Disk Jockey in Brain - A Prototype for Volume Control of Tracked Instrument during Playback

Ping-Keng Jao, Pei-I Chen, and Yi-Hsuan Yang*

Research Center for Information Technology Innovation, Academia Sinica, Taiwan
{nafraw, gooa1121, yang}@citi.sinica.edu.tw

Abstract. Brain-Computer Interface (BCI) has attracted much attention in the past years, and its application to music is just emerging. In this paper, we propose a novel Music Interaction System based on asynchronous BCI, which allows a user to make self-paced decisions on when to switch from one mental task to the next. Our specific goal is to provide an easy way to interact with music while a user immerses oneself in the music. We take volume control of instruments over sound streams as an example, and collect the electroencephalography (EEG) signals of a number of human subjects who were asked to pay their attentions to one of the two instruments being played in a piece of classical music. With these EEG recordings, we perform off-line analysis investigating the accuracy of a subject-dependent classifier in predicting the instrument that a subject is paying attention to, by using some features extracted from the EEG signals. Our result shows that for all the nine subjects the classifier can perform better than a random baseline (whose classification accuracy is 50%), and that for two of them the accuracy can reach 85%.

Keywords: Brain-computer interface, EEG, auditory, asynchronous BCI, music, human-computer interaction.

1 Introduction

Music evolves along with human history, and the advance of technology is one of the driving forces. For example, the invention of electronic device provides musicians with a higher degree of freedom in designing the timbre of music pieces. Nowadays, brain-computer interface (BCI) is a burgeoning technology, for it holds much potential in scientific research and practical application. As a result, some researchers have devoted themselves to the application of BCI to music. This can be done, for instance, by rendering the electroencephalography (EEG) signal as music [6], or by using BCI to send commands to play musical instruments [7]. In this work, we are interested in a BCI application that allows a listener to actively interact with live music, without compromising the joyful listening experience. In other words, the listener can simply immerse in the

* We would like to thank Steven Tsai, Bertram Liu, Chia-Hao Chung, and Yi-Wei Chen for their contribution in the pilot study of this work conducted and presented in a hackathon event in 2014 [1].

performance during the interactive listening experience, without the necessity of distracting herself or himself for generating commands for interaction.

The specific scenario considered in this work is that a listener would like to appreciate a specific *part* (in terms of pitch, not in time) in a polyphonic music. In many cases, a specific type of instrument represents the part. Hence, we would like to design an interactive system that captures the mind of the user, and then amplifies (attenuates) the attended (unattended) instrument. A kind of potential users would be a person currently learning a specific instrument, so the user may want to appreciate the instrument in more details from other performers in a polyphonic music. The users can also be merely music lovers who desire more control over, or interaction with, the music piece being played, or someone who would like to remix the audio as a disk jockey.

The goal of this paper is to assess the applicability of using consumer-grade EEG headset to realize such an application. The underlying assumption is that EEG signals carry sufficient information to identify the instrument the user is paying attention to. We would report an EEG data collection experiment where we asked a number of human subjects to wear an EEG headset while being asked to pay their attention to a particular instrument in a piece of two-instrument classical music. Based on this dataset, we extract a number of features from the EEG signals and evaluate the accuracy of a machine learning algorithm for recognizing the attended instrument.

This kind of paradigm should be categorized as *mental task*-based. Some mental tasks, for example, rotation, multiplication, counting, motor imaginary, have been found to trigger distinct cortex areas, and are therefore considered distinguishable [9]. An example of auditory stimuli is that discriminating the gender of the speaker, for which Xu *et al.* studied whether employing an active mental task improves the performance of the attention-related task [10]. Another popular paradigm is the *oddball* design such as the P300 speller [4], which is based on the presence of an infrequent event. Based on this paradigm, Treder *et al.* [8] recently published an article with a fairly similar goal to ours, i.e. using EEG to identify the musical instrument a human subject is attended. Our work differs from this prior art in the following aspects. First, they asked a human subject to count the deviant (i.e. the oddball) of the attended instrument. This is not a natural listening experience as the natural music piece in usual is not simply composed of both an infrequent deviant pattern and a standard pattern, not to say the deviants of different instruments do not appear simultaneously. Second, our work is based on the so-called *asynchronous BCI*, which allows a user to make self-paced decisions on when to shift his or her attention to another instrument, instead of *locking* the event (i.e. attention shifting) to the beginning of any epoch. Therefore, unlike [8] and [10], we do not assume any event-related potential (ERP) should be considered. Finally, although our subjects were actually not required to shift their attentions while listening to a music piece during the data collection experiment, we have indeed implemented such a system that allows a listener to attend to different instruments dynamically and accordingly changes the volume of the instruments through a mixer.

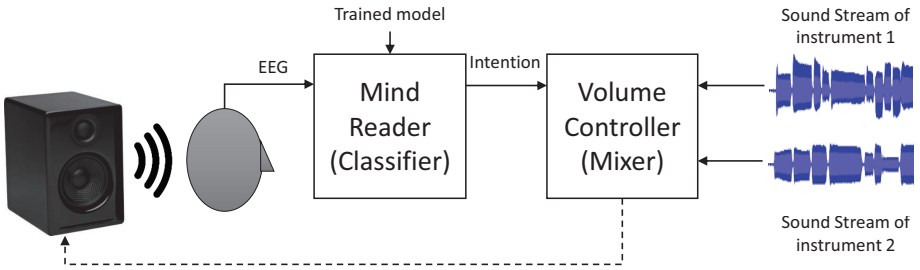

Fig. 1: The ideal online system. We performed off-line analysis in this work, which is equivalent to the case when the dashed arrow is removed.

2 The Proposed Interaction System

Fig. 1 illustrates the ideal (online) system. A user wearing an EEG headset is also listening to the music from the amplifier, and the EEG headset keeps transmitting EEG signals to a "mind reader," which is basically modeled as a classifier in this work. The classifier can be pre-trained in off-line, and (possibly) refined in the online stage by model adaptation techniques. The classifier would keep decoding the EEG signals to recognize the instrument the user is focusing. The output, then, is fed forward to a volume controller, whose goal is to smoothly adjust the volume of the sound streams from different instruments, so that the user will be satisfied with the quality of the remixed sound. The smoothing is important to avoid abrupt changes and to reduce the effect of potential classification errors in downgrading the remixing quality. In this work, we assume that we have clean sound streams, so the remixing of them becomes trivial. In practical situations where multi-track sources are not available, one would have to implement for example a source separation system (e.g. [2] for real time with high potential or [5] for better sound quality) to realize such an application.

2.1 Classifier

The "mind reader" is modeled as a multi-class classification problem solver. Toward making a time-efficient system, we simply employed a linear support vector machine (SVM) [3] for classification. Other than directly passing the EEG signals to the SVM, we applied some feature extraction and processing techniques. Specifically, for each time sample of an EEG signal, we output one classification result. For each time sample and spatial channel (the used 14 channels in the EEG headset will be described in Section 3), we use both the raw signal and signals in the following five frequency bands of interest in the time domain: δ (1-3 Hz), θ (4-7 Hz), α (8-12 Hz), β (13-30 Hz), and sleep spindle (12-14 Hz). Each band is filtered from the raw signal with a fifth order Butterworth bandpass filter according to their frequency range. Note we do not employ Fourier transform here for efficiency. Prior to feeding these features into a classifier, we further apply a sum-to-one normalization for each time sample, in the direction

of feature not in time so as to minimize system delay. Parameter tuning of SVM is done by inner-cross validation. Specifically, we use the l_2-regularized l_2-loss support vector classification (in primal form) implemented by the LIBLINEAR library, and optimize the parameter C with the search range, 2^{-10} to 2^{10}. We did not implement online model adaptation in this work.

2.2 Mixer

A simple volume controller is designed based on the output of the mind reader. Given N music streams, the waveform output at time t is $y(t) = \sum_{i=1}^{N} \alpha_i(t) x_i(t)$, where $x_i(t)$ is the waveform of stream i, $\alpha_i(t)$ is the weighting coefficient, and $\sum_{i=1}^{N} \alpha_i(t) = 1$. For each output of the classifier, we adjust a small quantity for each $\alpha_i(t)$. In this study, we have $N = 2$ (as described below), and we increase $\alpha_1(t)$ by 0.001 and decrease $\alpha_2(t)$ by 0.001 when the output favors stream 1, and vice versa. The value of $\alpha_i(t)$ is further bounded to $[0, 1]$.

3 Material and Experiment Protocol

To validate the idea, we designed an experiment to collect EEG signals for off-line analysis. The principal is to ask the subject to actively listen to (i.e. track) a specific instrument, rather than distracting herself or himself from the music.

Participants We recruited 9 subjects (2 females and 7 males) from our research institute, and their ages range from 24–39 (mean: 28.3, standard deviation: 4.72). All subjects are non-musicians and are right-handed. Before the experiment, each subject is informed how their data will be used and distributed.

Stimuli We selected four chorales and two instruments from the multi-track Bach10 dataset [2]. Originally, there are ten four-part chorales, where each instrument (including violin, clarinet, saxophone, and bassoon) represents a certain part. We selected the clarinet and saxophone from the four instruments, for simplifying the problem and for letting the non-musician participants to easily track the instrument. We did not consider violin for its volume is louder than the remaining three in Bach10, and we did not consider bassoon for its sound is dark and therefore harder to be tracked. Among the ten chorales, we selected '02-AchLiebenChristen,' '06-DieSonne,' '07-HerrGott,' and '10-NunBitten' for experiments. They were selected based on the volume and pitch difference between clarinet and saxophone. Roughly speaking, the four chorales can be categorized into two groups. The first group, chorale 2 and 10, has relatively close pitches for the instruments than the second group, chorales 6 and 7. In other words, the second group should be easier for subjects to track as the two instruments use more dissimilar pitches in these two chorales. The durations of the four chorales range from 32 to 40 seconds. All of them are sampled at 44,100 Hz.

We also picked a white noise from FreeSound[1] for neutralizing their memory of the listened music during the data collection experiments.

Equipment The used wireless EEG headset is the Emotiv EPOC model 1.0, which equipped with 14 EEG dry electrodes. The electrodes are placed at AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4, according to the international 10-20 system. The sampling rate is 128 Hz and the bandwidth is between 0.2 and 45 Hz (with a built-in digital notch filters at 50 and 60 Hz based on 5th order Sinc). The audio was played by a professional amplifier, Europort EPA300, to ensure high audio quality.

Arrangement The experiment was conducted in a bright, quiet and small room but without concrete sound-proofing. The subject wore the Emotiv headset and sat on a comfortable chair, which was approximately placed in the middle of the room. The left and right channels of the amplifier were evenly placed behind the subject, and the operator of the experiment sat at a place behind the subject and the amplifier. The operator controlled a laptop for collecting EEG signals and transmitting audio signals to the amplifier.

Experiment Flow The operator first played sample chorales (from the other six chorales) to let the subject balance the volumes of the left and right channels of the amplifier. The headset was moisturized before each subject wore it. Then, the subject wore the headset and tried the best to ensure all the 14 signal quality indicators of the Emotiv headset in "good." However, only three subjects completed the experiment with the indicators always in good. A few indicators for the other six subjects sometimes or mostly showed "poor," due to the difficulty of perfectly controlling the experiment. After the subject was ready, the operator played two sample audios, one for the clarinet and the other for the saxophone from the other six chorales, to let the subject have an idea about the specified instruments. The subject was instructed to close their eyes and to minimize body movements as much as possible during the recording of EEG signals. Then, the operator conducted the experiment with the following steps:

1. Shuffles randomly the order of the four chorales.
2. Plays the solo of the clarinet of the first chorale, to assist the subject to be more familiar with the melody of the target instrument.
3. Asks whether the subject is ready to actively listen to the clarinet in the next step. Goes to the next step once confirmed.
4. Plays the chorale again with both clarinet and saxophone, and records the EEG signal.
5. Repeats step 3-4 for the second time (i.e. *trial*).
6. Plays the white noise for 30 seconds for neutralization.
7. Repeats step 2-6 for the other three chorales.
8. Repeats step 1-7 with the target instrument being replaced by the saxophone.

[1] `https://www.freesound.org/people/theundecided/sounds/165058/`

Table 1: Averaged accuracy for evaluating the classifier.

	Chorale 10			Chorale 2			Chorale 6			Chorale 7			Mean		
	(clar.)		(saxo.)	(clar.)		(saxo.)	(clar.)		(saxo.)	(clar.)		(saxo.)	(clar.)		(saxo.)
Subject 1	(0.901)	0.709	(0.516)	(0.923)	0.950	(0.977)	(0.74)	0.604	(0.468)	(0.569)	0.632	(0.695)	(0.784)	0.724	(0.664)
Subject 2	(0.974)	0.957	(0.940)	(0.611)	0.631	(0.650)	(0.901)	0.900	(0.898)	(0.788)	0.825	(0.863)	(0.819)	0.828	(0.838)
Subject 3	(0.174)	0.458	(0.742)	(0.570)	0.563	(0.556)	(0.961)	0.688	(0.415)	(0.861)	0.587	(0.313)	(0.641)	0.574	(0.507)
Subject 4	(0.594)	0.679	(0.764)	(0.457)	0.649	(0.840)	(0.884)	0.837	(0.790)	(0.759)	0.698	(0.636)	(0.674)	0.716	(0.758)
Subject 5	(1.000)	0.970	(0.939)	(0.889)	0.934	(0.980)	(1.000)	0.857	(0.714)	(0.531)	0.742	(0.952)	(0.855)	0.876	(0.896)
Subject 6	(0.943)	0.956	(0.969)	(0.817)	0.874	(0.930)	(0.917)	0.803	(0.689)	(0.297)	0.571	(0.844)	(0.744)	0.801	(0.858)
Subject 7	(0.443)	0.600	(0.756)	(0.481)	0.651	(0.821)	(0.311)	0.575	(0.840)	(0.786)	0.690	(0.594)	(0.505)	0.629	(0.753)
Subject 8	(0.784)	0.802	(0.820)	(0.451)	0.560	(0.669)	(0.729)	0.704	(0.679)	(0.568)	0.611	(0.655)	(0.633)	0.669	(0.705)
Subject 9	(0.187)	0.593	(1.000)	(1.000)	0.979	(0.958)	(0.988)	0.878	(0.768)	(0.983)	0.989	(0.995)	(0.789)	0.860	(0.930)
Mean	(0.667)	0.747	(0.827)	(0.689)	0.754	(0.820)	(0.826)	0.761	(0.696)	(0.683)	0.705	(0.727)	(0.716)	0.742	(0.768)

4 Off-line System Verification

4.1 Validation Method

We employed leave-one-out as the cross validation method. Specifically, for each subject, there are in total 16 EEG trials (4 chorales $\times$ 2 trials $\times$ 2 instruments), and, for each chorale, we have 4 trials (1 chorale $\times$ 2 trials $\times$ 2 instruments) for testing, and the other 12 trials for training the SVM. The validation process repeats for each chorale and each subject.

We introduce two evaluation criteria to measure the system performance. The first one is aimed to evaluate the classifier. For an instrument of a chorale, we simply report the average classification accuracy over the entire chorale and the two trials. The second criterion is the of average 2-norm of the difference between the ideal and actual α_i (weighting coefficient of the sound stream), which indicates whether the sound streams are properly amplified (attenuated) by the mixer in the anticipated online system. It is calculated as $\frac{1}{NM}\sum_{t=1}^{M}||\hat{\alpha}(t)-\alpha(t)||_2^2$, where $\alpha(t)=[\alpha_1(t)\,\alpha_2(t)]^T$, and $\hat{\alpha}(t)$ is the ideal coefficient, which is equivalent to $[1\ 0]^T$ or $[0\ 1]^T$ for the data we collected, depending on the attended instrument. The value of this quantity falls between 0 and 1 in our case (N=2), and a smaller value indicates better performance.

4.2 Result and Analysis

Table 1 shows the average accuracy of each chorale and subject, for each of the two instruments and the average result (i.e. the middle one). The average value over all subjects and chorales is 0.742, and the accuracies are 0.716 and 0.768 respectively for the clarinet and saxophone. The difference between the two instruments is not too much, suggesting that the classifier indeed learns a model that helps recognizing the attended instrument. Although the average accuracy is 0.742, the large dynamic range between subjects deserves attention. For example, the average accuracy of subject 5 is as high as 0.876 while subject 3 is as low as 0.574. This implies that either the used features are not robust enough across subjects, or some technical difficulties occurred during the trials for some subjects. Actually, for subject 3, we indeed observed that most of the

Table 2: An objective evaluation for the re-mixed audio based on the difference between ideal and actual weighting coefficients α.

	Chorale 10			Chorale 2			Chorale 6			Chorale 7			Mean		
	(clar.)		(saxo.)	(clar.)		(saxo.)	(clar.)		(saxo.)	(clar.)		(saxo.)	(clar.)		(saxo.)
Subject 1	(0.047)	0.222	(0.397)	(0.033)	0.027	(0.020)	(0.238)	0.286	(0.333)	(0.241)	0.165	(0.088)	(0.14)	0.175	(0.210)
Subject 2	(0.014)	0.012	(0.010)	(0.141)	0.090	(0.039)	(0.027)	0.022	(0.018)	(0.037)	0.034	(0.031)	(0.055)	0.040	(0.025)
Subject 3	(0.857)	0.499	(0.141)	(0.327)	0.249	(0.171)	(0.013)	0.326	(0.639)	(0.04)	0.327	(0.615)	(0.309)	0.350	(0.391)
Subject 4	(0.179)	0.106	(0.032)	(0.526)	0.280	(0.034)	(0.068)	0.056	(0.044)	(0.043)	0.144	(0.245)	(0.204)	0.146	(0.089)
Subject 5	(0.009)	0.013	(0.017)	(0.019)	0.014	(0.008)	(0.010)	0.091	(0.173)	(0.207)	0.108	(0.01)	(0.061)	0.057	(0.052)
Subject 6	(0.036)	0.022	(0.009)	(0.066)	0.038	(0.010)	(0.033)	0.044	(0.056)	(0.575)	0.325	(0.075)	(0.177)	0.107	(0.037)
Subject 7	(0.441)	0.263	(0.086)	(0.585)	0.307	(0.029)	(0.785)	0.398	(0.011)	(0.113)	0.170	(0.228)	(0.481)	0.285	(0.088)
Subject 8	(0.023)	0.060	(0.097)	(0.343)	0.299	(0.255)	(0.094)	0.127	(0.160)	(0.242)	0.172	(0.103)	(0.176)	0.165	(0.154)
Subject 9	(0.845)	0.427	(0.009)	(0.008)	0.008	(0.008)	(0.011)	0.012	(0.014)	(0.011)	0.011	(0.010)	(0.219)	0.115	(0.010)
Mean	(0.272)	0.180	(0.089)	(0.228)	0.146	(0.064)	(0.142)	0.151	(0.161)	(0.168)	0.162	(0.156)	(0.203)	0.160	(0.118)

EEG signal quality indicators showed "poor" during most trials, and some other subjects have similar but not that severe situations, so, perhaps the low accuracy for some subjects can be attributed to this reason. Subject 6 also experienced a technical issue during a trial of the clarinet in chorale 7. The computer lost its connection to the headset, and this may explain why the accuracy is significantly lower than others. On the other hand, even subject 5 has a very high accuracy, the accuracy of clarinet in chorale 7 for this subject turns out to be much lower than the other chorales. Future work might be needed to take a closer look at the result of chorale 7 to understand the limitation of the current classifier.

Rather than the subject-oriented report of these accuracies, we also found a phenomenon in chorale-oriented examination. There is no instrument dominating the other over all chorales in terms of accuracy. For example, the accuracy of the clarinet is apparently better than the saxophone in chorale 6, but we see the opposite for the other three chorales. This may imply that pitch is also an important latent factor for linking EEG signal to the goal of the classifier, in addition to musical timbre.

Table 2 shows the result when we evaluate the performance using the second criterion. Generally, the performance trend is similar to that in Table 1. For example, 0.008 is the lowest error in Table 2 and the corresponding accuracies are all larger than 0.95. The largest error, 0.857, also maps to the lowest accuracy in Table 1. However, we need to emphasize that a smaller error (in Table 2) does not necessarily mean a higher accuracy (in Table 1). For example, the clarinet of chorale 6 for subject 4 and the chorale 7 for subject 3. This indicates that the user might have a better listening experience even when the average accuracy is lower, possibly because classification errors are distributed in a relatively scattering manner over time. In this case, the user will not feel the system totally goes into the wrong direction for a noticeable period.

5 Conclusion

We have proposed a simple yet innovative system potentially applicable for music interaction based on asynchronous BCI in mobile platforms. This is based on

the facts that only an 84-dimension (14 channels $\times$ 6 features) feature vector is used, and operations are based on IIR filters, a linear product (classifier), and some divisions for the normalization. However, a difficulty in implementing such a system for practical usage would be the unavailability of multi-track sources of music which may lead to high computational cost in performing source separation. By sidestepping this source separation issue, we built a small dataset using the Bach10 dataset to verify our idea, and the result seems to be promising. In the future, we plan to implement the system in a mobile platform for practical scenario, and to study the relevance between the EEG signals and music features.

6 Acknowledgement

This work was supported by a grant from the Ministry of Science and Technology under the contract MOST 102-2221-E-001-004-MY3 and the Academia Sinica Career Development Program.

References

1. Dynamical remixing system based on user intention. `http://labrosa.ee.columbia.edu/hamr_ismir2014/proceedings/doku.php?id=dynamical_remixing_system_based_on_user_intention`, [Date of Access: 09-March-2015]
2. Duan, Z., Pardo, B.: Soundprism: an online system for score-informed source separation of music audio. IEEE Journal of Selected Topics in Signal Processing 5(6), 1205–1215 (2011)
3. Fan, R.E., Chang, K.W., Hsieh, C.J., Wang, X.R., Lin, C.J.: LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874 (2008)
4. Guan, C., Thulasidas, M., Wu, J.: High performance P300 speller for brain-computer interface. In: IEEE International Workshop on Biomedical Circuits and Systems (2004)
5. Jao, P.K., Yang, Y.H., Wohlberg, B.: Informed monaural source separation of music based on convolutional sparse coding. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (2015)
6. Miranda, E.R., Brouse, A.: Interfacing the brain directly with musical systems: on developing systems for making music with brain signals. Leonardo 38(4), 331–336 (2005)
7. Mullen, T., Warp, R., Jansch, A.: Minding the (transatlantic) gap: An internet-enabled acoustic brain-computer music interface. In: International Conference on New Interfaces for Musical Expression (2011)
8. Treder, M.S., Purwins, H., Miklody, D., Sturm, I., Blankertz, B.: Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification. Journal of Neural Engineering 11(2), 26009 (2014)
9. Wang, L., Xu, G., Wang, J., Yang, S., Yan, W.: Feature extraction of mental task in bci based on the method of approximate entropy. In: 29th Annual International Conference of the IEEE, Engineering in Medicine and Biology Society (2007)
10. Xu, H., Zhang, D., Ouyang, M., Hong, B.: Employing an active mental task to enhance the performance of auditory attention-based brain-computer interfaces. Clinical Neurophysiology 124(1), 83 – 90 (2013)