Detection of Note Onsets From EEG While Listening to Music

Yuiko Kumagai and Toshihisa Tanaka
Tokyo University of Agriculture and Technology, Tokyo, Japan
kumagai15@sip.tuat.ac.jp, tanakat@cc.tuat.ac.jp

***Abstract*—This paper proposes an approach to predicting the onsets of notes in music from electroencephalogram (EEG) signals. Participants listened to 45 kinds of single-tone melodies produced by piano sounds set on the same tempo. Training labels (onset or not-onset) were given by 100 ms using the scores of the melodies. An EEG while listening music was divided into segments with a window width of 500 ms and an overlap of 100 ms. Then, we solve the classification problems using logistic regression (LR) or support vector machine (SVM). We report that five out of fourteen participants' areas under the curve (AUC) indicated more than 0.7. Furthermore, when the predicted onset sequence was used to predict the musical stimulus being listened to, the maximum classification accuracy was 91.7%. These results suggest that each note can be decoded from brain response. The proposed approach can measure brain responses to each note or adapted for brain-computer interface (BCI) using natural music.**

I. Introduction

The music consists of complex structures such as rhythm, tempo, timbre, and so on. When people are listening to music, they utilize cognitive processes such as perception, multi-modal integration, learning, memory, syntactic processing, and processing of meaning information [1]. Many researchers studied event-related potentials (ERP) in numerous contexts in the music and speech domains to understand the auditory mechanisms. However, ERPs, usually obtained from averaging over trials, are suitable for isolated stimuli due to the low signal-to-noise ratio of EEG. Therefore, decoding technologies were needed to study brain response to continuous stimuli such as speech, music, or environmental sounds [2].

Decoding methods on EEG signals are pretty standard in the domain of BCI research. For instance, Schaefer et al. reported that seven types of 3-s music-fragment could be classified from EEG [3]. O'Sullivan et al. proposed a speech-reconstruction model for decoding attentional selection in a multi-speaker environment [4].

In addition to this, decoding techniques might enable us to solve the mechanism of music perception, apply music therapy, and train music learners. Thus, researchers are focusing on the relationship between EEG and musical components. Cong et al. have reported that there are some links between musical components and EEG [5]. Recently, Vinay et al. [6] have predicted the onset sequence from an 1-s EEG using a deep neural network (DNN). To our knowledge, however, there is no research to decode each music component one by one from EEG.

In this paper, we proposed a novel approach to decoding each onset of the note from EEG. The EEG signal for each trial was divided into 500-ms windows as features and classified by LR or SVM. We set the tempo of music stimuli to 2.5 Hz, and the lowest note was the sixteenth note. Hence, we annotated the training labels by each sixteenth note using digital scores of the melodies.

The remaining part of the paper proceeds as follows: In Section II, we review past research for the brain response while listening to music. Section III describes the dataset we used, the preprocessing of the data, and the classification model. In Section IV, we present the results and discussion. In Section V, conclusions are drawn.

II. Related Work

The brain response while listening to music has been studied using ERP for a long time. For rhythm perception, it has been known that P1, N1, and P2 components were relevant [7]. As for music syntax, previous studies have used deviant speech sounds [8], rhythmic sequences [9], and melodies [10] to elicit the mismatch negativity.

ERPs have also been used for BCI. For example, Treder et al. proposed a novel simulation approach for BCI using polyphonic music [11]. In their analysis, P300 was measured, and the attended instrument can be classified offline with a mean accuracy of 91%.

As for BCI research, the auditory decoding method has also been studied. In the past decade, several studies have sought to determine the relationship between components of music and EEG to utilize BCI. For instance, Schaefer et al. classified 3s-fragment music from seven kinds of music using EEG [3]. Cong et al. applied independent component analysis (ICA) to extract EEG components and reported that the ICA-related components correlate with rhythm components [5]. Nowadays, new approaches to decode auditory stimuli from EEG are studying. For instance, Cheveigné et al. proposed an approach based on Canonical Correlation Analysis that finds the optimal transform to apply to both the stimulus and response. [2].

Recently, researchers have focused on cortical entertainment because entrainment reflects the rhythmic structure of stimuli. Specifically, cortical entrainment to beat, meter, rhythm, and even the envelope of melody has been demonstrated [12], [13], [14], [15].

On the other hand, some researchers have focused on decoding onsets of music because estimating audio onsets must be

fundamental for decoding more complex musical components from EEG. Sturm et al. proposed a multivariate regression-based method to extract onset-related brain responses from EEG [16]. The brain response resembles canonical onset-related ERPs, such as the N1-P2 complex. This EEG projection was then utilized to determine the Cortico-Acoustic Correlation (CACor), a synchronization measure between EEG signal and stimulus. Their results showed significant CACor in an individual listener's EEG of a single presentation of a full-length complex naturalistic music stimulus. Vinay et al. conducted a feasibility study for decoding onset in music from EEG using an open dataset called "NMED-T" [17]. The dataset does not provide onset data for the music stimuli as training data; thus, they annotated the onset using MADMOM library [18]. Then, the recurrent neural network model was trained by the 1-s window of EEG to decode onset sequences. As a result, F1-socre of their model achieved 0.54 assessed with mir_eval toolkit's onset evaluation method [19].

The limitation of previous studies was that onset labels were annotated, and the prediction was based on the onset of a specific interval, not the prediction of the brain's response to each sound. A new approach is therefore needed for future research to predict onsets one by one from EEG.

III. Materials and Methods

A. Dataset

The EEG data used in this study were collected in our previous work [20]. We used a subset of the data (control task conditions) to classify the note's onset. A brief description of the dataset follows.

1) Participants: Fifteen males (mean age 23.1 ± 1.11; range 21–25 years old) who had no professional music education participated in this experiment. All participants were healthy; none reported any history of hearing impairment or neurological disorders. They signed written informed consent forms, and the study was approved by the Human Research Ethics Committee of the Tokyo University of Agriculture and Technology.

2) EEG: EEGs were measured using an EEG gel head cap with 64 scalp electrodes (Twente Medical Systems International (TMSi), Oldenzaal, The Netherlands) following the international 10–10 placement system. For patient grounding, a wetted TMSi wristband was used. To measure eye movement, we recorded an electrooculogram (EOG) with two bipolar electrodes at the corner of the right eye (referenced to the right ear) and above the right eye (referenced to the left ear). All channels were amplified using a Refa 72-channel amplifier (TMSi) against the average of all connected inputs. The signals were sampled at a sampling rate of 2,048 Hz, and they were recorded with TMSi Polybench. At the same time, we recorded the audio signals to validate the onset timing of the presented musical stimuli.

3) Musical stimuli: We used musical stimuli (MIDI) synthesized by Sibelius (Avid Technology, USA), a music computation and notation program. We created forty-five pieces that consisted of melodies produced by piano sounds without harmony. The sound intensities of all of the generated musical pieces were identical. The length of each musical piece was 34 s, with the tempo set to 150 beats per minute (bpm) (i.e., the frequency of a quarter of a note was 2.5 Hz). The sampling frequency was set to 44,100 Hz.

4) Tasks: We used only data from the control task including thirty trials. In each task, a 34-s-long musical stimulus was presented 1 s after the onset of the trial. Participants were instructed to focus on the musical stimulus while fixating on the screen. The order of the stimuli was random across the subjects.

B. Data Representation

We hypothesized that the brain responds to each note onset. Based on previous studies, musical tones elicit ERP such as N1-P2 complex and P300. Therefore, we classified "onset" or "not-onset" from the 500-ms window of EEG.

1) Onset Labels: In this paper, we used the 30 s musical pieces with the frequency of a quarter of a note was 2.5 Hz. The minimum length of note as a sixteen note was 100 ms. Therefore, we gave labels (onset/not-onset) to EEG segments every 100 ms, so that each stimulus has 296 labels. We excluded five musical stimuli since they included triplet, which cannot be divided. The total number of musical stimuli was forty pieces, as shown in Table I. We labeled "onset" if the 100 ms-segment includes a note onset, and "not-onset" if the 100 ms-segment includes a sustained note or rest. A visual representation of the music label generation is shown in Fig. 1, and the number of onset labels was shown in the third columns of the Table I. Music label generation was implemented through a custom-written Python program that operated on an XML file generated in Sibelius.

2) Preprocessing: We classified note's onset from the 500-ms segment of EEG. One participant (s5ka) was excluded from the analysis because, due to technical difficulties, the audio signals related to this participant could not be recorded. First, a zero-phase second-order infinite impulse response notch digital filter (50 Hz) and a zero-phase fifth-order Butterworth digital highpass filter (1 Hz) were applied to the recorded EEG. Second, the trials contaminated with a large number of artifacts were removed by visual inspection. Third, to remove artifacts caused by EOG, we applied a blind source separation algorithm called second-order blind identification to the recorded EEGs [21], [22], [23]. We then re-referenced the filtered EEGs from the average reference to the average of ear references (M1 and M2) and extracted seven channels (FC5, FC1, FC2, FC6, C3, Cz, C4) for classification. Moreover, the re-referenced EEGs were resampled to 400 Hz, and a zero-phase 25th-order Butterworth digital lowpass filter (40 Hz) was applied. We utilized the 30-s EEG data, from 4 s to 34 s after onset, the same as our previous study [20]. The filtered EEGs were divided into segments with a width of 500 ms and an overlap of 100 ms. Finally, each segment was baseline corrected (using the average EEG from -100 ms to 0 ms) and was downsampled to 80 Hz. The number of trials across participants and onset rates is shown in Table II. The last

TABLE I
MUSIC FOR THE AUDIO STIMULI. FORTY SEGMENTS WERE EXTRACTED BASED ON THE MUSIC MENTIONED IN THIS TABLE. THE THIRD COLUMN SHOWS THE RATE OF ONSET.

Composer	Title	Onset Rate [%]
L. v. Beethoven	Symphony "Ode to Joy"	24.7
G. Bizet	Carmen "Toreador Song"	26.7
J. Brahms	Hungarian Dances No.5	40.9
S. E. W. Elgar	Pomp and Circumstance Marches	17.3
E. H. Grieg	In the Hall of the Mountain King	39.5
G. F. Handel	Messiah "Hallelujah"	32.7
G. Holst	Planets "Mercury"	31.6
W. A. Mozart	Eine Kline Nachtmusik	33.2
W. A. Mozart	Piano Sonata No.11–3 "Turkish March"	65.5
H. Necke	Csikos Post	33.8
J. Offenbach	Orpheus in the Underworld	36.9
J. Pachelbel	Canon	38.1
S. S. Prokofiev	Romeo and Juliet "Montagues and Capulets"	41.8
G. A. Rossini	William Tell Overture	58.2
E. A. L. Satie	Gymnopedie No.1	15.0
E. A. L. Satie	Je te veux	14.4
J. Strauss	Voices of Spring Waltz	23.0
P. I. Tchaikovsky	Swan Lake "Scene"	29.8
I. Albeniz	Piano Sonata Op.82	34.7
L. v. Beethoven	Piano Sonata Op.14–1	23.3
A. Diabelli	Sonatina Op.151–2	38.6
A. Dvorak	Waltz	32.2
A. Dvorak	Serenade for Strings Op. 22–3 "Scherzo"	70.3
A. Dvorak	Serenade for Strings Op. 22–5 "Finale"	37.1
G. U. Faure	Dolly Suite Op. 56 "Kitty-valse"	31.0
E. H. Grieg	Lyric Pieces Op.47–6 "Spring dance"	47.1
F. J. Haydn	Piano Sonata No.12	44.9
F. J. Haydn	Piano Sonata No.28	39.3
F. J. Haydn	Piano Sonata No.33	52.3
F. Kuhlau	Sonatina Op.55–1	40.1
T. Leschetizky	Humoresque	38.4
J. L. F. Mendelssohn Bartholdy	Songs Without Words Op.19–1	22.9
W. A. Mozart	Piano Sonata KV309	31.8
S. S. Prokofiev	10 Pieces Op.12–2 Gavotte	33.8
S. S. Prokofiev	10 Pieces Op.12–3 Rigaudon	33.2
F. P. Schubert	Piano Sonata No.4 Scherzo	23.9
F. P. Schubert	Piano Sonata No.6–3	20.1
F. P. Schubert	String Quartet No.2	27.3
F. P. Schubert	String Quartet No.3	63.7
W. R. Wagner	Piano Sonata Op. 1	26.7

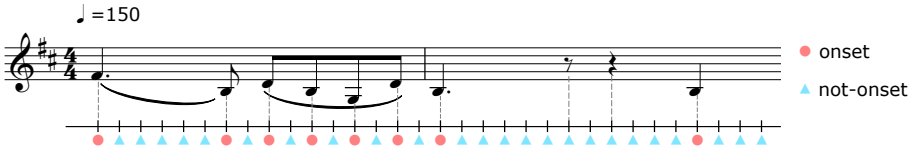

Fig. 1. Example of labels. Circles indicate "onset", and rectangles indicate "not-onset".

column shows the mean and standard deviation of the onset rate across participants.

C. Classification and Evaluation

Onset classification was achieved using three different methods, LR with L_1 regularization (LR–L1), LR with L_2 regularization (LR–L2), and linear SVM (LSVM). These three models were utilized as implemented in the open-source machine learning library scikit-learn [24]. The input feature of each model was the preprocessed EEG data with vectorization (80 Hz $\times$ 0.5 s $\times$ 7 channels $=$ 280 features per segment). Since our dataset was imbalanced, we set the class weight parameter of each model to "balanced", and selected the parameter C via a grid-search approach. Finally, we conducted five-fold cross-validation with the parameters mentioned above. The classification analysis was performed individually for each participant.

In this paper, two different evaluations: segment-level and music-level, were conducted. For the segment-level evaluation, F1-score and area under the curve (AUC) were calculated for each predicted result from a 500-ms segment. For music-level evaluation, we predicted the musical stimulus being listened to. Firstly, we predicted the onset sequence using 30 seconds of EEG while listening to an entire musical stimulus. Secondly, similarities between the predicted sequence and generated

TABLE II
THE NUMBER OF TRIALS AND ONSET RATE ACROSS PARTICIPANTS IS SHOWN. THE LAST COLUMN INDICATES THE MEAN AND STANDARD DEVIATION OF THE ONSET RATE ACROSS PARTICIPANTS.

Participant	Number of trials	Rate of onset
s1ka	22	34.2 ± 11.4
s2ka	24	35.7 ± 14.7
s3ka	24	34.2 ± 13.5
s4ka	26	34.9 ± 14.4
s6ka	25	35.5 ± 13.1
s7ka	22	36.0 ± 11.2
s8ka	26	30.9 ± 9.9
s9ka	26	38.8 ± 11.9
s10ka	27	37.2 ± 14.2
s11ka	23	36.1 ± 13.5
s12ka	27	34.8 ± 13.9
s13ka	24	36.6 ± 15.0
s14ka	26	35.3 ± 14.7
s15ka	24	33.5 ± 11.3

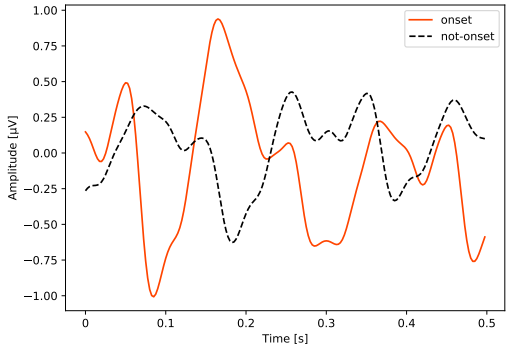

Fig. 2. Grand average ERP from channel Cz ($n = 14$) is shown. The red line indicates onset, and the black dashed line indicates not-onset.

sequence of each musical stimuli participants listened to in the control task were calculated with all combinations. The similarities were assessed with correlation coefficient (Pearson's r) or hamming distance. Then, the indexes of the musical stimuli were ordered by degree of similarity. Finally, top-1 and top-3 accuracy were calculated based on the degree of similarity.

To confirm the chance level of music-level evaluation, we conducted an additional experiment below. Firstly, as a prediction of each musical stimuli, onset sequences were calculated using a uniform distribution. Secondly, similarities were calculated between the predicted sequences and generated sequences of musical stimuli. Then, the indexes of the musical stimuli were ordered by degree of similarity, and top-1 and top-3 accuracy were extracted. Lastly, we repeated the process 10,000 times and averaged accuracies across the process.

IV. RESULTS AND DISCUSSION

The purpose of our experiment was to predict onset or not-onset from the 500-ms segment of EEG. Firstly, we compared ERP responses between onset and not-onset. Fig. 2 shows grand average ERP from channel Cz. Both N1 and P2 amplitudes of onset were larger than that of not-onset. This finding is consistent with that of Sturm et al., who reported N1-P2 complex elicited as a reaction to sound onset [16]. Secondly, we assessed the decoding model as segment-level (by 500-ms window). The summary of segment-level classification accuracy (F1-score and AUC) for each classification method are shown in Table III, which shows the mean, minimum, and

TABLE III
A SUMMARY OF THE RESULTS OF THE SEGMENT-LEVEL EVALUATION IS SHOWN. F1-SCORE AND AUC WERE CALCULATED AS THE CLASSIFICATION ACCURACY OF LR–L1, LR–L2, AND LSVM.

model	F1-score			AUC		
	mean	min	max	mean	min	max
LR–L1	0.546	0.496	0.594	0.673	0.607	0.729
LR–L2	0.546	0.499	0.591	0.673	0.612	0.729
LSVM	0.538	0.488	0.594	0.661	0.593	0.729

TABLE IV
A DETAIL OF THE RESULTS OF THE SEGMENT-LEVEL EVALUATION IS SHOWN. WE CALCULATED F1-SCORE AND AUC FOR CLASSIFICATION ACCURACY FOR EACH PARTICIPANT-SPECIFIC MODEL.

Participant	model	C	F1-score	AUC
s1ka	LR–L1	12.743	0.500	0.620
	LR–L2	12.743	0.500	0.620
	LSVM	112.884	0.493	0.612
s2ka	LR–L1	1.438	0.546	0.663
	LR–L2	0.695	0.544	0.663
	LSVM	0.004	0.488	0.594
s3ka	LR–L1	0.336	0.521	0.658
	LR–L2	2.976	0.534	0.665
	LSVM	6.158	0.530	0.665
s4ka	LR–L1	483.293	0.564	0.691
	LR–L2	26.367	0.565	0.697
	LSVM	26.367	0.565	0.696
s6ka	LR–L1	2.976	0.594	0.723
	LR–L2	0.695	0.591	0.723
	LSVM	2.976	0.594	0.724
s7ka	LR–L1	0.695	0.518	0.626
	LR–L2	0.078	0.521	0.627
	LSVM	0.038	0.519	0.624
s8ka	LR–L1	0.695	0.548	0.721
	LR–L2	0.695	0.548	0.721
	LSVM	0.336	0.550	0.722
s9ka	LR–L1	54.556	0.529	0.627
	LR–L2	112.884	0.531	0.627
	LSVM	112.884	0.531	0.627
s10ka	LR–L1	2.976	0.519	0.634
	LR–L2	26.367	0.514	0.626
	LSVM	1.438	0.521	0.636
s11ka	LR–L1	26.367	0.560	0.692
	LR–L2	233.572	0.554	0.685
	LSVM	0.002	0.510	0.593
s12ka	LR–L1	0.336	0.585	0.723
	LR–L2	1.438	0.583	0.725
	LSVM	6.158	0.586	0.726
s13ka	LR–L1	2.976	0.585	0.709
	LR–L2	6.158	0.585	0.709
	LSVM	2.976	0.585	0.709
s14ka	LR–L1	2.976	0.496	0.607
	LR–L2	0.162	0.499	0.612
	LSVM	0.018	0.490	0.604
s15ka	LR–L1	0.695	0.577	0.729
	LR–L2	0.162	0.576	0.729
	LSVM	0.336	0.577	0.729

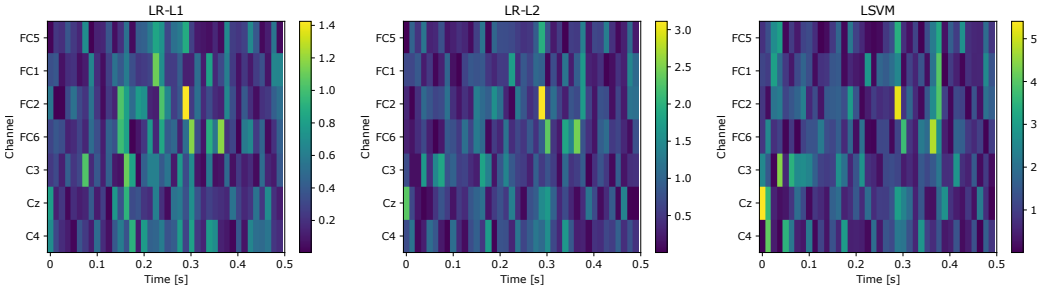

Fig. 3. The absolute values of the coefficients of LR–L1, LR–L2, and LSVM for s12ka are shown. The horizontal line indicates a time in seconds, and the vertical line indicates the EEG channel. In all models, the coefficients around 0.3 s were more significant than those at other times.

maximum values of accuracy across subjects. In addition to this, the classification accuracy by each participant was shown in Table IV. According to Tables III–IV, the averaged F1-socre was 0.54, and five out of fourteen participants' AUC indicated more than 0.7. Besides, there are no significant differences between the models. The averaged F1-scores were the same as the previous study [6], although the evaluation methods were different from ours.

In order to discuss which time of the EEG is essential for classification, we plotted the magnitude of the weight of the LR–L1, LR–L2, and LSVM averaged across folds. Fig. 3 shows the coefficients of a participant (s12ka). The horizontal axis indicates the channel index corresponding to Fig. 3, and the vertical axis indicates the time. As shown in Fig. 3, the weights of the three models were largest just before at 0.3 s after onset. These results suggest that the onset-related response occurs around 300 ms after the onset of the stimulus. According to the previous study [25], P200 components were correlated with the magnitude of the rapid change in the musical feature. Also, Haumann et al. demonstrated that average P2 responses could be extracted at sound onsets in real musical pieces [26]. Since P200 is evoked 200–300 ms after onset, the present results are consistent with previous studies. Moreover, these results match with the segment-level accuracies shown in Tables III–IV. Thus, considering our paradigm to predict onset one by one can be challenging, our proposed model seems appropriate.

Furthermore, we conducted a music-level evaluation to assess the trained models. Fig. 4 shows the results of music-level evaluation with LSVM. Although accuracy depended on participants, top-1 accuracy using correlation showed the best performance with an accuracy of 91.7%, and top1-accuracy using hamming distance showed 79.2%. As for the evaluation using correlation coefficients, the accuracy of both top-1 and top-3 were above chance level. The results were basically above the chance level in the evaluation using the Hamming distance, although some subjects were below the chance level. In addition, the maximum accuracies, or outliers, in the validation process are indicated by red crosses. Overall, nine subjects' accuracy was above the maximum in all cases, while the remaining five participants' accuracy was below the maximum in some cases As well as the segment-level evaluation, the music-level evaluations of the LR showed similar results.

One unanticipated finding was that the classification accuracies were low for some participants. In BCI studies, there is a phenomenon called BCI-illiteracy, which is the incompatibility between the user and the BCI occurring when a user cannot attain adequate control of a BCI [27], [28]. One of the possible causes of BCI-illiteracy is a low signal-to-noise ratio of EEG. It is known that although brain activity is occurring, it may not be observed as scalp EEG due to the shape of the cerebrum and scalp. According to Ahn et al., subjects in the illiterate group showed higher noise than subjects in the literate group [29]. The low performance in our results may be due to noise in the EEG. Also, Lotte et al. argued that it could not be decoded by signal processing and machine learning algorithms if the user cannot encode the command [30]. Therefore, it can be assumed that participants with a low accuracy may not recognize the music they listen to. One possible solution to improve the classification accuracies is learning end-to-end DNN. As several researchers have proposed DNN architectures, models that integrate feature extraction may precisely decode EEG [31], [32], [33]. Further studies are needed to better understand brain responses during listening to music.

V. Conclusion

We proposed an approach to predict onsets of notes in music using EEG. The baseline-corrected EEG was divided into segments of 500 ms each, and the models of LR–L1, LR–L2, and LSVM were trained. From Segment-level evaluation and the coefficients of the classifiers, we found no significant difference between the models. Most accuracies were above the chance level for music-level evaluation, and the best participant achieved a classification rate of 87.5%. These results suggest that the brain's response to each note of music can be decoded from the brain waves while listening to music. Further, the proposed method can be adapted for BCIs using natural music, solving music perception, music therapy, and training for music learners.

Acknowledgment

This work was supported by JSPS KAKENHI Grant No. 21K18311.

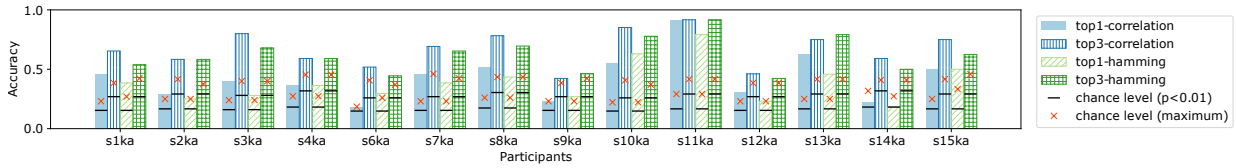

Fig. 4. Accuracy for music-level evaluation with LSVM of each participant using correlation or hamming distance was shown. The black dashed lines indicate the chance level ($p < 0.01$), and the red crosses indicate the maximum accuracies of the verification process.

REFERENCES

[1] S. Koelsch, "Toward a neural basis of music perception–a review and updated model," *The Relationship Between Music and Language*, p. 169, 2011.

[2] A. de Cheveigné, D. E. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018.

[3] R. S. Schaefer, J. Farquhar, Y. Blokland, M. Sadakata, and P. Desain, "Name that tune: Decoding music from the listening brain," *NeuroImage*, vol. 56, no. 2, pp. 843–849, 2011.

[4] J. A. O'Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, B. G. Shinn-Cunningham, M. Slaney, S. A. Shamma, and E. C. Lalor, "Attentional selection in a cocktail party environment can be decoded from single-trial EEG," *Cerebral Cortex*, vol. 25, no. 7, pp. 1697–1706, 2015.

[5] F. Cong, V. Alluri, A. K. Nandi, P. Toiviainen, R. Fa, B. Abu-Jamous, L. Gong, B. G. W. Craenen, H. Poikonen, M. Huotilainen, and T. Ristaniemi, "Linking brain responses to naturalistic music through analysis of ongoing EEG and stimulus features," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1060–1069, 2013.

[6] A. Vinay, A. Lerch, and G. Leslie, "Mind the beat: detecting audio onsets from eeg recordings of music listening," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 231–235.

[7] J. Stupacher, G. Wood, and M. Witte, "Neural Entrainment to Polyrhythms: A Comparison of Musicians and Non-musicians," *Frontiers in Neuroscience*, vol. 11, 2017.

[8] G. Dehaene-Lambertz, "Electrophysiological correlates of categorical phoneme perception in adults," *NeuroReport*, vol. 8, no. 4, pp. 919–924, 1997.

[9] C. Lappe, O. Steinsträter, and C. Pantev, "Rhythmic and melodic deviations in musical sequences recruit different cortical areas for mismatch detection." *Frontiers in Human Neuroscience*, vol. 7, no. June, p. 260, 2013.

[10] P. Virtala, M. Huotilainen, E. Partanen, and M. Tervaniemi, "Musicianship facilitates the processing of western music chords–an ERP and behavioral study," *Neuropsychologia*, vol. 61, no. 1, pp. 247–258, 2014.

[11] M. S. Treder, H. Purwins, D. Miklody, I. Sturm, and B. Blankertz, "Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification." *Journal of Neural Engineering*, vol. 11, no. 2, p. 026009, 2014.

[12] T. Fujioka, L. J. Trainor, E. W. Large, and B. Ross, "Internalized timing of isochronous sounds is represented in neuromagnetic beta oscillations," *The Journal of Neuroscience*, vol. 32, no. 5, pp. 1791–1802, 2012.

[13] S. Nozaradan, "Exploring how musical rhythm entrains brain activity with electroencephalogram frequency-tagging," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 369, no. 1658, p. 20130393, 2014.

[14] B. Meltzer, C. S. Reichenbach, C. Braiman, N. D. Schiff, A. Hudspeth, and T. Reichenbach, "The steady-state response of the cerebral cortex to the beat of music reflects both the comprehension of music and attention," *Frontiers in Human Neuroscience*, vol. 9, p. 436, 2015.

[15] Y. Kumagai, M. Arvaneh, and T. Tanaka, "Familiarity affects entrainment of EEG in music listening," *Frontiers in Human Neuroscience*, vol. 11, p. 384, 2017.

[16] I. Sturm, S. Dähne, B. Blankertz, and G. Curio, "Multi-variate EEG analysis as a novel tool to examine brain responses to naturalistic music stimuli," *PLoS ONE*, vol. 10, no. 10, pp. 1–30, 2015.

[17] S. Losorelli, D. T. Nguyen, J. P. Dmochowski, and B. Kaneshiro, "Nmed-t: A tempo-focused dataset of cortical and behavioral responses to naturalistic music." in *ISMIR*, 2017, pp. 339–346.

[18] S. Böck, F. Korzeniowski, J. Schlüter, F. Krebs, and G. Widmer, "Madmom: A new python audio and music signal processing library," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1174–1178.

[19] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, D. P. Ellis, and C. C. Raffel, "mir_eval: A transparent implementation of common mir metrics," in *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014.

[20] Y. Kumagai, R. Matsui, and T. Tanaka, "Music familiarity affects EEG entrainment when little attention is paid," *Frontiers in Human Neuroscience*, vol. 12, p. 444, 2018.

[21] A. Belouchrani, K. Abed-Meraim, J. Cardoso, and E. Moulines, "Second-order blind separation of temporally correlated sources," in *Proc. Int. Conf. Digital Signal Processing*. Citeseer, 1993, pp. 346–351.

[22] A. Belouchrani and A. Cichocki, "Robust whitening procedure in blind source separation context," *Electronics Letters*, vol. 36, no. 24, pp. 2050–2051, 2000.

[23] A. Cichocki and S.-I. Amari, *Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications*. John Wiley & Sons, 2002, vol. 1.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[25] H. Poikonen, V. Alluri, E. Brattico, O. Lartillot, M. Tervaniemi, and M. Huotilainen, "Event-related brain responses while listening to entire pieces of music," *Neuroscience*, vol. 312, pp. 58–73, 2016.

[26] N. T. Haumann, M. Lumaca, M. Kliuchko, J. L. Santacruz, P. Vuust, and E. Brattico, "Extracting human cortical responses to sound onsets and acoustic feature changes in real music, and their relation to event rate," *Brain Research*, vol. 1754, p. 147248, mar 2021.

[27] "Towards EEG-based BCI driven by emotions for addressing BCI-Illiteracy: a meta-analytic review," *Behaviour and Information Technology*, vol. 37, no. 8, pp. 855–871, 2018.

[28] M. H. Lee, O. Y. Kwon, Y. J. Kim, H. K. Kim, Y. E. Lee, J. Williamson, S. Fazli, and S. W. Lee, "EEG dataset and OpenBMI toolbox for three BCI paradigms: An investigation into BCI illiteracy," *GigaScience*, vol. 8, no. 5, pp. 1–16, 2019.

[29] M. Ahn, H. Cho, S. Ahn, and S. C. Jun, "High theta and low alpha powers may be indicative of BCI-illiteracy in motor imagery," *PLoS ONE*, vol. 8, no. 11, p. 80886, 2013.

[30] F. Lotte, C. Jeunet, J. Mladenović, B. N'Kaoua, and L. Pillette, *A BCI challenge for the signal processing community: considering the user in the loop*, ser. Signal Processing and Machine Learning for Brain-Machine Interfaces. IET, 2018.

[31] R. T. Schirrmeister, J. T. Springenberg, L. D. J. Fiederer, M. Glasstetter, K. Eggensperger, M. Tangermann, F. Hutter, W. Burgard, and T. Ball, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[32] V. J. Lawhern, A. J. Solon, N. R. Waytowich, S. M. Gordon, C. P. Hung, and B. J. Lance, "EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces," *Journal of Neural Engineering*, vol. 15, no. 5, p. 056013, 2018.

[33] W. Ko, E. Jeon, S. Jeong, and H. I. Suk, "Multi-Scale Neural Network for EEG Representation Learning in BCI," *IEEE Computational Intelligence Magazine*, vol. 16, no. 2, pp. 31–45, 2021.