Decoding of Attentional Selection in a Cocktail Party Environment from Single-Trial EEG is Robust to Task

Timo Lauteslager, James A. O'Sullivan, Richard B. Reilly, *IEEE Senior Member*, and Edmund C. Lalor

Abstract— **Recently it has been shown to be possible to ascertain the target of a subject's attention in a cocktail party environment from single-trial (~60 s) electroencephalography (EEG) data. Specifically, this was shown in the context of a dichotic listening paradigm where subjects were cued to attend to a story in one ear while ignoring a different story in the other and were required to answer questions on both stories. This paradigm resulted in a high decoding accuracy that correlated with task performance across subjects. Here, we extend this finding by showing that the ability to accurately decode attentional selection in a dichotic speech paradigm is robust to the particular attention task at hand. Subjects attended to one of two dichotically presented stories under four task conditions. These conditions required subjects to 1) answer questions on the content of both stories, 2) detect irregular frequency fluctuations in the voice of the attended speaker 3) answer questions on both stories and detect frequency fluctuations in the attended story, and 4) detect target words in the attended story. All four tasks led to high decoding accuracy (~89%). These results offer new possibilities for creating user-friendly brain computer interfaces (BCIs).**

I. Introduction

In recent years, a number of studies have utilized linear regression methods in order to estimate response functions that index how neural activity changes in response to the presentation of natural continuous speech [1-4]. This has been very useful for studying the cocktail party problem; that is, our ability to attend to a single speaker in a multi-speaker environment [5]. For example, using electroencephalography (EEG) data, it has been shown to be possible to derive separate temporal response functions (TRFs) to the amplitude envelopes of two competing speech streams, with attentional effects evident at around 200ms post-stimulus [6]. However, such effects were only discernible after averaging across many trials and subjects, a lack of sensitivity not atypical of EEG-based cognitive neuroscience studies. More recently, it has been shown to be possible to determine which speaker a subject is attending to on a single-trial basis (~60 s) [7,8]. Specifically, this was shown in the context of a dichotic listening paradigm where subjects were cued to attend to a story in one ear while ignoring a story in the other and were required to answer questions on both stories [7]. This paradigm resulted in a high decoding accuracy that correlated with task performance across subjects. Across 40 subjects, a mean decoding-accuracy of 89% was achievable. Here, with the long term goal of increasing decoding-accuracy within this relatively naturalistic stimulus paradigm, we investigate the effects of manipulating the attention task on decoding-accuracy. Our hypothesis is that tasks that require subjects to attend to low- or high-level information, or a combination of both, may lead to more easily distinguished attentional effects in the EEG data. We show that attentional selection of one speech stream in our dichotic listening paradigm can be accurately decoded using single-trial EEG across four different attention tasks. These findings have implications for the future development of naturalistic and user-friendly brain computer interfaces (BCIs) and for future studies into the difficulties encountered by certain cohorts when attempting to solve the cocktail party problem [9].

II. Methods

A. Participants

Fourteen human subjects took part (mean ± standard deviation (SD) age, 24.4 ± 4.1 years; 7 male; 1 left-handed). The experiment was undertaken in accordance with the Declaration of Helsinki. The Ethics Committee of the School of Psychology at Trinity College Dublin approved the experimental procedures and each subject provided written informed consent. Subjects reported no history of hearing impairment or neurological disorder.

B. Stimuli and Procedures

Subjects undertook 40 trials, each of ~60 s in length, where they were presented with 2 classic works of fiction: one to the left ear, and the other to the right ear. Each story was read by a different male speaker and, for both stories, each trial began where the story ended on the previous trial. Each subject attended to the story in either the left or right ear throughout all 40 trials (7 subjects to the left, 7 subjects to the right). However the specific attention task was changed after every 10 trials. The four different attention tasks required subjects to: 1) answer questions on the content of the attended and unattended stories, 2) detect irregular frequency fluctuations in the voice of the attended speaker, 3) answer questions on the content of the stories and detect frequency fluctuations in the attended speaker's voice, and 4) detect target words in the attended stream. Specifically, in

Research supported by Science Foundation Ireland (09-RFP-NES2382).

T. Lauteslager is with the School of Engineering, Trinity Centre for Bioengineering and Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland, (e-mail: lauteslt@tcd.ie).

J. A. O'Sullivan is with the School of Engineering, Trinity Centre for Bioengineering and Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland, (e-mail: osullij8@tcd.ie).

R. B. Reilly is with the School of Engineering, School of Medicine, Trinity Centre for Bioengineering and Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland, (e-mail: richard.reilly@tcd.ie).

E. C. Lalor is with the School of Engineering, Trinity Centre for Bioengineering and Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland, (e-mail: edlalor@tcd.ie).

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

condition 1, subjects were required to answer between 4 and 6 multiple-choice questions on both stories after each 60 s trial. Each question had 4 possible answers. In condition 2, frequency fluctuation targets consisted of segments of the original audio, which were periodically shifted forward and backward in time at a rate of 12 Hz, resulting in a ‘vibrato’ effect (Figure 1). Targets lasted 250 ms and occurred 7 times per trial on average. The intensity of the vibrato effect, and thereby the difficulty of detecting the targets, was adjusted in real time to maintain a target detection rate of approximately 75%. In condition 3, subjects had to answer the same style of question as that from condition 1 and to detect the same type of frequency fluctuation as that in condition 2. In condition 4, subjects were required to respond by button-click within 2.5 s of the presentation of the word “and”, which occurred on average approximately 4 times per trial. The order in which each condition was presented was randomized between subjects. Stimulus amplitudes in each audio stream within each trial were normalized to have the same root mean squared (RMS) intensity, and silent gaps exceeding 0.5 s in the speech streams were truncated to 0.5 s in duration. Stimuli were presented using Sennheiser HD650 headphones and Presentation software from Neurobehavioral Systems (http://www.neurobs.com). Subjects were instructed to maintain visual fixation on a crosshair centered on the screen for the duration of each trial, and to minimize eye blinking and all other motor activities.

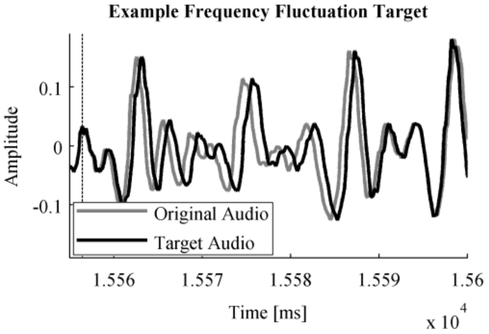

Figure 1. Example of a Frequency Fluctuation Target. Original audio is shown in grey, the target audio is shown in black. The periodical delay causes a transient fluctuation in the frequency content of the audio, resulting in a ‘vibrato’ effect. Onset of the target is denoted by a dashed line.

C. Data Acquisition and Preprocessing

Electroencephalography data were recorded using 130 electrode positions (128 scalp plus 2 mastoids). The data were filtered over the range 0–134 Hz and digitized at the rate of 512 Hz using a BioSemi Active Two system. Data were re-referenced offline to the average of all scalp channels.

In order to decrease the processing time required, all EEG data were downsampled by a factor of 8 to give an equivalent sampling rate of 64 Hz, after applying a zero phase-shift antialiasing filter. The amplitude envelopes of the speech signals were obtained using a Hilbert transform, and then downsampled to the same sampling rate of 64 Hz to allow us to relate their dynamics to those of the EEG. Because envelope frequencies between 0.5 and 8 Hz are linearly relatable to the EEG [5,10], the EEG data were digitally filtered offline with a band-pass filter between 0.5 and 8 Hz. The speech envelopes were low-pass filtered with a cutoff frequency of 8 Hz.

D. Stimulus-Reconstruction

We wished to determine how accurately we could estimate to which of the two speakers each subject was attending. Our strategy for this was centered on the approach of stimulus-reconstruction. This approach attempts to reconstruct an estimate of the input stimulus S using recorded neural data R via a linear reconstruction model g. For a set of N electrodes, we represent the response of electrode n at time t = 1 … T as $R(t,n)$. The reconstruction model, $g(\tau, n)$, is a function that maps $R(t,n)$ to stimulus $S(t)$ as follows:

$$\hat{\mathrm{S}}(t) = \sum_{n}\sum_{\tau} g(\tau, n) R(t-\tau, n)$$

where $\hat{\mathrm{S}}$ denotes the estimated stimulus. The function g is estimated by minimizing the mean-squared error between the actual and reconstructed stimulus

$$\min \;\; e = \sum_{t} [S(t) - \hat{\mathrm{S}}(t)]^2$$

Solving this analytically results in calculation of the normalized reverse correlation [11,12]

$$g = [RR^T]^{-1} RS^T$$

where R and S are defined as

$$R = \begin{bmatrix} r_1(0) & r_1(1) & \cdots & r_1(\tau_{max}) & \cdots & r_1(T) \\ \vdots & \vdots & & & & \vdots \\ 0 & 0 & \cdots & r_1(0) & \cdots & r_1(T-\tau_{max}) \\ \vdots & \vdots & & & & \vdots \\ r_n(0) & r_n(1) & \cdots & r_n(\tau_{max}) & \cdots & r_n(T) \\ \vdots & \vdots & & \vdots & & \vdots \\ 0 & 0 & & r_n(\tau_{max}) & & r_n(T-\tau_{max}) \end{bmatrix}$$

and

$$S = [S(0)\ S(1)\ S(2)\ \ldots\ S(T)].$$

All 128 channels of EEG data were used, since the relative contribution from each electrode is weighted by the model. Because previous research indicates that EEG activity reflects the dynamics of the speech envelope at latencies up to 250 ms post-stimulus [1], we attempted to maximize the accuracy of our speech reconstruction using EEG at time-lags $\boldsymbol{\tau}$ from 0 to 250 ms post-stimulus. As we calculated a mapping from the neural data back to the stimulus, in practice we used time-lags from -250 to 0 ms.

As there were two input speech streams (attended and unattended), we trained two decoders per trial: one where linear-regression was performed between the EEG and the attended stream alone, and another where linear-regression was performed between the EEG and the unattended stream alone. We refer to these as Attended and Unattended decoders, respectively. As each subject undertook 10 trials per condition, this resulted in 20 decoders for each subject per condition (10 Attended and 10 Unattended).

Each decoder is a two-dimensional matrix (electrode channels x time-lags). Stimulus-reconstruction is performed by convolving this matrix with EEG data. Multiple decoders can be combined by averaging these matrices together. A leave-one-out cross-validation approach was used, whereby the reconstruction for each trial was made by convolving the average of all decoders for that subject from every other trial within the same condition.

For each reconstruction, we evaluated the reconstruction-accuracy by determining a correlation coefficient (Pearson's r) between the reconstructed speech envelope and the actual attended and unattended speech envelope, which we will refer to as $r_{attended}$ and $r_{unattended}$, respectively.

E. *Combining Attended and Unattended Decoders*

Both the Attended and Unattended decoders can be used to decode attention, with varying success. We combined the two decoders into a single, more robust algorithm, using the difference between $r_{attended}$ and $r_{unattended}$ as a weighting factor.

For each condition, we plotted separate trials in 2D space, with $r_{attended}$ - $r_{unattended}$ from the Attended decoder on the horizontal axis and $r_{unattended}$ - $r_{attended}$ from the Unattended decoder on the vertical axis (Figure 2). The classifier was a straight line crossing the origin of the plot, and determined to be perpendicular to the resultant vector of all the trials within the condition. That way, if either the Attended or the Unattended decoder performed better than the other decoder for a specific condition, it would contribute more to the classification process.

Decoding accuracy was calculated as the percentage of correctly classified trials for each condition separately.

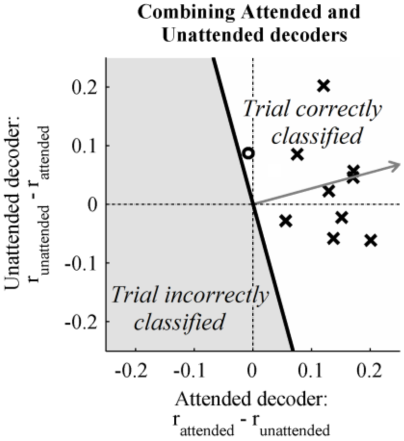

Figure 2. Illustration of the combination of the Attended and Unattended decoders. Data from 10 trials from 1 subject are plotted as crosses and a circle, with $r_{\_attended} – r_{\_unattended}$ from the Attended decoder on the horizontal axis and $r_{\_unattended} – r_{\_attended}$ from the Unattended decoder on the vertical axis. If only the Attended decoder was used, a decoding accuracy of 90% would be achieved, because for 9 out of 10 trials $r_{\_attended} > r_{\_unattended}$ (crosses). By incorporating the Unattended decoder, the classifier is changed and the 10th trial (circle) is correctly classified as well, resulting in 100% decoding accuracy.

III. Results

A. *Behavioral Results*

Our behavioral results are summarized in Table I. In condition 1, subjects were clearly compliant in the task. On average, subjects correctly answered 77.9 ± 8.6% of questions on the attended story which was statistically greater ($P < 0.001$) than the 25.3 ± 7.2% questions answered correctly on the unattended story. This is in line with previous reports on dichotic listening behavior [6,7,13]. For conditions 2, 3 and 4 subjects demonstrated that they could perform the target detection task with an accuracy in line with our design. For condition 3, the percentage of questions answered correctly was significantly lower than that for condition 1 ($P < 0.001$). This was likely due to the increased difficulty of this dual-task condition.

TABLE I. Behavioral Performance across Conditions

Attention Condition	Task Performance (Mean ± SD)		
	Attended Questions	*Unattended Questions*	*Target Detection Accuracy*
1	77.9 ± 8.6%	25.3 ± 7.2%	-
2	-	-	76.3 ± 4.7%
3	66.3 ± 10.2%	23.8 ± 9.1%	77.1 ± 3.2%
4	-	-	75.0 ± 13.0%

B. *Decoding-Accuracy*

Across all four attention conditions, we were able to decode attentional selection with a high degree of accuracy (Table II). We successfully decoded attentional selection with an average accuracy ranging from 87.9% to 90.7%, depending on the condition. A boxplot (Figure 3) shows the distribution of Decoding Accuracies for each condition. Using a non-parametric Friedman test, no significant differences were found in Decoding Accuracy across conditions ($\chi^2(3) = 0.448$, $P = 0.93$).

TABLE II. Decoding Accuracy for each Condition

	Attention condition			
	1	*2*	*3*	*4*
Mean Decoding Accuracy ± SE	90.0 ± 2.8%	88.6 ± 3.6%	90.7 ± 3.0%	87.9 ± 5.4%

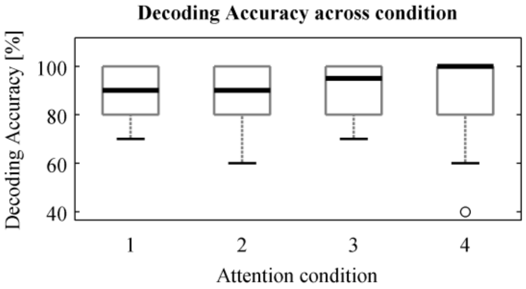

Figure 3. Boxplot of Decoding Accuracies across all four conditions. For each box, the central line is the median, the edges of the box are the 25th and 75th percentiles, the whiskers extend to the most extreme datapoints not considered to be outliers, and the outliers are plotted as individual circles.

IV. Discussion

We have successfully replicated previously published research [7] showing that it is possible to accurately decode attentional selection in a cocktail party environment from unaveraged EEG data. In particular we have shown that this is possible in a high-level task where subjects are required to attend to one of two simultaneously presented stories and to answer questions on the content of the story. Importantly, we have gone further than this previous research by showing that accurate decoding can also be performed under different task conditions: a low-level task involving detection of frequency fluctuations, a medium-level task involving detection of target words, and a dual task (low and high-level) involving both the detection of frequency fluctuations and the answering of questions on the story content.

Behaviorally, all four of our task conditions clearly demonstrated subject compliance with high accuracy on the frequency and word detection tasks, and a significant difference between the performance on answering questions to the attended and unattended stories.

In keeping with the hypothesis underlying the study, this behavior was reflected in the ability to accurately decode attention from the EEG data. The ability to decode attentionial selection is very clear, with average decoding accuracies ranging from 87.9% to 90.7%, depending on the condition. There were no statistical differences in decoding accuracy between any of the attention conditions. This was despite the fact that performance on answering questions on the attended story in condition 3 was significantly lower than that in condition 1. One may have expected a difference in decoding accuracy between these conditions given previous reports of a correlation between behavior on this task and stimulus reconstruction accuracy [7]. However, it is likely that decoding accuracy remained high because of the attention required to perform the concurrent frequency fluctuation detection task.

One other consideration when comparing decoding accuracy across conditions is the relatively small variability in the data. Future work will aim to investigate task-related differences in decoding accuracy, when smaller amounts of EEG data are used to decode attentional selection. One plausible hypothesis is that, because of the increased difficulty of condition 3, it may require increased attentional engagement on the part of the subjects and thereby result in higher decoding accuracies than either condition 1 or 2. In addition, future work will investigate which particular time-lags and electrode channels are most important for accurately decoding attention. Previous research using the same high-level task as that used here (condition 1), has shown that time-lags at around 200 ms post-stimulus are most important for decoding attention and that data on channels over temporal regions contribute most to stimulus reconstruction [7]. Based on previous literature suggesting that different attention tasks lead to effects on EEG at different latencies [14, 15], we hypothesize that decoding accuracy for our low-, medium- and high- level tasks (conditions 2, 4, and 1) will be mostly driven by EEG data at short, medium and long time-lags, respectively. For condition 3 we expect a broader distribution of attention effects across time-lags given that the task required both low- and high-level attentional engagement.

Acknowledgments

The authors thank Emma-Jane Forde for assistance with preparation of the stimuli. The authors thank Michael Crosse, Gerard Loughnane and Giovanni Di Liberto for useful discussions.

References

[1] E. C. Lalor, and J. J. Foxe, “Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution,” *European Journal of Neuroscience*, vol. 31, no. 1, pp. 189-193, Jan, 2010.

[2] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma, “Influence of Context and Behavior on Stimulus Reconstruction From Neural Activity in Primary Auditory Cortex,” *Journal of Neurophysiology*, vol. 102, no. 6, pp. 3329-3339, Dec, 2009.

[3] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, R. T. Knight, and E. F. Chang, “Reconstructing Speech from Human Auditory Cortex,” *Plos Biology*, vol. 10, no. 1, Jan, 2012.

[4] N. Ding, and J. Z. Simon, “Neural coding of continuous speech in auditory cortex during monaural and dichotic listening,” *Journal of Neurophysiology*, vol. 107, no. 1, pp. 78-89, Jan, 2012.

[5] E. M. Zion Golumbic, N. Ding, S. Bickel, P. Lakatos, C. A. Schevon, G. M. McKhann*, et al.*, “Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”,” *Neuron,* vol. 77, pp. 980-991, 2013.

[6] A. J. Power, J. J. Foxe, E. J. Forde, R. B. Reilly, and E. C. Lalor, “At what time is the cocktail party? A late locus of selective attention to natural speech,” *European Journal of Neuroscience*, vol. 35, no. 9, pp. 1497-1503, May, 2012.

[7] J. A. O’Sullivan, A. J. Power, N. Mesgarani, S. Rajaram, J. J. Foxe, M. Slaney, B. G. Shinn-Cunningham, S. A. Shamma, and E. C. Lalor, “Attentional selection in a cocktail party environment can be decoded from single-trial EEG,” *Cerebral Cortex*, doi:10.1093/cercor/bht355, 2014.

[8] C. Horton, R. Srinivasan, & M. D’Zmura, (In review) “Envelope responses in single-trial EEG indicate attended speaker in a “cocktail party”,”

[9] D. Ruggles, H. Bharadwaj, B. G. Shinn-Cunningham. “Why middle-aged listeners have trouble hearing in everyday settings,” *Current Biology*, vol. 22, no. 15, pp.1417-1422, 2012.

[10] B. N. Pasley, S. V. David, N. Mesgarani, A. Flinker, S. A. Shamma, N. E. Crone, et al., “Reconstructing Speech from Human Auditory Cortex,” *PLoS Biology*, vol. 10, p. e1001251, 2012.

[11] W. Bialek, F. Rieke, R. de Ruyter van Steveninck, and D. Warland, “Reading a neural code,” *Science*, vol. 252, pp. 1854-1857, June 28, 1991 1991.

[12] G. B. Stanley, F. F. Li, and Y. Dan, “Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus,” *Journal of Neuroscience*, vol. 19, pp. 8036-42, Sep 15 1999.

[13] C. Cherry, “Some Experiments on the Recognition of Speech, with One and with Two Ears,” *Journal of the Acoustical Society of America*, vol. 25, pp. 975-979, // 1953.

[14] E. K. Vogel, G. F. Woodman, and S. J. Luck, "Pushing around the locus of selection: evidence for the flexible-selection hypothesis," *Journal of Cognitive Neuroscience*, vol. 17, pp. 1907-22, Dec 2005.

[15] S. P. Kelly, M. Gomez-Ramirez, and J. J. Foxe, "Spatial attention modulates initial afferent activity in human primary visual cortex," *Cerebral Cortex*, vol. 18, pp. 2629-36, Nov 2008.