visual cortex decoding. One reason is that human perceptual information is predominantly visual, which has historically led to greater emphasis on the visual cortex. In addition, scanner noise during fMRI acquisition makes it more difficult to detect neural activity in the auditory cortex. However, improvements in fMRI technology, including noise-canceling mechanisms, have made auditory cortex research increasingly feasible.

In studies of auditory cortex decoding, many investigations have focused on music-related targets and have reported the identification or classification of genres and moods, such as upbeat songs, somber songs, or mood-lifting songs, from brain activity [4][5][6]. For example, Casey et al. [6] developed and validated an SVM-based classification method using fMRI data acquired during auditory stimulation with 25 songs from five musical genres. Many of these studies have focused on qualitative musical characteristics such as mood and genre. In contrast, relatively few studies have examined quantitative musical characteristics, particularly the three fundamental elements of sound: frequency, sound pressure, and timbre.

To better understand fine-grained auditory processing, it is important to investigate not only qualitative but also quantitative sound characteristics. Accordingly, our research group has focused on decoding quantitative musical characteristics. Shigemoto et al. [7], for example, focused on frequency and used deep learning to discriminate between two sounds differing by 124.5 Hz from brain activation images, achieving an accuracy of 75%. Building on this result, the present study aims to examine even smaller frequency differences. At present, this study is positioned as fundamental neuroscience, and possible clinical application remains a topic for future investigation.

Our research group has defined the auditory cortex as a Region of Interest (ROI) based on the concept of tonotopy in order to capture brain responses to auditory stimuli. Tonotopy refers to the spatial organization of frequency-specific responses within the auditory cortex. Tonotopy has been reported in many previous studies [8][9][10][11], particularly in Brodmann Areas (BA) 41 and 42, corresponding to the primary auditory cortex in Heschl's gyrus. These studies typically examined a wide frequency range and often focused on continuous frequency modulation. However, the possibility that tonotopy may remain informative even for very small frequency differences has not been sufficiently explored.

Therefore, to investigate smaller frequency differences than those examined by Shigemoto et al. [7], this study hypothesizes that tonotopy may still function even for

frequency differences that are effectively inaudible to humans. Healthy individuals can typically distinguish frequency differences of approximately 5 Hz, whereas differences of 2–3 Hz are more strongly influenced by musical experience and individual variation. In this study, a 1 Hz difference between two sounds is treated as a frequency difference that is difficult for humans to perceive.

Accordingly, the objective of this study is to examine a machine-learning-based method for discriminating between two sounds differing by 1 Hz from fMRI activation patterns. Machine learning methods were adopted because the difference in brain activation between such closely spaced frequencies is expected to be subtle. To capture these subtle activation patterns in high-dimensional fMRI data, machine learning techniques, as part of the MVPA framework, were employed. In addition, such techniques may be relevant to future methodological developments in medical and assistive technologies.

The present study builds on two lines of our previous work. First, the imaging design was examined in earlier studies [12][13]. Two major imaging designs are commonly used in fMRI experiments: the event-related design and the block design. The event-related design uses shorter stimulus durations and can provide a larger number of data points, whereas the block design uses longer stimulus durations and generally yields clearer activation images. Because the present study focuses on very small frequency differences, the higher image clarity of the block design was considered advantageous for feature extraction. At the same time, the limited number of data points obtainable with the block design raises concerns about the amount of data available for model training. Comparative experiments in our previous work [12][13] showed discrimination accuracies of 55.9% for the event-related design and 63.4% for the block design, suggesting that the block design was more suitable for 1 Hz difference discrimination.

Second, based on these findings, our previous study [1] re-examined the ROIs to improve discrimination accuracy. In the earlier studies [12][13], the ROIs were limited to BA41 and BA42. In the later study [1], the ROI definition was expanded based on the hypothesis that additional activation information would improve feature extraction for subtle activation patterns. In addition to BA41 and BA42, BA22 was included as an ROI because previous work has also suggested tonotopic organization in BA22 [14]. Using this expanded ROI definition, discrimination accuracy reached 60.4% with the event-related design and 100% with the block design in participant-wise analysis. These findings suggested that using a block design with ROIs in BA41, BA42, and BA22 was an effective configuration for 1 Hz difference discrimination using a 3DCNN.

Based on these previous findings, the objective of the present study is to further evaluate the 1 Hz difference discrimination framework from two perspectives.

The first objective is to examine reproducibility across multiple participants. Because the previous studies [1][12][13] involved only a single participant, it remained unclear whether the observed high accuracy was specific to that participant. Therefore, the first objective of this study is to determine whether similar results can be obtained across multiple participants under the same experimental framework.

The second objective is to examine the suitability of a 3DCNN for 1 Hz difference discrimination by comparing it with an SVM. In general, studies employing machine learning methods often compare a proposed method with another learning algorithm in order to assess its relative usefulness [2][15][16]. In this study, the 3DCNN was selected as the main discrimination method because it can extract features while preserving the spatial structure of brain activation images. However, because fMRI requires a specialized experimental environment, the amount of data available for training is limited. Since deep learning methods generally require larger datasets, an SVM, which is often effective for smaller datasets, was selected as a baseline method for comparison.

To address this second objective, both individual and group analyses were conducted. In the individual analysis, training and testing were performed using data from the same participant, representing a participant-wise discrimination framework applicable when training data can be acquired in advance. In the group analysis, one participant was used for testing and the remaining participants were used for training, thereby providing an initial evaluation of cross-participant applicability in cases where participant-specific prior data may not be available.

The remainder of this paper is organized as follows. Section II describes the acquisition of fMRI data and the discrimination procedures. Section III presents the experimental results. Section IV discusses the findings from the perspectives of reproducibility and comparative model performance. Section V concludes the paper and outlines future work.

II. METHOD

The following procedure is to be followed. Acoustic stimuli consisting of between two sounds with a 1 Hz difference are presented using an fMRI scanner to acquire brain activation images. These images are annotated using SPM12 to provide input to the 3DCNN and SVM. The annotated brain activation images are converted into 3D data for the 3DCNN and into 1×N dimensional data for the SVM. The analysis is performed using the converted data for each of the four methods (3DCNN Method in Individual Analysis, SVM Method in Individual Analysis, 3DCNN Method in Group Analysis, and SVM Method in Group Analysis). A detailed explanation is provided in the subsequent section.

A. Participants

Six adult male participants without hearing impairments took part in this study. One of these participants was the same individual included in our previous study [1], for whom fMRI data had already been acquired; therefore, new experiments were conducted for the remaining five participants. This study was conducted in accordance with the ethical standards of the Ethics Review Committee of Kochi University of Technology, and informed consent was obtained from all participants after a full explanation of the experimental procedures.

B. fMRI Experiment

In this study, a SIEMENS 3-Tesla fMRI scanner (MAGNETOM Prisma 3T) [17] and a 64-channel head coil [17], both housed at the Brain Communication Research Center of Kochi University of Technology, were used, as shown in Figures 1 and 2. The functional imaging parameters are listed in Table I, and the structural imaging parameters are listed in Table II. During the fMRI experiment, each participant lay in a supine position inside the scanner bore, and the head was stabilized within the head coil to reduce motion artifacts during image acquisition. This arrangement was intended to maintain a consistent measurement condition throughout the experiment and to improve the reliability of the acquired activation images.

Auditory stimuli were presented to the participants through an RME FireFace UCX USB audio interface connected to a Windows PC. The OptoACTIVE headphones [18] used in this study, shown in Figure 3, were equipped with active noise cancellation. This auditory presentation setup was adopted to deliver the stimuli as consistently as possible in the MRI environment, where scanner noise can interfere with auditory perception and the stability of baseline brain activity.

Figure 1. 3-Tesla fMRI scanner (MAGNETOM Prisma 3T) [17].

Figure 2. 64-channel head coil [17].

TABLE I. FUNCTIONAL BRAIN IMAGING PARAMETERS.

Echo time (TE), ms	48
Repetition time (TR), ms	3000
Field of view (FOV), mm	192×192
Flip angle (FA), °	90
Matrix size, mm	2.0×2.0×3.0
Slice thickness, mm	3.0
Slice gap, mm	0.75
Number of slices	36
Slice acquisition order	Ascending

TABLE II. STRUCTURAL BRAIN IMAGING PARAMETERS.

Echo time (TE), ms	2.52
Repetition time (TR), ms	1900
Field of view (FOV), mm	250×250
Flip angle (FA), °	9
Matrix size, mm	1.0×1.0×1.0
Slice thickness, mm	0.5
Slice gap, mm	0.50
Number of slices	176
Slice acquisition or der	Ascending

Figure 3. OptoACTIVE headphones [18].

The use of active noise-canceling headphones therefore played an important role in improving the listening condition during data acquisition. White noise was continuously presented throughout the experiment in order to stabilize the auditory background during scanning and to reduce fluctuations in the baseline auditory state caused by scanner noise and other environmental sounds. By maintaining a more consistent auditory background, the response evoked by the target stimuli could be captured more clearly as a relative change from that baseline condition. In addition, the onset of auditory stimulation was controlled in accordance with the experimental timing so that the task periods could be

presented under a stable acquisition schedule throughout the scanning session. This design was intended to reduce unnecessary variability in the measured responses and to improve the consistency of the data used for subsequent analysis. The auditory stimuli administered to the participants were generated using Steinberg Nuendo 10.3. The stimuli consisted of two pure tones at 523 Hz and 524 Hz, with sound pressure levels ranging from 78 to 85 dB. As in our previous studies [1][12][13], a block design was adopted. Each task period lasted 9 s, and each rest period lasted 15 s. The block design was selected because it was considered advantageous for obtaining clearer activation patterns from subtle auditory differences than an event-related design. Since the present study focused on discriminating a very small frequency difference of 1 Hz, the use of a design expected to provide relatively stable activation images was regarded as appropriate for subsequent machine-learning-based analysis.

C. Annotation

Brain activation images obtained by fMRI are typically stored in the Digital Imaging and Communications in Medicine (DICOM) format, which is widely used in medical imaging devices such as computed tomography (CT). In this study, preprocessing, statistical image creation, and ROI definition were performed using Statistical Parametric Mapping 12 (SPM12). For compatibility with SPM12, the fMRI data were converted from DICOM format to Neuroimaging Informatics Technology Initiative (NIfTI-1) format before analysis.

During preprocessing, three main steps were carried out: realignment, normalization, and smoothing. The first step was realignment. Although fMRI experiments are generally conducted over relatively short periods to reduce participant burden, head motion may still occur during scanning. Therefore, motion correction was required. Realignment corrected variations in the fMRI data caused by head movement by aligning all subsequent images to the first acquired fMRI image. The second step was normalization. Because brain structures differ among participants, the data were transformed into a standard brain space to reduce structural variability across individuals. SPM12 uses the Montreal Neurological Institute (MNI) standard brain model for this purpose. The third step was smoothing, which reduces noise introduced during the realignment and normalization procedures and helps stabilize the spatial distribution of activation signals.

During statistical image creation, brain activation features were extracted by selecting multiple preprocessed images, performing statistical analysis, and generating contrast images. In this study, each statistical image used for training was generated from two preprocessed brain activity scans. This setting was adopted on the basis of our previous studies, in which the use of two scans was found to provide a practical balance between image stability and the number of available training samples [1][12][13]. If only a single scan was used, the resulting activation image tended to be more sensitive to scan-to-scan fluctuation. In contrast, combining more than two scans would reduce the number of available samples for machine-learning-based analysis. Because four scans were available within each condition, four different combinations of two scans could be constructed, and these combinations were used to increase the number of training samples while preserving the basic experimental structure.

The generated statistical images were then restricted to the ROIs, and both activation and spatial information were obtained in CSV format. Based on our previous findings [1], the ROIs were set to BA41, BA42, and BA22, which are closely related to auditory information processing. After the CSV data were obtained, the input data for the 3DCNN were converted into three-dimensional data of size H41 × W50 × D15 on the basis of spatial position information so that the spatial arrangement of the activation pattern could be preserved. In contrast, the input data for the SVM consisted of 1 × N-dimensional activation vectors that could be directly used for classification. Thus, the 3DCNN and SVM differed not only in classifier structure but also in the way spatial information was incorporated into the analysis.

This annotation procedure yielded 320 training samples and 32 test samples for each participant. Table III summarizes the numbers of training and test samples used in the individual and group analyses. These data were subsequently used to construct the datasets for both the 3DCNN- and SVM-based discrimination experiments described in the following sections.

D. Learning Algorithms

In this study, a 3DCNN, a deep learning method, and a SVM, a conventional machine learning method, were used to discriminate brain activation images. These two methods were selected because they differ substantially in the way spatial information is handled. The 3DCNN can directly process volumetric activation images while preserving their spatial structure, whereas the SVM uses vectorized input data for classification.

The 3DCNN is an extension of the CNN architecture into three dimensions and is designed to process volumetric image data. It alternates convolution and pooling operations in the spatial directions, thereby enabling feature extraction while preserving the spatial structure of brain activation images. The structure of the 3DCNN used in this study is shown in Figure 4. It consisted of three sets of convolutional and pooling layers, followed by fully connected layers immediately before the output layer. Because the task involved discrimination between two classes, the output layer contained two units. The Softmax function was used as the activation function in the output layer so that the outputs could be interpreted as class probabilities.

The training parameters used for the 3DCNN are listed in Table IV. For each parameter, several values were tested, and a grid search was conducted over the combinations shown in Table IV. Training was terminated when the training error fell below 0.1, at which point the model was saved. The trained model was then evaluated using the test data in the two-sound discrimination task. In this study, the parameter search was performed in order to examine the behavior of the model under multiple settings within the present experimental framework.

TABLE III. NUMBER OF TRAINING AND TEST SAMPLES FOR EACH ANALYSIS METHOD (THE SECOND ROW SHOWS INDIVIDUAL ANALYSIS, AND THE THIRD ROW SHOWS GROUP ANALYSIS).

Analysis	Training samples	Test samples
Individual	320 samples	32 samples
Group	1600 samples	32 samples

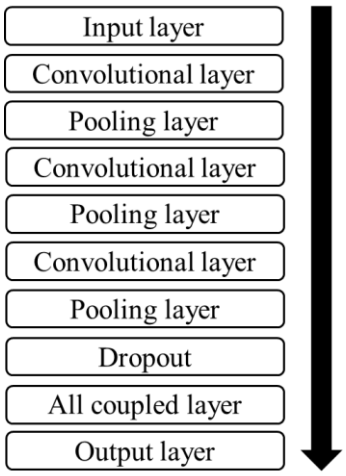

Figure 4. Architecture of the 3DCNN used in this study.

An advantage of the 3DCNN is that it can preserve the spatial structure of three-dimensional brain activation images during feature extraction. This characteristic was considered particularly relevant in the present study because the activation patterns were derived from localized regions in the auditory cortex. In contrast, the SVM treats the input as one-dimensional feature vectors and therefore does not directly use spatial structure in the same manner.

The SVM is a machine learning method that learns a decision boundary from the training data and then uses that boundary for classification. The decision boundary is determined by maximizing the margin with respect to the support vectors. In this study, the SVC (support vector classification) function [19] from the scikit-learn library was used to implement the SVM. Table V lists the hyperparameter combinations examined for the SVM. As in the 3DCNN experiments, multiple parameter settings were tested in order to compare model behavior under the present experimental conditions.

As described above, in the individual analysis, training and testing were performed separately for each participant using participant-specific annotated data. This setting was intended to evaluate whether the subtle 1 Hz frequency difference could be discriminated within each participant on the basis of that participant's own activation patterns. In the group analysis, one participant was designated as the test participant, while the remaining participants were used to construct the training model. The trained model was then evaluated using the data from the single test participant. This setting was used to examine how well the learned discrimination model generalized when applied to data from a participant who was not included in the training process.

TABLE IV. TRAINING PARAMETERS FOR THE 3DCNN.

Layer			Set value
Input Layer			41×50×15×1
Condition parameter	Number of Layer		6
	Convolutional layer	Stride	1
	Pooling layer	Filter size	2×2×2
		Stride	2
	Output layer	Class	2
	Learning rate		0.01
	Error rate		0.1
	Dropout		0.2
	Termination condition		Error rate
Hyperparameter	Convolutional layer	Kernel size (Ks)	3, 4, 5, 6, 7
		Filters (F)	16, 32
	Batch size (Bs)		8, 16, 32

TABLE V. HYPERPARAMETERS USED FOR THE SVM.

Kernel (K)	linear	rbf
C	0.01, 0.1, 1, 10, 100	0.01, 0.1, 1, 10, 100
Gamma (G)	—	0.01, 0.1, 1, 10, 100

III. RESULTS

Only models that completed training successfully were evaluated. For each trained model, the test data were input, and the model produced output probabilities for the two class labels (523 Hz or 524 Hz). The label with the higher probability was assigned as the predicted label and compared with the ground-truth label of the test data. This procedure was repeated for all 32 test samples, and the discrimination accuracy was defined as the number of correct predictions divided by 32.

Tables VI-IX present the discrimination accuracies and the corresponding hyperparameter combinations. Table VI shows the results for the 3DCNN method in individual analysis, Table VII shows those for the SVM method in individual analysis, Table VIII shows those for the 3DCNN method in group analysis, and Table IX shows those for the SVM method in group analysis. In all tables, the first column indicates the Participant ID used for training, and the second column indicates the Participant ID used for testing. The third column reports the highest discrimination accuracy observed in the grid search, and the fourth column lists the hyperparameter combinations that achieved this result. In some cases, the same discrimination accuracy was obtained with multiple hyperparameter combinations. For example, in Table VI for Participant A (Test A), a discrimination accuracy

of 100% was obtained with two combinations: Ks = 4, F = 32, Bs = 8, and Ks = 7, F = 32, Bs = 8.

Participant A in Table VI corresponds to the participant reported in the previous study [1], whereas Participants B-F represent the additional participants examined in the present study. In the individual-analysis setting, the 3DCNN generally showed high discrimination accuracy across participants, and high values were observed not only for Participant A but also for the additional participants. The SVM also achieved relatively high accuracies in several cases, but the overall tendency was less favorable than that of the 3DCNN. These results suggest that, under the present participant-specific setting, the 3DCNN was better able to capture the activation patterns associated with the subtle auditory difference. At the same time, the participant-wise results indicate that the favorable performance of the 3DCNN was not restricted to the participant examined in the previous study, but was also observed in the newly added participants under the same analytical framework.

A closer comparison of the individual-analysis results also suggests that the favorable performance of the 3DCNN was not limited to a single exceptionally good case. Rather, high accuracies were repeatedly observed across multiple participants, whereas the SVM showed a wider spread of values depending on the participant. This tendency is consistent with the summary statistics in Table X, in which the 3DCNN showed both a higher average accuracy and a smaller standard deviation in the individual-analysis setting. Thus, the participant-wise results support the view that preserving the spatial structure of activation images was advantageous under the present experimental conditions.

In the group analyses shown in Tables VIII and IX, the first column lists the five Participant IDs used for training. In contrast to the individual-analysis results, the group-analysis results showed substantial variation depending on the participant combination. Some train-test combinations produced discrimination accuracies above the chance level, whereas others yielded much lower values. This tendency suggests that the difficulty of discrimination increased when inter-participant variability was introduced into the training and test data. In addition, the variation observed in the group analyses was larger than that in the individual analyses, indicating that cross-participant discrimination was more sensitive to differences among participants. Thus, although certain participant combinations yielded relatively favorable results, the group-analysis setting as a whole showed less stable behavior than the individual-analysis setting.

The reported accuracies correspond to the highest observed results among the tested hyperparameter combinations under the present protocol.

Tables VI-IX each contain six train-test combinations. The average and standard deviation of the six discrimination accuracies in each table are summarized in Table X. Because the observed discrimination accuracies varied depending on both the hyperparameter settings and the train-test combinations, these average and standard deviation values should be interpreted as useful summary measures of the overall behavior of each method under the present experimental setting. In this study, these values are referred to as the performance and stability of each method. A higher average generally indicates better overall discrimination performance, whereas a lower standard deviation indicates smaller variation across train-test combinations and therefore greater overall stability.

According to Table X, the 3DCNN showed better overall performance than the SVM in the individual-analysis setting, as indicated by its higher average discrimination accuracy. In addition, the standard deviation for the 3DCNN in the individual analysis was relatively small, suggesting that its performance was comparatively stable across participants under this setting. In the group-analysis setting, however, both methods showed lower overall performance and greater variability than in the individual analysis. These results are consistent with the tendency observed in Tables VIII and IX and indicate that generalization across participants remained more difficult than discrimination within individual participants. Taken together, the results summarized in Table X support the view that the 3DCNN was particularly advantageous in the individual-analysis setting, whereas neither method showed uniformly stable performance in the group-analysis setting.

IV. DISCUSSION

In this study, the discrimination task was formulated as a binary classification problem, for which the nominal chance level is 50%. In the neuroscience literature, classification performance above chance level is often regarded as a meaningful indication that stimulus-related information is represented in brain activity, and accuracy levels around 60% have also been discussed as practically informative in some decoding contexts [20][21]. From this perspective, the observed results, especially those obtained in the individual analysis, are of clear interest.

First, regarding the reproducibility of the 1 Hz difference discrimination method across multiple participants, the individual-analysis results showed that high classification accuracy was observed not only for Participant A, reported in the previous study [1], but also for Participants B-F. As shown in Table VI, the 3DCNN method in individual analysis achieved 100 ± 0.0% for Participants A, B, C, E, and F, and 95.6 ± 3.2% for Participant D. This tendency supports the view that the previously reported result was not limited to a single participant and that similar outcomes can be obtained across multiple participants under the same experimental framework. In this sense, the present findings strengthen the participant-wise reproducibility of the proposed method. They also indicate that the favorable result observed in the earlier study was not an isolated case limited to one particularly well-performing participant.

At the same time, the present results should be interpreted within the scope of the current evaluation protocol. Because only 32 test samples were used per participant and permutation testing or label-randomization testing was not conducted, the observed accuracies should not be treated as a fully independent statistical demonstration by themselves. Rather, they should be interpreted as encouraging results under the present experimental setting, while stricter statistical validation remains an important subject for future work. In

particular, the present findings should be viewed as supporting the promise of the framework under the current protocol, rather than as definitively establishing statistical robustness.

Next, to examine the suitability of the 3DCNN for 1 Hz difference discrimination, its performance was compared with that of the SVM in both individual and group analyses. As summarized in Table X, the 3DCNN method in individual analysis achieved 99.3 ± 1.6%, whereas the SVM method in individual analysis achieved 85.9 ± 16.5%. In the individual analysis, the 3DCNN showed higher average accuracy and smaller variation than the SVM. This result suggests that preserving the spatial structure of brain activation images may be advantageous for participant-wise discrimination. Because the training and test data were derived from the same participant in the individual analysis, structural and activation features were relatively consistent, which likely facilitated feature extraction by the 3DCNN. In contrast, the SVM handled the input as one-dimensional vectors and therefore could not directly exploit spatial structure in the same manner. This interpretation is consistent with previous findings such as those reported by Vu et al. [22]. Vu et al. examined fMRI volume classification using three methods, including a 3DCNN and an SVM, across several task conditions, and reported that the 3DCNN outperformed the comparison methods because it was able to preserve and utilize the spatial organization of brain activation patterns. Their interpretation is therefore consistent with the present finding that the preservation of spatial structure is likely to be advantageous in participant-wise analysis.

In the group analysis, however, the performance of both methods was lower and more variable than in the individual analysis. As shown in Tables VIII and IX, both the 3DCNN and the SVM showed participant combinations with accuracies above 60%, whereas other combinations remained at or near the chance level. The summary values in Table X also show that average group-analysis performance was much lower than that in individual analysis. Accordingly, the group-analysis results should not be interpreted as demonstrating uniformly reliable discrimination across participants. Rather, they indicate that cross-participant discrimination is possible in some cases under the present framework, but that its stability remains limited. This contrast between the individual and group analyses is one of the most important findings of the present study, because it suggests that participant-specific consistency plays a substantial role in successful discrimination.

One plausible reason for this difficulty is the lack of uniformity in participant-specific brain structure and activation patterns. Because the 3DCNN preserves structural information during feature extraction, it may learn not only the sound-related features necessary for discrimination but also participant-specific differences that are not directly relevant to the task. In individual analysis, this property is advantageous because the model is trained and tested on the same participant. In group analysis, however, inter-participant differences may interfere with stable discrimination performance. As described in Section II-C, normalization to a standard brain was applied to reduce anatomical differences across participants; however, some degree of residual misalignment likely remained. As Kikuchi et al. [23] noted, complete alignment across participants is difficult to achieve, and such remaining differences may affect subsequent analysis. Such inter-participant variability may affect the spatial features extracted by the 3DCNN. In addition, differences in physiological response characteristics across participants may have influenced activation patterns, as discussed by Tavor et al. [24]. These factors may have reduced the consistency of the features relevant to the two auditory stimuli in the group-analysis setting. Thus, the lower stability in the group analysis may reflect not only limitations of the classifier itself but also the inherent difficulty of harmonizing participant-specific activation patterns within a single learning framework.

Another point requiring careful interpretation is the model-selection procedure. In this study, the reported accuracies correspond to the highest observed results among the tested hyperparameter combinations. This procedure was useful for exploratory comparison of model behavior under multiple settings; however, a stricter separation between hyperparameter selection and final evaluation, such as nested cross-validation or an independent held-out test set, would provide a more rigorous estimate of generalization performance. This remains an important direction for future work. Accordingly, the present results should be understood as reflecting the best observed behavior under the tested settings rather than as a definitive estimate of out-of-sample performance.

The issue of overfitting is also relevant to possible future improvements of the proposed framework. As many researchers have noted, common causes of overfitting include an excessive number of learnable parameters relative to the available data and insufficient training data [25][26]. In the present study, these conditions may have affected performance, particularly in group analysis. One possible countermeasure is stronger regularization. As Srivastava et al. [27] reported, Dropout is a representative technique for mitigating overfitting. Although Dropout was already used in the present study, it was fixed at 0.2, as shown in Table IV. Further examination of the dropout setting, together with additional training data and improved evaluation design, may help improve the stability of group-analysis performance. In addition, the stopping criterion for 3DCNN training in the present study was based on training error. Although this criterion enabled training to be completed consistently within the present framework, validation-based early stopping would offer a more rigorous approach to controlling overfitting. Together with stronger regularization and improved evaluation design, this may further improve the stability and interpretability of the results.

The participant cohort in the present study was also limited in size and diversity, consisting of six adult males in their twenties. Therefore, while the findings are meaningful as a methodological examination under controlled conditions, they should not yet be interpreted as establishing broad population-level robustness or practical applicability. Further validation using larger and more diverse cohorts will be necessary. Such expansion will also be important for determining whether the present framework remains effective under more

TABLE VI. DISCRIMINATION ACCURACY AND HYPERPARAMETERS FOR THE 3DCNN METHOD IN INDIVIDUAL ANALYSIS (THE SECOND ROW SHOWS THE RESULT REPORTED IN THE PREVIOUS STUDY [1]).

Participant ID (Training)	Participant ID (Testing)	Discrimination accuracy	Hyperparameters (Ks: kernel size; F: number of filters; Bs: batch size. See Table IV for details.)
A	A	100 ± 0.0%	Ks:4, 7, F:32, Bs:8
B	B	100 ± 0.0%	Ks:3, 4, 5, 6, 7, F:16, 32, Bs:8, 16, 32
C	C	100 ± 0.0%	Ks:3, 6, 7, F:32, Bs:8, 32 Ks:4, F:16, Bs:8 Ks:4, F:32, Bs:8, 16, 32 Ks:5, 6, 7, F:16, Bs:8, 16 Ks:5, F:32, Bs:16, 32
D	D	95.6 ± 3.2%	Ks:7, F:32, Bs:8
E	E	100 ± 0.0%	Ks:3, 4, 5, 6, 7, F:16, 32, Bs:8, 16, 32
F	F	100 ± 0.0%	Ks:3, 4, 5, 6, 7, F:16, 32, Bs:8, 16, 32

TABLE VII. DISCRIMINATION ACCURACY AND HYPERPARAMETERS FOR THE SVM METHOD IN INDIVIDUAL ANALYSIS.

Participant ID (Training)	Participant ID (Testing)	Discrimination accuracy	Hyperparameters (K: kernel; G: gamma. See Table V for details.)
A	A	84.4%	K:rbf, C:1, 10, 100, G:0.01
B	B	96.9%	K:rbf, C:1, 10, 100, G:0.01
C	C	90.6%	K:linear, C:0.1, 1, 10, 100
D	D	87.5%	K:linear, C:0.1, 1, 10, 100
E	E	50.0%	K:linear, C:0.01, 0.1, 1, 10, 100 K:rbf, C:1, 10, 100, G: 0.01, 0.1, 1, 10, 100 K:rbf, C:0.01, 0.1, G: 0.01, 1, 10, 100
F	F	100%	K:rbf, C:1, 10, 100, G:0.01

TABLE VIII. DISCRIMINATION ACCURACY AND HYPERPARAMETERS FOR THE 3DCNN METHOD IN GROUP ANALYSIS.

Participant ID (Training)	Participant ID (Testing)	Discrimination accuracy	Hyperparameters (Ks: kernel size; F: number of filters; Bs: batch size. See Table IV for details.)
BCDEF	A	50.0 ± 0.0%	Ks:4, 6, 7, F:16, Bs:8 Ks:5, 6, F:16, Bs:16 Ks:4, 5, F:32, Bs:32 Ks:4, 6, 7, F:16, Bs:8 Ks:6, F:16, Bs:32 Ks:6, F:32, Bs:8
ACDEF	B	85.9 ± 7.8%	Ks:5, F:16, Bs:16
ABDEF	C	57.8 ± 14.1%	Ks:7, F:32, Bs:32
ABCEF	D	50.0 ± 0.0%	Ks:3, 7, F:16, Bs:8, 16 Ks:3, 4, 5, 7, F:32, Bs:8 Ks:3, 5, 6, 7, F:32, Bs:16 Ks:4, 5, F:16, Bs:32 Ks:4, 6, 7, F:32, Bs:32 Ks:5, 6 F:16, Bs:8
ABCDF	E	60.9 ± 7.8%	Ks:7, F:16, Bs:32
ABCDE	F	50.0 ± 0.0%	Ks:3, F:16, Bs:8, 16 Ks:3, F:32, Bs:8 Ks:4, 5, 6, 7, F:16, 32, Bs:8, 16, 32

TABLE IX. DISCRIMINATION ACCURACY AND HYPERPARAMETERS FOR THE SVM METHOD IN GROUP ANALYSIS.

Participant ID (Training)	Participant ID (Testing)	Discrimination accuracy	Hyperparameters (K: kernel; G: gamma. See Table V for details.)
BCDEF	A	59.4%	K:linear, C: 0.1, 1, 10, 100
ACDEF	B	50.0%	K:rbf, C:0.01, 0.1, G: 0.01, 0.1, 1, 10, 100 K:rbf, C:1, 10, 100 G: 0.1, 1, 10, 100
ABDEF	C	59.4%	K:linear, C:1, 10, 100 K:rbf, C:10, 100, G:0.01
ABCEF	D	96.9%	K:rbf, C:1, 10, 100, G:0.01
ABCDF	E	50.0%	K:rbf, C:0.01, G:0.1, 1, 10, 100 K:rbf, C:0.1, 1, 10, 100, G:0.01, 0.1, 1, 10, 100
ABCDE	F	50.0%	K:rbf, C:0.01, 0.1, G:0.1, 1, 10, 100 K:rbf, C:1, 10, 100, G:1, 10, 100

TABLE X. AVERAGE DISCRIMINATION ACCURACY AND STANDARD DEVIATION FOR EACH METHOD, CALCULATED FROM THE SIX DATASET COMBINATIONS SHOWN IN TABLES VI–IX.

Method	Performance and stability (average ± standard deviation)
3DCNN Method in Individual Analysis （TABLE VI）	99.3 ± 1.6%
SVM Method in Individual Analysis （TABLE VII）	85.9 ± 16.5%
3DCNN Method in Group Analysis （TABLE VIII）	59.1 ± 12.7%
SVM Method in Group Analysis （TABLE IX）	60.6 ± 16.6%

heterogeneous real-world conditions.

The present findings are also meaningful from the viewpoint of methodological development in auditory brain decoding. In particular, the contrast between the individual and group analyses indicates that the choice of analytical framework substantially affects the observed performance when subtle auditory differences are examined. This suggests that future studies should not only seek higher classification accuracy, but also carefully consider how training-test design, participant variability, and feature representation influence the interpretability of the decoding results. From this perspective, the present study provides a useful basis for refining experimental and analytical strategies in future auditory fMRI studies. It also highlights the importance of evaluating not only whether subtle auditory information can be decoded, but also under what conditions such decoding remains stable across different analytical settings.

Taken together, the present findings indicate that the proposed framework is particularly promising for individual analysis, where the 3DCNN achieved high and comparatively stable performance across participants under the current setting. In contrast, the group-analysis results suggest that cross-participant application remains more challenging and requires further methodological refinement. Thus, although the present study does not yet establish robust generalization across participants, it provides additional support for the possibility that subtle frequency differences, even at the 1 Hz level, are reflected in fMRI activation patterns in the auditory cortex and can be captured to some extent by machine learning methods.

V. CONCLUSIONS AND FUTURE WORK

The aim of this study was to examine a machine-learning-based method for discriminating between two auditory stimuli differing by 1 Hz from fMRI activation patterns in the auditory cortex. To further evaluate the method beyond our earlier study, we investigated its reproducibility across multiple participants and compared the performance of a 3DCNN with that of a SVM.

In the individual analysis, high classification accuracy was observed across all participants, and the 3DCNN showed higher performance and smaller variation than the SVM under the present experimental setting. These results support the usefulness of the 3DCNN-based framework for participant-wise discrimination of very small frequency differences from fMRI activation patterns. More specifically, the present findings suggest that subtle frequency differences can be reflected in auditory-cortex activation patterns and can be captured under the present experimental setting when training and testing are performed within the same participant. The reproducibility observed across Participants A-F also suggests that the present framework has value as a participant-specific decoding approach under controlled conditions.

In the group analysis, both the 3DCNN and the SVM showed more variable performance across participants. While some train-test combinations achieved accuracies around or above 60%, several cases remained at the chance level. These results suggest that cross-participant discrimination is more difficult than participant-wise discrimination under the current framework and that further methodological improvement is needed before stronger conclusions can be

drawn regarding broader applicability. In particular, the present results indicate that stable generalization across participants remains a more demanding problem than within-participant discrimination of subtle auditory differences.

The present study also has several methodological limitations that should be recognized. The test size per participant was limited, permutation testing was not conducted, hyperparameter selection and final evaluation were not fully separated, and the participant cohort was small and homogeneous. Accordingly, the present findings should be understood as evidence obtained under restricted experimental conditions, while still providing meaningful support for the feasibility of fine-grained auditory discrimination from fMRI activation patterns. These limitations do not negate the interest of the present findings, but they do indicate that additional validation will be necessary before broader claims can be justified.

Future work will include more rigorous statistical validation, improved model-selection and stopping procedures, examination of regularization settings such as Dropout, and evaluation using larger and more diverse participant cohorts. In addition, extending the framework to other frequency ranges, environmental sounds, and musical stimuli may help clarify the broader scope of auditory decoding from human brain activity. Further investigation along these lines may also contribute to a better understanding of how subtle acoustic differences are represented in human brain activation patterns. In this respect, the present findings may serve as a useful reference point for future methodological refinement in auditory fMRI decoding studies.

Overall, the present study is meaningful not only because it examined discrimination of an extremely small auditory difference of 1 Hz, but also because it extended the previous work by evaluating reproducibility across multiple participants and by comparing two different classification frameworks. Although the current results should be interpreted within methodological limitations, they provide a useful basis for future studies aimed at clarifying how fine-grained auditory information is represented in human brain activity and how such information can be decoded more reliably from fMRI activation patterns. Such a line of research may also contribute to future efforts to establish more reliable frameworks for decoding fine-grained auditory information from human brain activity.

ACKNOWLEDGMENTS

The authors would like to express their sincere gratitude to all members of the laboratory for their valuable advice and support throughout this study. The authors also thank all participants for their kind cooperation in the experiments.

REFERENCES

[1] Y. Ooyashiki, and K. Shibata, “Discrimination by Deep Learning of 1Hz Difference in Auditory Cortex Using fMRI Activation Patterns ”, The Tenth International Conference on Informatics and Assistive Technologies for Health-Care, Medical Support and Wellbeing (HEALTHINFO 2025), ISBN: 978-1-68558-312-5, pp. 1-4, 2025.

[2] Y. Liang, K. Bo, S. Meyyappan, and M. Ding, “Decoding fMRI data with support vector machines and deep neural networks”, Journal of Neuroscience Methods, vol. 401, pp. 110004, 2024.

[3] T. Horikawa and Y. Kamitani, “Hierarchical Neural Representation of Dreamed Objects Revealed by Brain Decoding with Deep Neural Network Features”, Frontiers in Computational Neuroscience, vol. 11, pp. 00004, 2017.

[4] M. Guilhem, M. D. L. Giovanni, and A. S. Shihab, “The Music of Silence: Part I: Responses to Musical Imagery Encode Melodic Expectations and Acoustics”, Journal of Neuroscience, vol. 41 (35), pp. 7435-7448, 2021.

[5] I. Daly, “Neural decoding of music from the EEG”, Nature, Scientific Reports. 13, pp. 624, 2023.

[6] M. A. Casey, “Music of the 7Ts: Predicting and Decoding Multivoxel fMRI Responses with Acoustic, Schematic, and Categorical Music Features”, Front, Psychol, vol. 8:1179, 2017.

[7] N. Shigemoto, H. Satoh, K. Shibata, and Y. Inoue, “Study of Deep Learning for Sound Scale Decoding Technology from Human Brain Auditory Cortex”, IEEE, Global Conference on Life Sciences and Technologies, pp. 212-213, 2019.

[8] D. R. M. Langer, W. H. Backes, and P. V. Dijk, ” Representation of lateralization and tonotopy in primary versus secondary human auditory cortex”, NeuroImage, vol. 34, pp. 264-273, 2007.

[9] W. Guo et al., “Robustness of Cortical Topography across Fields, Laminae, Anesthetic States, and Neurophysiological Signal Types”, The Journal of Neuroscience, vol. 32(27), pp. 9159–9172, 2012.

[10] V. A. Kalatsky, D. B. Polley, M. M. Merzenich, C. E. Schreiner, and M. P. Stryker, “Fine functional organization of auditory cortex revealed by Fourier optical imaging”, PNAS vol. 102 No. 37, pp. 13325-13330, 2005.

[11] S. Romero et al., “Cellular and Widefield Imaging of Sound Frequency Organization in Primary and Higher Order Fields of the Mouse Auditory Cortex”, Cerebral Cortex, vol. 30, pp. 1603-1622, 2020.

[12] Y. Ooyashiki, K. Shibata, and H. Satoh, “Identification of Small Frequency Differences in Primary Auditory Cortex of Human Brain by Deep Learning Using fMRI”, JSME Chugoku-Shikoku Student Association 54th Student Member Graduation Research Presentation Lecture, pp. 01c2, 2023, (in Japanese).

[13] Y. Ooyashiki, K. Shibata, and H. Satoh, “Discrimination fMRI Learning of 1Hz Difference in Primary Auditory Cortex of Human Brain using fMRI”, JSME Mechanical Engineering Congress, pp. J162p-21, 2024, (in Japanese).

[14] J. B. Issa et al., “Multiscale optical Ca2+ imaging of tonal organization in mouse auditory cortex”, Neuron, vol. 83, pp. 944-959, 2014

[15] G. C. Oliveira et al., “Robust deep learning for eye fundus images: Bridging real and synthetic data for enhancing generalization”, Biomedical Signal Processing and Control, vol. 94, pp. 106263, 2024.

[16] J. Wang, Z. Wang, and G. Liu, “Recording brain activity while listening to music using wearable EEG devices combined with Bidirectional Long Short-Term Memory Networks”, Alexandria Engineering Journal, vol. 109, pp. 1–10, 2024

[17] SIEMENS <https://www.siemens-healthineers.com/jp/about> 2026. 05.27

[18] OptoACTIVE <https://www.optoacoustics.com/medical/optoactive-inflatable> 2026.05.27

[19] scikit-learn SVC <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html> 2026.05.27

[20] T. A. Carlson, T. Grootswagers, and K. A. Robinson, “An introduction to time-resolved decoding analysis for M/EEG”, arXiv, qbio.NC, 1905.04820, 2019.

[21] A. K. Robinson, G. L. Quek, and T. A. Carlson, “Visual Representations: Insights from Neural Decoding”, Annual Review of Vision Science, vol. 9, pp. 313-335, 2023.

[22] H. Vu, H.-C. Kim, M. Jung, and J.-H. Lee, “fMRI volume classification using a 3D convolutional neural network robust to shifted and scaled neuronal activations”, NeuroImage, vol. 223, pp. 117328, 2020.

[23] Y. Kikuchi, and Y. Shiraishi, Practical Analysis of Brain Activity and Functional Connectivity in fMRI Data: Mastering SPM, SnPM, and CONN, Ishiyaku Publishers, 2019, (in Japanese).

[24] I. Tavor et al. ,"Task-free MRI predicts individual differences in brain activity during task performance", Science, vol.352, pp. 216-220, 2016.

[25] D. M. Hawkins, “The Problem of Overfitting”, Journal of Chemical Information and Computer Sciences, vol. 44, no. 1, pp. 1–12, 2004.

[26] K. Saito, Deep Learning from Scratch: Theory and Implementation with Python, O’Reilly Japan, pp.190-197, 2016.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting”, Journal of Machine Learning Research, vol. 15, no. 1, pp. 1929–1958, 2014.