Decoding auditory attention from single-trial EEG for a high-efficiency brain-computer interface*

Winko W. An[1], Alexander Pei[1], Abigail L. Noyce[2], and Barbara Shinn-Cunningham[1,2,3]

Abstract— **Brain-computer interface (BCI) systems enable humans to communicate with a machine in a non-verbal and covert way. Many past BCI designs used visual stimuli, due to the robustness of neural signatures evoked by visual input. However, these BCI systems can only be used when visual attention is available. This study proposes a new BCI design using auditory stimuli, decoding spatial attention from electroencephalography (EEG). Results show that this new approach can decode attention with a high accuracy (>75%) and has a high information transfer rate (>10 bits/min) compared to other auditory BCI systems. It also has the potential to allow decoding that does not depend on subject-specific training.**

I. Introduction

Electroencephalography (EEG) offers a noninvasive and portable method for monitoring brain activity, making it a popular technology for brain-computer interfaces (BCIs) [1]. Many successful BCI systems use visual stimuli as the sensory input, and decode a user's attention from neural signatures such as event-related potentials (ERPs). These visual paradigms efficiently transmit information to a computer, as quantified by their information transfer rate (ITR). For example, one recent study on visual ERP-based BCI reported an average ITR as high as 20.26 bits/min [2].

Though visual BCI systems are efficient, they cannot be used in scenarios where visual attention is already engaged by real world demands (e.g. walking or driving), or by users with visual impairment. Some previous studies developed auditory BCI systems to tackle these problems. For example, Kim et al. [3] used multiple streams of spatialized modulated signal with a constant-frequency carrier as the stimuli, and decoded attention from auditory steady-state response (ASSR). An et al. [4] used more user-friendly stimuli, synthesized melodies, and developed a novel BCI paradigm. In another study to reduce user fatigue, Huang et al. [5] proposed using drip drop sounds as the input. However, the efficiency of these auditory systems is substantially lower than most visual-based BCIs. For example, ITRs of the three aforementioned studies were all below 3 bits/min, making them less useful in real applications.

The current study proposed an auditory BCI system with high efficiency. It used spatialized human-voiced syllables as the stimuli, and decoded selective attention from EEG. Inspired by previous studies on auditory attention [6], [7], [8], [9], both ERP and EEG spectrogram were used as features to train and test a series of linear classifiers for attention decoding.

*This work was supported by the Office of Naval Research (Project ID: N00014-18-1-2069).

[1]Winko W. An, Alexander Pei and Barbara Shinn-Cunningham are with the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA (email: bgsc@andrew.cmu.edu)

[2]Abigail L. Noyce and Barbara Shinn-Cunningham are with the Neuroscience Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA

[3]Barbara Shinn-Cunningham is also with the Department of Biomedical Engineering, Boston University, MA 02215, USA

II. Methods

A. Participants

Thirty adults (21.95 ± 4.95 years old, 15 female) participated in this study. No participant reported hearing loss or any history of neurological disorders. The Institutional Review Board of Boston University approved this study. All participants gave written informed consent, and were paid for taking part in the study.

B. Experiment

Before the experiment started, participants were asked to sit comfortably in a soundproof booth in front of a computer monitor. The syllables /ba/, /da/ and /ga/, spoken by native English talkers, were used as stimuli. The syllables were spatialized by a set of generic head-related transfer functions (Media Lab, MIT), and played through a pair of insert earphones (ER1, Etymotic Research). The intensity of sound was adjusted to a comfortable listening level for each individual to ensure syllable intelligibility. The simulated locations of these syllables were 90°from the left (L90), center, or 90°from the right (R90, Fig. 1a), in the horizontal plane.

At the beginning of each trial, a visual cue (VC) was shown on the screen for one second, which could be one of the two words: "Space" or "Relax" (Fig. 1b). "Space" indicated that participants should direct spatial attention in the upcoming trial, while "Relax" represented a control trial where no attention would be required. An auditory cue (AC) — a spatialized /a/ sound — was given after the VC to direct the participant's attention. In "Space" attention trials, the AC specified the target location (either L90 or R90). In no-attention "Relax" trials, the AC always came from the center (i.e. a neutral value). After a 1000 ms silent period, a 4-syllable mixture was played. All syllables were 600 ms long, and their onsets were separated by 300 ms. In "space" attention trials, the first and the last syllables were always distractors (D1) played from the center. Of the second and third syllables, one was the target (T), which came from the AC location. The other syllable was the second distractor (D2) that came from the opposite side. Syllables /ba/, /da/, and /ga/ were randomly permuted among T, D1 and D2. The task was to ignore D1 and D2, and to identify T using the

978-1-7281-1990-8/20/$31.00 ©2020 IEEE

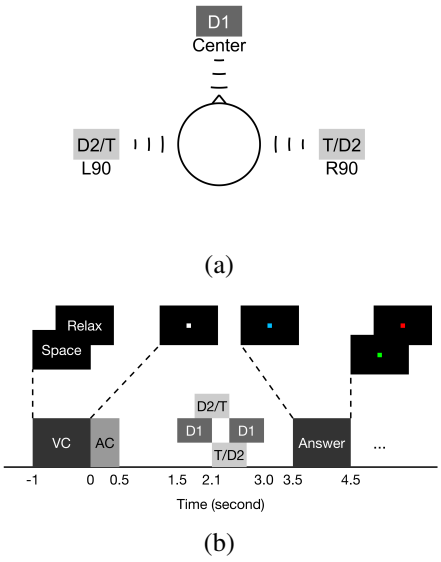


Fig. 1. (a) Spoken syllables were spatialized to center and to 90°left (L90) or right (R90), always in the horizontal plane. The syllable from the center was always a distractor (D1). The target (T) might be at L90 or R90, with a second distractor (D2) at the opposite side. (b) Illustration of the events within a trial. A visual cue (VC) was followed by an auditory cue (AC). A 4-syllable mixture was played 1 second after the AC. The participants should respond when the fixation dot turned blue. A green or red dot was presented at the end of the trial, showing the correctness of the response.

keyboard ("1" for /ba/, "2" for /da/, and "3" for /ga/). Visual feedback was given after each response to show whether the answer was correct. In no-attention trials, the participants were asked to ignore all syllables, and give a random answer at the end. The inter-trial interval was set to be 2 seconds with jitter.

Each participant completed 756 trials in total. These trials differed in the locations and talkers of the spoken syllables, and in the type of attention (spatial or non-spatial) required. Data from only four conditions (72 trials in each condition) are presented in this analysis: 1) selective spatial attention trials in which the four syllables occur in location order Center-Left-Right-Center (Spa_LR); 2) selective spatial attention trials with syllables in order Center-Right-Left-Center (Spa_RL); 3) a control no-attention condition with syllables in order Center-Left-Right-Center (Ctr_LR); 4) a control no-attention condition with syllables in order Center-Right-Left-Center (CTR_RL). From these four conditions, we attempted two binary attention decoding problems: 1) Spa_LR vs Ctr_LR; and 2) Spa_RL vs Ctr_RL. Note that for each of these comparisons, the stimuli presented were exactly matched; only the instructions given to the participant differed. The following sections of this paper will explore the differences in neural signatures for each pair, and how effectively attention can be decoded.

C. EEG processing

EEG was collected using a 64-channel Biosemi system throughout the experiment, sampled at 2048 Hz. Raw EEG data first were bandpass filtered (0.1 - 50 Hz), and were then downsampled to 256 Hz. An independent component analysis was conducted subsequently using EEGLAB [10], [11]. Components that represented eye blinks, eye movements, and muscle artifacts were removed from further analysis.

D. ERP and time-frequency analysis

The continuous EEG data were segmented into epochs to study differences in ERPs and oscillation activity between conditions. In this study, ERP is defined as the condition-wise average EEG waveform time-locked to the onset of the first syllable. The spectro-temporal representation of EEG was studied using a continuous wavelet transform (CWT) implemented using a custom MATLAB script. The wavelet bases (Morlet wavelet with $\omega_0 = 6$) were normalized to have unit total energy at all scales [12].

A group-level cluster-based permutation test [13], implemented with FieldTrip [14], was used to examine the difference in ERP and in CWT between each spatial attention condition (i.e. Spa_LR and Spa_RL) and its corresponding control condition (i.e. Ctr_LR and Ctr_RL, respectively). Both cluster-forming and cluster-significance thresholds were set at 0.05.

E. Feature extraction and classification

Subject-specific linear discriminant analysis (LDA) models were used to decode attention from single-trial EEG data for each of the two classification problems (i.e., Spa_LR vs Ctr_LR, and Spa_RL vs Ctr_RL). Inspired by the results in Section III-A and III-B, the feature used for training and testing the model contained multi-channel EEG time-courses as well as the magnitude of the CWT, averaged within each 100 ms interval between 1500 ms and 2700 ms after the onset of the AC (i.e., from the onset of the first syllable to the offset of the third syllable). The CWT magnitudes were also averaged within five frequency bands: delta (2 – 4 Hz), theta (4 – 8 Hz), alpha (8 – 14 Hz), beta (14 – 30 Hz) and gamma (30 – 40 Hz). The decoding accuracy of each binary classification was derived from a leave-one-trial-out cross-validation with 1000 repetitions.

III. Results

A. ERP analysis

The differences in ERPs between spatial attention and control conditions are shown in Fig. 2. Significant differences were observed in frontal and parietal channels at multiple time instances. The topographic pattern of the ERP difference was similar for the two contrasts, with a slight difference in the lateralization of the positivity at 1900 ms and 2200 ms. Such lateralization is likely affected by the spatialized location of the syllable being played at those moments.

B. Time-frequency analysis

Event-related synchronization (ERS) and desynchronization (ERD), defined as the percent change in value from one condition to a baseline (i.e., Ctr_LR and Ctr_RL in this study), were used to evaluate the signal change in

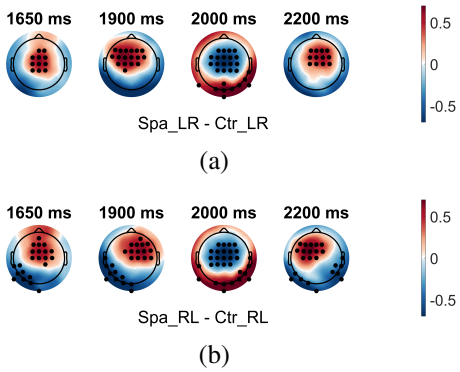


Fig. 2. Topographic maps of ERP differences between spatial attention and control conditions. Time stamps are with respect to the onset of the auditory cue. Solid dots represent channels with significant effects ($p < 0.05$). Unit: μV

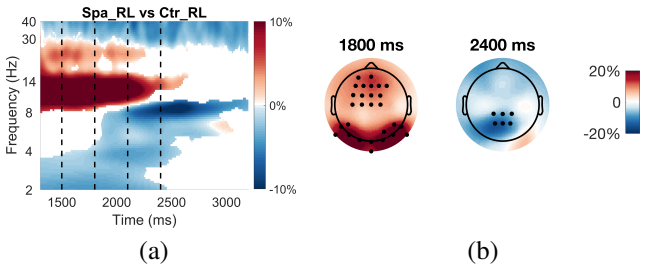


Fig. 3. (a) Average event-related synchronization (positive values) or desynchronization (negative values) across all channels. The values are masked by their significance derived from a non-parametric statistical test ($p < 0.05$). Black dashed lines represent the onset of four syllables. (b) Topographic maps of the alpha power difference between Spa_RL and Ctr_RL. Time stamps are with respect to the onset of the auditory cue. Solid dots represent channels with significant effects for at least one frequency bin ($p < 0.05$).

the time-frequency domain when attention was engaged. Strong ERS was seen in the alpha band before the onset of the last distractor (2400 ms, Fig. 3a). Higher values of alpha ERS were seen in the frontal and parieto-occipital sensors (Fig. 3b). In addition, the ERS in the beta band, and the ERDs in the delta, theta and gamma band were also significant in at least one channel throughout the stimuli period.

C. Decoding accuracy

Attention can be decoded accurately from EEG in most participants. All results were above 50%, the absolute chance level for a binary classification (Fig. 4). However, since studies on brain signal classification are generally susceptible to a high false positive rate, Combrisson and Jerbi [15] proposed a method to correct the chance level based on sample size, number of classes, and the desired confidence interval. Even with the corrected chance level (56.94%, 95% confidence), only one classification fell below chance (Fig. 4a). Table I shows the average decoding accuracy, their equivalent ITR, and the best ITR among all participants.

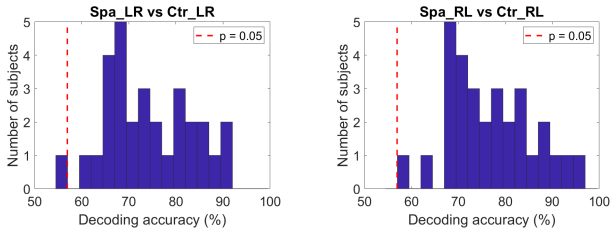


Fig. 4. Histogram of decoding accuracy. The red dashed lines at 56.94% represent the corrected chance level.

TABLE I

DECODING RESULTS & INFORMATION TRANSFER RATE (ITR)

Conditions	Average accuracy	Average ITR (bits/min)	Best ITR (bits/min)
Spa_LR vs Ctr_LR	74.15%	9.44	23.53
Spa_RL vs Ctr_RL	75.83%	10.25	31.70

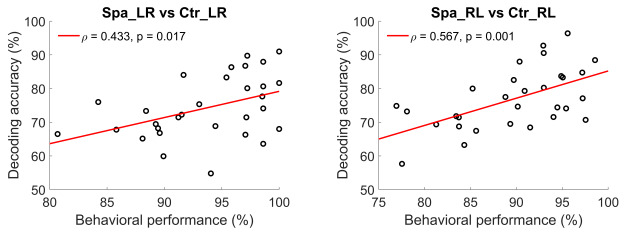
D. Behavioral correlate

In order to explore the relationship between decoding and attentional effort, the decoding accuracy for each participant was correlated with behavioral performance. In this study, behavioral performance is defined as the percent correct of the syllable identification tasks in spatial attention trials (see Section II-B), which represents a proxy for the participant's mental engagement during the task. The results showed strong correlation between behavior and decoding accuracy for both Spa_LR vs Ctr_LR (ρ = 0.433, p = 0.017) and Spa_RL vs Ctr_RL (ρ = 0.567, p = 0.001, Fig. 5).

IV. DISCUSSION

This study introduced a new auditory BCI system that can generate a binary output within 2 seconds. Human-voiced syllables were used as the stimuli, which are natural, user-friendly, and unlikely to cause fatigue even with extensive usage. Users can voluntarily attend or ignore these stimuli to control the value of the output (e.g., "yes" or "no"). The efficiency of the proposed system is substantially greater than that reported in previous studies that used modulated signals [3], [16], melodies [4], or drip drop sounds [5] as the stimuli. The best ITRs across participants even outperformed some visual BCI systems [2], [17]. The high ITR achieved in this study was due in part to the use of short trials — the classifications were run with only 1.2 seconds of EEG data. Such a brief delay between attentional control and a BCI output may even enable a conversation-level interaction with a computer. To achieve even higher efficiency, in the future, we will explore the feasibility of decoding the direction of spatial attention (left or right) from single-trial EEG. Together with the no-attention condition, we can build a 3-way classifier, which may have better value in real applications.

The current decoding method uses high-dimensional features for classification. Inspired by results in the ERP and the time-frequency analysis, these features contain information



(a) (b)

Fig. 5. Scatter plots showing each subject's behavioral performance and decoding accuracy.

represented in either the time domain or spectro-temporal domain. However, these features may not contribute equally to classification. Including irrelevant features may even decrease the accuracy of the model. Similarly, some EEG channels may contribute more than the others. Shrinking the number of channels while maintaining a high decoding score, if possible, would be important to building unobtrusive BCI systems with few channels. In the future, we will conduct a feature selection analysis by estimating feature weights, and reduce the dimensionality of features used for classification.

It is nearly impossible for participants to sustain full attention throughout the whole experiment. At least some of their incorrect responses during spatial attention tasks are likely due to attention drifting. The strong correlation between decoding accuracy and behavioral performance suggests that some of the wrong classifications might simply originate from a lack of attentional effort during such trials. Therefore, the proposed BCI system has the potential to achieve even higher efficiency if the user is always fully engaged, which is usually the case during real-life applications.

Significant differences in ERP and CWT were shown in group-level statistics, suggesting that some of the contrasting features are common across subjects. Although user-specific classifiers were our main focus, this suggests that a general decoder might be feasible that is not trained on individual subjects. Such a decoder would largely reduce the amount of time and data required to implement a system for a new user. A future study on the feasibility of building a general classifier for all participants is warranted.

V. Conclusions

The current study proposed a new BCI system based on auditory attention. It not only yielded high efficiency compared with previously reported auditory BCI systems, but also presented pleasant, user-friendly stimuli that allow comfortable long-term use. The system also has the potential to allow decoding that does not depend on subject-specific training.

References

[1] Reza Abiri, Soheil Borhani, Eric W. Sellers, Yang Jiang, and Xiaopeng Zhao, "A comprehensive review of EEG-based brain-computer interface paradigms," *Journal of Neural Engineering*, vol. 16, no. 1, pp. 011001, 2019.

[2] Zhimin Lin, Chi Zhang, Ying Zeng, Li Tong, and Bin Yan, "A novel P300 BCI speller based on the Triple RSVP paradigm," *Scientific Reports*, vol. 8, no. 1, pp. 3350, dec 2018.

[3] 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.

[4] 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, IEEE.

[5] 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.

[6] Inyong Choi, Le Wang, Hari Bharadwaj, and Barbara Shinn-Cunningham, "Individual differences in attentional modulation of cortical responses correlate with selective attention performance," *Hearing Research*, vol. 314, pp. 10–19, 2014.

[7] Yuqi Deng, Robert MG Reinhart, Inyong Choi, and Barbara G Shinn-Cunningham, "Causal links between parietal alpha activity and spatial auditory attention," *eLife*, vol. 8, pp. e51184, nov 2019.

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

[9] Martin Spüler and Simone Kurek, "Alpha-band lateralization during auditory selective attention for brain–computer interface control," *Brain-Computer Interfaces*, vol. 5, no. 1, pp. 23–29, jan 2018.

[10] Arnaud Delorme and Scott Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, mar 2004.

[11] Jason A. Palmer, Kenneth Kreutz-Delgado, and Scott Makeig, "Super-Gaussian Mixture Source Model for ICA," Lecture Notes in Computer Science, pp. 854–861. Springer Berlin Heidelberg, mar 2006.

[12] 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.

[13] Eric Maris and Robert Oostenveld, "Nonparametric statistical testing of EEG- and MEG-data," *Journal of Neuroscience Methods*, vol. 164, no. 1, pp. 177–190, 2007.

[14] Robert Oostenveld, Pascal Fries, Eric Maris, and Jan Mathijs Schoffelen, "FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data," *Computational Intelligence and Neuroscience*, vol. 2011, pp. 156869, 2011.

[15] 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.

[16] Hyun Jae Baek, Min Hye Chang, Jeong Heo, and Kwang Suk Park, "Enhancing the usability of brain-computer interface systems," *Computational Intelligence and Neuroscience*, vol. 2019, pp. 5427154, 2019.

[17] G Townsend, B K LaPallo, C B Boulay, D J Krusienski, G E Frye, C K Hauser, N E Schwartz, T M Vaughan, J R Wolpaw, and E W Sellers, "A novel P300-based brain-computer interface stimulus presentation paradigm: moving beyond rows and columns.," *Clinical neurophysiology*, vol. 121, no. 7, pp. 1109–20, jul 2010.