DECODING MUSIC ATTENTION FROM "EEG HEADPHONES": A USER-FRIENDLY AUDITORY BRAIN-COMPUTER INTERFACE

Winko W. An[⋆†]*, Barbara Shinn-Cunningham*[†]*, Hannes Gamper*[§]*,*
Dimitra Emmanouilidou[§]*, David Johnston*[§]*, Mihai Jalobeanu*[§]*, Edward Cutrell*[§]*,*
Andrew Wilson[§]*, Kuan-Jung Chiang*[⋆‡]*, Ivan Tashev*[§]

[†] Carnegie Mellon University, Pittsburgh, PA, USA
[‡] University of California San Diego, San Diego, CA, USA
[§] Microsoft Research, Redmond, WA, USA

ABSTRACT

People enjoy listening to music as part of their life. This makes music an excellent choice for designing a user-friendly brain-computer interface (BCI) for long-term use. We propose a novel BCI system using music stimuli that relies on brain signals collected via Smartfones, an EEG recording device integrated into a pair of headphones. In a user study of the proposed system, participants were asked to pay attention to one of three musical instruments playing simultaneously from separate spatial directions. We used a stimulus reconstruction method to decode attention from EEG signals. Results show that the proposed system can achieve good decoding accuracy (>70%) while providing superior user-friendliness compared to a traditional EEG setup.

Index Terms— Auditory attention decoding, BCI, EEG, music

1. INTRODUCTION

A brain-computer interface (BCI) offers a covert and non-verbal way to communicate with a computer. BCIs have great potential in applications including assistive technology and emotion monitoring [1]. Electroencephalography (EEG), due to its mobility, low cost, and proven relevance to cognitive functions [2, 3], has become a popular choice for BCI design. Previous studies have demonstrated great success in building EEG-based BCI systems using visual or auditory stimuli. Chen et al. [4] designed a high-throughput visual BCI system using flickering objects. When the user focuses on one of them, a neural signature known as the steady-state visual evoked potential (SSVEP) appears in EEG signals. However, SSVEP requires a stable line of sight, which may not be available due to permanent or situational impairment (e.g., while driving). As an alternative solution, researchers applied a similar idea to designing auditory BCI systems, where the users were presented with multiple streams of pure tones modulated at different frequencies. The modulation frequency of the attended stream may result in a strong EEG component known as the auditory steady-state response (ASSR) [5].

One major disadvantage of SSVEP or ASSR paradigms is the use of flickering objects or modulated pure tones, which can cause fatigue in users. Recent studies endeavored to use more naturalistic and pleasant stimuli to improve the user-friendliness of BCI systems. Huang et al. [6] used drip-drop sounds in their BCI design, creating a relaxing auditory scene for the users. An et al. [7] designed an

---

⋆ Work done as research intern at Microsoft Research, Redmond.

attention task with human-voiced syllables, and achieved a high accuracy in detecting whether the user is paying attention. In another study, An et al. [8] built an auditory BCI system with a sequence of tones forming melodic patterns.

Here we explore the feasibility of using music stimuli for BCI design. We decode a user's attention to a particular musical instrument while listening to polyphonic music. This idea was previously attempted by Treder et al. [9], who embedded oddballs in music streams and used the oddball-evoked response for attention decoding. Despite achieving a high accuracy, their system averages 40 seconds of data to generate one output, which may be slow for real-time applications. Here, we adopt a different decoding method called auditory attention decoding (AAD) [10] and decode attention within a time window of just 8 seconds. The AAD method linearly combines multi-channel EEG signals to reconstruct a stimulus envelope, which tracks the envelope of the attended stimulus more strongly than the unattended one. This method has been successfully applied in decoding attention to continuous speech for BCI purposes [11, 12]. To further improve the user-friendliness of the design, we used Smartfones (mBrainTrain, Serbia) as the form factor, which is a compact EEG recording device integrated into a pair of headphones. It is a saline-based system with three sensors on top of the head and four on each side around the ear, for a total of 11 sensors. It has less coverage than a traditional EEG cap, but is a good option for this study for its all-in-one design.

2. MATERIALS AND METHODS

2.1. Participants and Stimuli

Nine adults (34.0 ± 3.1 years old, 4 female) volunteered to participate in this study. No participants reported a known history of neurological disorder or hearing loss. The study was reviewed and approved by the Institutional Review Board of Microsoft Research. A written consent was obtained upon participation.

The stimulus used in this study was a four-bar polyphonic piece composed of short melodic excerpts adapted from three popular songs (see Fig. 1a). Each excerpt was assigned to a separate voice and instrument using MuseScore 3: vibraphone for "I'm yours" by Jason Mraz, piano for "Wherever you will go" by The Calling, and harmonica for "Forever young" by Alphaville. We hypothesized that using melodic excerpts from different songs for the three voices and assigning a different instrument to each voice would help listeners to pay attention to one voice at a time. The excerpts chosen followed
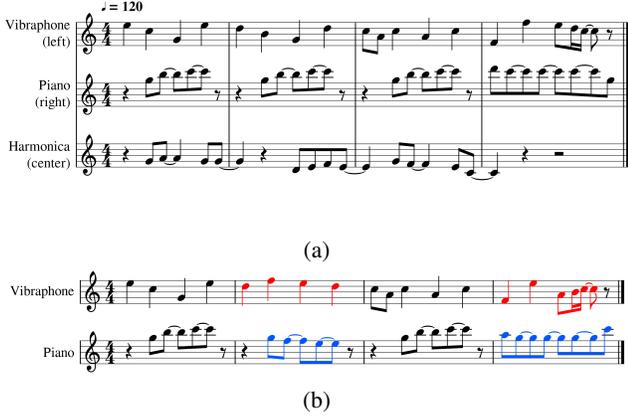
(a)

(b)

Fig. 1. (a) Score sheet of the standard stimuli. (b) The 2nd or 4th bars of the standard stimuli were modified to create oddball bars, colored in red and blue.

the same chord progression (C major - G major - A minor - F major), which would ensure an overall pleasant listening experience.

Each excerpt consisted of four bars, for a total duration of 8 seconds. Besides the original excerpts (Standards), we generated oddball excerpts (Oddballs) by altering the second or the fourth bar of the Standards (Fig. 1b). We created an oddball recognition task (see Sec. 2.2) using these stimuli to motivate participants to listen attentively. The excerpts were spatialized using a set of generic head-related transfer functions [13] to form three streams, where the perceived positions of vibraphone, harmonica, and piano were left, center, and right, respectively. The loudness of these streams was normalized using A-weighting, after which the streams were combined into polyphonic mixtures.

2.2. Experiment

At the start of the experiment, the participants were asked to sit comfortably in front of a computer, read the instructions from the screen, and familiarize themselves with the stimuli. The experiment consisted of 28 trials for attention to vibraphone, 28 trials for attention to piano, and 14 trials for attention to harmonica. For this study, we only focused on generating binary outputs, i.e., distinguishing attention to vibraphone from attention to piano. The data from the attention to harmonica condition were only used for calculating the decoder (see Sec. 2.3) and as a sanity check. All trials were divided into 5 blocks with 14 trials per block, and their order was randomized for each participant. In the beginning of a trial, a left, right or up arrow, was presented on the screen as a visual cue (VC) to direct attention to the instrument on the left, right or center, respectively (see Fig. 2). After a 1-second delay we played two repetitions of the music mixtures through Smartfones. In the stream to be attended, the first repetition was always a Standard, while the second repetition could be either a Standard or an Oddball. The task for participants was to identify whether the two repetitions were the same or different in the attended stream, and answer with a mouse click. Visual feedback (FB) was provided by a green dot displayed for a correct answer, or a red dot for an incorrect answer.

2.3. Auditory attention decoding

EEG signals, sampled at 500 Hz, were passed through a Hamming windowed sinc FIR bandpass filter (2–8 Hz), and were split into
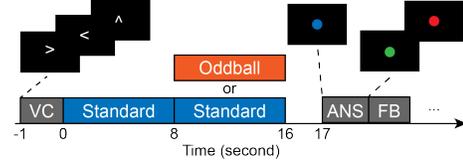

Fig. 2. A trial started with a visual cue (VC) directing attention to the instrument on the left, right or center. It was followed by two music stimuli. Visual feedback (FB) was provided after an answer (ANS) was received.

epochs starting from the onset of each stimulus. Attention was decoded using AAD [10]. The envelopes of the individual voices in the stimuli (cf. Fig. 1) were extracted using the Hilbert Transform, and then lowpass filtered at 8 Hz and downsampled to 64 Hz to derive the stimulus feature $s(t)$ (Fig. 3). The response feature, $r(t)$, was derived by downsampling the bandpass filtered EEG signals to 64 Hz. The AAD algorithm sought to find a decoder $g(\tau, n)$ that could linearly map $r(t)$ back to $s(t)$ [14] as:

$$\hat{s}(t) = \sum_n \sum_\tau r(t + \tau, n) g(\tau, n), \qquad (1)$$

where $\hat{s}(t)$ is the reconstructed stimulus feature, n denotes the EEG channel index, and $0 \leq \tau \leq 600$ ms specifies a range of time-lags relative to the instantaneous occurrence of the stimulus feature, which is used to model the latency between a stimulus envelope and its corresponding envelope-following response in EEG signals. The decoder $g(\tau, n)$ is essentially a spatial-temporal filter that linearly transforms the EEG signals at time-lags τ from 0 to 600 ms post-stimulus to predict the corresponding auditory input. We can estimate $g(\tau, n)$ by minimizing the mean-square-error between the actual stimulus envelope $s(t)$ and the reconstructed envelope $\hat{s}(t)$ plus a regularization term:

$$\min \sum_t [s(t) - \hat{s}(t)]^2 + \lambda \sum_n \sum_\tau g(\tau, n)^2, \qquad (2)$$

where λ is the regularization parameter set to avoid over-fitting. The optimal λ can be determined through cross-validation [15]. The decoder g can be computed using the following equation:

$$\mathbf{g} = (\mathbf{R^T R} + \lambda \mathbf{I})^{-1} \mathbf{R^T s}, \qquad (3)$$

where $\mathbf{R}$ is the matrix of response features $r(t)$ delayed by all possible values in τ with zero padding [14].

Auditory attention was decoded from each epoch. For each participant, we first pooled all epochs of EEG signals except for the one to be decoded to form the response feature $r(t)$. The envelopes of corresponding target voices were concatenated to form the stimulus feature $s(t)$. With the decoder g calculated via (3), we reconstructed a stimulus envelope $\hat{s}(t)$ using (1). We then correlated $\hat{s}(t)$ with the envelopes of each of the vibraphone, piano and harmonica voices in that epoch to generate three correlation coefficients using Pearson's correlation: ρ_{vibr}, ρ_{pian}, ρ_{harm}, respectively. We hypothesize that the correlation between the reconstruction and the envelope of the target instrument to be higher than the ones with the unattended instrument. To verify this hypothesis, we examined the difference between ρ_{vibr} and ρ_{pian} using a paired t-test. The Benjamini-Hochberg method was used to control the false discovery rate (FDR) in multiple comparisons (alpha = 0.05).
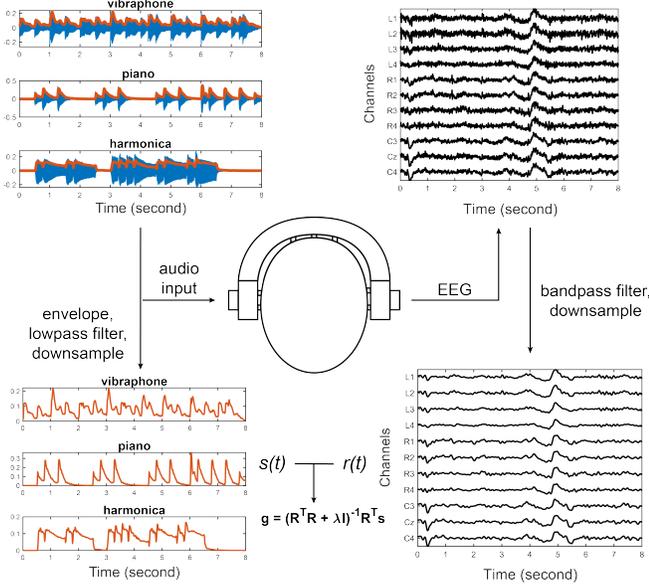
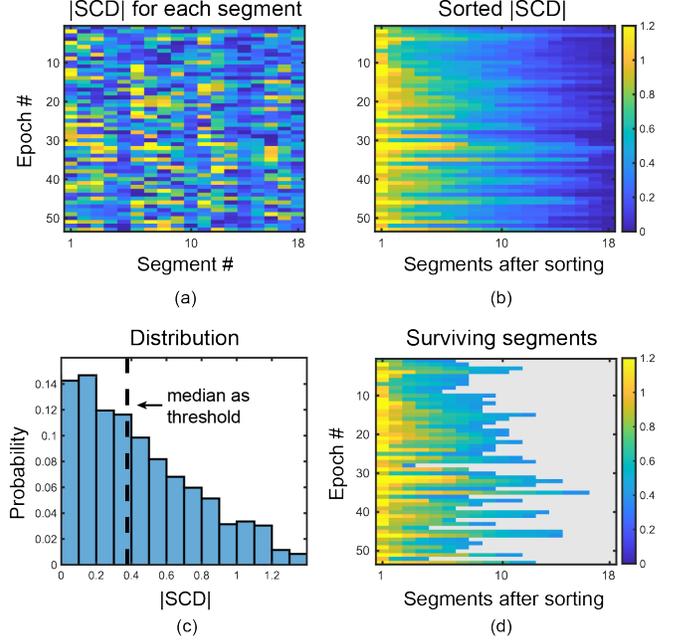**Fig. 3**. Illustration of the auditory attention decoding algorithm

Fig. 4. Illustration of the process of segment-based feature selection. (a) The correlation difference (SCD) was calculated for each segment in each epoch. (b) |SCD| sorted for each epoch (for visualization only). (c) The median of the distribution of all |SCD| values was used as the threshold. (d) The same figure as in (b), but with sub-threshold values masked by grey.

2.4. Segment-based feature selection

The AAD method introduced in Sec. 2.3 uses an entire 8-second epoch to decode auditory attention. However, participants may not sustain their attention throughout the whole epoch, for example due to interference from a distracting stream, or due to the way they scheduled their attention to perform the task. During periods of reduced attention to the target instrument, the neural representation of the masking stimuli might interfere with or mask the target stimulus. We hypothesize that excluding data from periods of reduced attention may reduce noise and improve the overall decoding performance. We added a segment-based feature selection step to exclude irrelevant time segments from decoding. After an epoch-specific decoder g was calculated, we applied it on multiple segments of EEG signals instead of the whole epoch. These segments were 2 s in duration with an overlap of 80%, resulting in a total of 18 segments per epoch. The strength of attention during each segment was estimated by comparing the strength of the correlation of the EEG with the vibraphone and the piano envelopes. Specifically, we calculated the absolute value of the segment-wise correlation difference (|SCD|, Fig. 4a), defined as:

$$|\text{SCD}_k| = |\rho_{\text{vibr},k} - \rho_{\text{pian},k}|, \tag{4}$$

where $k = \{1, \cdots, 18\}$ is the segment index, and $\rho_{\text{vibr},k}$ or $\rho_{\text{pian},k}$ represent the correlation between a segment-wise reconstruction with its corresponding segment-wise vibraphone envelope or piano envelope, respectively. If attention to either the vibraphone or piano is strong during a particular segment, the neural response for that segment should resemble the attended instrument voice more than the unattended one, i.e., |SCD| should be non-zero. During segments with reduced attention, |SCD| should approach zero.

Segments with small |SCD| values were excluded from analysis. The threshold was determined by the distribution of all |SCD| values in the training data (see Fig. 4c). Values above the median of the distribution were retained (see Fig. 4d). The correlations $\rho_{\text{vibr},k}$ and $\rho_{\text{pian},k}$ of surviving segments in each epoch were averaged to calculate ρ_{vibr} and ρ_{pian} after feature selection, respectively. A paired

t-test was conducted to reveal any statistically significant change in these correlation measures with and without feature selection (alpha = 0.05, FDR corrected).

2.5. Classification

We used ρ_{vibr} and ρ_{pian} as the features to decode attention to vibraphone and attention to piano. We trained and tested a subject-specific linear support vector machine using leave-one-out cross-validation with 1000 repetitions. Classification was run on data with and without feature selection separately.

3. RESULTS AND DISCUSSION

3.1. Correlation with envelopes

The correlation between the reconstructed envelope and the attended stimulus envelope is strongly modulated by attention, even without feature selection. When the participants were paying attention to the vibraphone, their average ρ_{vibr} was significantly higher than ρ_{pian} (p<0.001, Fig. 5a). When attention was on the piano, ρ_{pian} was greater than ρ_{vibr}. However, the difference was not found to be statistically significant (p=0.063). In both conditions, ρ_{harm} was around 0 for all participants.

With the segment-based feature selection, the differences between ρ_{vibr} and ρ_{pian} were magnified. For the attention to vibraphone condition, feature selection significantly boosted ρ_{vibr} (p<0.001, Fig. 5b) and suppressed ρ_{pian} (p<0.001). Similarly, ρ_{vibr} was suppressed by feature selection when attention was on piano (p=0.007), with ρ_{pian} statistically unchanged (p=0.906). We
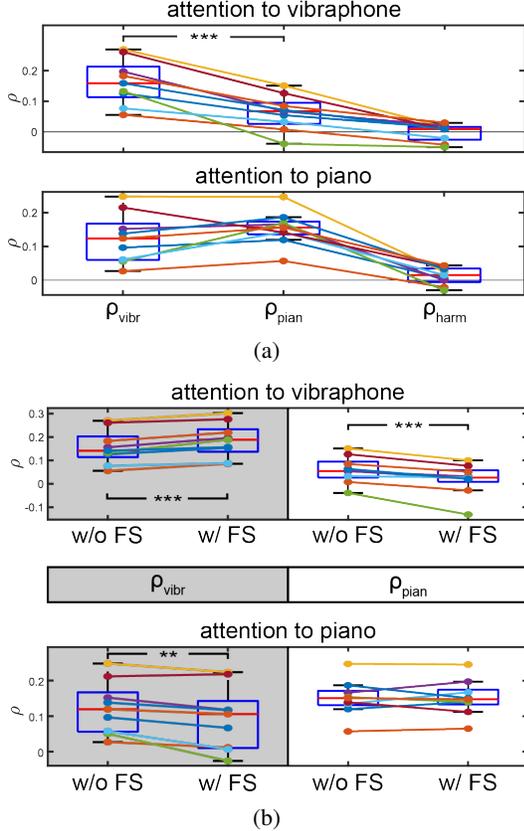
Fig. 5. (a) Correlation between the reconstruction and the envelope of vibraphone (ρ_{vibr}), piano (ρ_{pian}) or harmonica (ρ_{harm}) without feature selection. Each line represents a subject. (b) Comparison of correlations with feature selection (w/ FS) and without (w/o FS). **, p< 0.01; ***, p< 0.001; FDR corrected for multiple comparisons.

conclude that the proposed feature selection method identified segments relevant for the classifier to determine which instrument the participant paid attention to.

3.2. Decoding accuracy

Experimental results indicate that the proposed method allows decoding of attention to music. Without feature selection, the average decoding accuracy was 63.77%, which is above the significant chance level (60.71%) with 95% confidence [16] (see Fig. 6). We also observed great individual variability in the results, a known observation in many auditory BCI studies [7, 17].

The positive effect of feature selection on correlation measures (see Sec. 3.1) resulted in a boost in decoding accuracy. With segment-based feature selection, the average decoding accuracy improved to 71.23% (Fig. 6), with a performance gain observed for all participants. Notably, this gain was more remarkable for subjects with a low decoding score before feature selection was implemented — the subjects with a decoding accuracy below 65% (Subject 4, 8, 2 and 1, see Fig. 6) benefited an average of 11.0% from feature selection, which led to much smaller individual variability in the results. The decoding performance achieved in this study is comparable to previous works on auditory BCI using the same linear decoding method [11, 12], despite the use of a user-friendly EEG recording
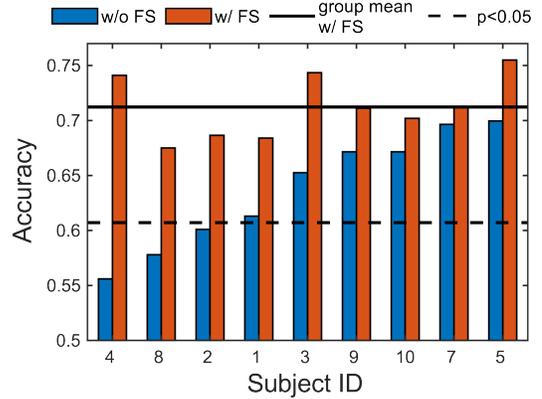

Fig. 6. Decoding accuracy with feature selection (w/ FS) and without (w/o FS). The average for w/ FS is 71.23%. The subjects are sorted by their decoding accuracy w/o FS in ascending order.

Table 1. Comparison with previous studies using AAD

Study	Sensors (#, type)	Sample length (s)	Accu. (%)	ITR (bits/min)
O'Sullivan et al. [10]	128, gel	~60	89.0	0.50
Ciccarelli et al. [12]	64, gel	10	66.0	0.45
	18, dry	10	59.0	0.14
here	11, saline	8	71.2	1.01

device with fewer sensors, less spatial coverage and lower signal-to-noise ratio compared to a conventional EEG cap. In addition, since we decoded attention with short data (8 seconds), the overall efficiency of the BCI system, evaluated by its information transfer rate (ITR) [18], is higher than similar studies with longer decoding windows (1.01 bit/min compared to ≤0.50 bits/min) [10, 12] (see Table 1).[1] One limitation of this study, however, is the small number of participants recruited (nine), which will be improved in follow-up studies in the future.

4. CONCLUSIONS

This study investigated the feasibility of building a user-friendly BCI system by decoding auditory attention. The proposed system relies on short musical stimuli with three voices. Due to its harmonic nature, this stimulus type may be more pleasant to listen to than previously proposed auditory stimuli like modulated pure tones or tone sequences and thus better suited for long-term use in a BCI system. Furthermore, the proposed system uses a compact headphone-based form factor with fewer sensors and requires much less effort in system setup than a traditional EEG system, which may be an appealing feature for novel users.

---

[1] Only results obtained from linear AAD were compared with results in this study. ITR was calculated based on the number of classes, sample length and decoding accuracy reported in these studies.

5. REFERENCES

[1] Joong Hoon Lee, Hannes Gamper, Ivan Tashev, Steven Dong, Siyuan Ma, Jacquelin Remaley, James D. Holbery, and Sang Ho Yoon, "Stress monitoring using multimodal bio-sensing headset," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–7.

[2] Yuqi Deng, Inyong Choi, and Barbara Shinn-Cunningham, "Topographic specificity of alpha power during auditory spatial attention," *NeuroImage*, vol. 207, pp. 116360, feb 2020.

[3] Winko W. An, Kin-Hung Ting, Ivan P. H. Au, Janet H. Zhang, Zoe Y. S. Chan, Irene S. Davis, Winnie K. Y. So, Rosa H. M. Chan, and Roy T. H. Cheung, "Neurophysiological Correlates of Gait Retraining With Real-Time Visual and Auditory Feedback," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 6, pp. 1341–1349, jun 2019.

[4] Jingjing Chen, Dan Zhang, Andreas K. Engel, Qin Gong, and Alexander Maye, "Application of a single-flicker online SSVEP BCI for spatial navigation," *PLoS ONE*, vol. 12, no. 5, pp. e0178385, may 2017.

[5] Do Won Kim, Jae Hyun Cho, Han Jeong Hwang, Jeong Hwan Lim, and Chang Hwan Im, "A vision-free brain-computer interface (BCI) paradigm based on auditory selective attention," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2011, pp. 3684–3687.

[6] Minqiang Huang, Jing Jin, Yu Zhang, Dewen Hu, and Xingyu Wang, "Usage of drip drops as stimuli in an auditory P300 BCI paradigm," *Cognitive Neurodynamics*, vol. 12, no. 1, pp. 85–94, 2018.

[7] Winko W An, Alexander Pei, Abigail L Noyce, and Barbara Shinn-cunningham, "Decoding auditory attention from single-trial EEG for a high-efficiency brain-computer interface," in *42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2020, pp. 3456–3459.

[8] Winko W. An, Hakim Si-Mohammed, Nicholas Huang, Hannes Gamper, Adrian KC Lee, Christian Holz, David Johnston, Mihai Jalobeanu, Dimitra Emmanouilidou, Edward Cutrell, Andrew Wilson, and Ivan Tashev, "Decoding auditory and tactile attention for use in an EEG-based brain-computer interface," in *2020 8th International Winter Conference on Brain-Computer Interface (BCI)*, feb 2020, pp. 1–6.

[9] 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, pp. 026009, 2014.

[10] James A. O'Sullivan, Alan J. Power, Nima Mesgarani, Siddharth Rajaram, John J. Foxe, Barbara G. Shinn-Cunningham, Malcolm Slaney, Shihab A. Shamma, and Edmund 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.

[11] Ali Aroudi, Tobias De Taillez, and Simon Doclo, "Improving Auditory Attention Decoding Performance of Linear and Non-Linear Methods using State-Space Model," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing*, may 2020, pp. 8703–8707.

[12] Gregory Ciccarelli, Michael Nolan, Joseph Perricone, Paul T. Calamia, Stephanie Haro, James O'Sullivan, Nima Mesgarani, Thomas F. Quatieri, and Christopher J. Smalt, "Comparison of Two-Talker Attention Decoding from EEG with Nonlinear Neural Networks and Linear Methods," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[13] Bill Gardner and Keith Martin, "HRTF Measurements of a KEMAR Dummy-Head Microphone MIT Media Lab Perceptual Computing-Technical Report #280," Tech. Rep., 1994.

[14] Michael J. Crosse, Giovanni M. Di Liberto, Adam Bednar, and Edmund C. Lalor, "The multivariate temporal response function (mTRF) toolbox: A MATLAB toolbox for relating neural signals to continuous stimuli," *Frontiers in Human Neuroscience*, vol. 10, pp. 1–14, nov 2016.

[15] Stephen V. David and Jack L. Gallant, "Predicting neuronal responses during natural vision," *Network: Computation in Neural Systems*, vol. 16, no. 2-3, pp. 239–260, 2005.

[16] Etienne Combrisson and Karim Jerbi, "Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy," *Journal of Neuroscience Methods*, vol. 250, pp. 126–136, 2015.

[17] Netiwit Kaongoen and Sungho Jo, "A novel hybrid auditory BCI paradigm combining ASSR and P300," *Journal of Neuroscience Methods*, vol. 279, pp. 44–51, 2017.

[18] Jonathan R. Wolpaw, Niels Birbaumer, Dennis J. McFarland, Gert Pfurtscheller, and Theresa M. Vaughan, "Brain-computer interfaces for communication and control," *Clinical Neurophysiology*, vol. 113, no. 6, pp. 767–791, 2002.