35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

The Effects of Attention and Visual Input on the Representation of Natural Speech in EEG*

James A. O'Sullivan, Michael J. Crosse, Alan J. Power, and Edmund C. Lalor, *Member, IEEE*

Abstract— **Traditionally, the use of electroencephalography (EEG) to study the neural processing of natural stimuli in humans has been hampered by the need to repeatedly present discrete stimuli. Progress has been made recently by the realization that cortical population activity tracks the amplitude envelope of speech stimuli. This has led to studies using linear regression methods which allow the presentation of continuous speech. One such method, known as stimulus reconstruction, has so far only been utilized in multi-electrode cortical surface recordings and magnetoencephalography (MEG). Here, in two studies, we show that such an approach is also possible with EEG, despite the poorer signal-to-noise ratio of the data. In the first study, we show that it is possible to decode attention in a naturalistic cocktail party scenario on a single trial (≈60 s) basis. In the second, we show that the representation of the envelope of auditory speech in the cortex is more robust when accompanied by visual speech. The sensitivity of this inexpensive, widely-accessible technology for the online monitoring of natural stimuli has implications for the design of future studies of the cocktail party problem and for the implementation of EEG-based brain-computer interfaces.**

I. Introduction

Analyzing electroencephalography (EEG) data on a single trial basis is extremely difficult due to its inherently poor signal-to-noise ratio and large inter-trial variability. The most widely used method of analyzing auditory processing has been to use Auditory Evoked Potentials (AEPs) which require the repeated presentation of discrete stimuli, followed by the averaging of the resulting responses [1-4]. This severely inhibits the ecological validity of experiments examining natural speech, which is continuous by nature. Recent research has shown that neuronal activity in auditory cortex tracks the amplitude envelope of natural speech [5-7]. Using this result, novel system identification approaches have been developed such as the AESPA (Auditory Evoked Spread Spectrum Analysis) [8]. This method calculates a linear forward mapping between the amplitude envelope of speech and the EEG data. As such, it has been useful in addressing important questions such as the cocktail party problem [9]. Using a mapping approach in the backward direction, several recent studies have used invasive neurophysiological recordings in animals and MEG in humans to estimate a reconstruction of the causal input speech stimulus on a subject by subject basis [10-13]. Invasive recordings are not an option for most human subjects. Also, the cost, lack of portability and availability of MEG recording facilities make population specific research somewhat difficult. Thus, it would be extremely useful if this approach could be employed in EEG which is cheaper, more widely accessible and easier to use in many specific cohorts and in brain-computer interface applications. Also, MEG and EEG detect slightly different aspects of the simultaneous electromagnetic brain activity; for example, localization of cerebral sources of brain activity may be simpler and more accurate with MEG [14]. However, EEG shows many attention-related components that are not clearly detected with MEG [15, 16]. Here, in two experiments, we use a stimulus reconstruction approach to quantify the effects of attention and visual input on the representation of natural speech in EEG. In the first study, we attempt to decode EEG on a single trial (≈60 s) basis in order to determine which speaker a subject is attending to in a natural cocktail party like scenario. In the second, we examine the multisensory effects of visual input on the representation of auditory speech in the cortex.

II. Methods

A. EEG Acquisition and Pre-processing

EEG data were recorded at 128 electrode positions, filtered online above 134 Hz and digitized at a rate of 512 Hz using a BioSemi ActiveTwo system. The data were re-referenced offline to the average of the left and right mastoid channels. The EEG data were then digitally band-pass filtered between 2 and 20 Hz.

In order to decrease the processing time required, all EEG data were downsampled by a factor of 6, whilst ensuring that aliasing and phase distortion were avoided. The amplitude envelope of the speech signals was also converted to this sampling rate to allow us to relate its dynamics to those of the EEG. Further to this, because envelope frequencies between 2 and 16 Hz contribute most to speech intelligibility [17, 18], the envelope was low-pass filtered with a corner frequency of 20 Hz.

B. Stimulus Reconstruction

Our strategy for analyzing EEG data centers 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) [10-12, 19-21]. For a set of N electrodes, we represent the response of electrode n at time $t = 1 \ldots T$ as $R(t,n)$. The reconstruction model $g(\tau,n)$ is a function that maps $R(t,n)$ to stimulus $S(t)$ as follows:

*Research supported by a grant from the United States National Science Foundation (BCS0642584) and a grant from Science Foundation Ireland (09-RFP-NES2382).

J. A. O'S. and M. J. C. contributed equally to this paper.

J. A. O'Sullivan and M. J. Crosse are with the School of Engineering and Trinity Centre for Bioengineering, Trinity College Dublin, Dublin 2, Ireland (e-mail: osullij8@tcd.ie; crossemj@tcd.ie).

A. J. Power is with the Department of Psychology, Centre for Neuroscience in Education, University of Cambridge, Cambridge, UK (e-mail: ajp218@cam.ac.uk).

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 (phone: +353-1-8961743; e-mail: edlalor@tcd.ie).

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

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

where $\hat{S}$ denotes the estimated stimulus representation. In our case, we generated an estimate of the amplitude envelope of the input stimulus using our 128 channels of EEG data. We chose to use the *optimal prior* method of stimulus reconstruction. This approach optimally minimizes the mean-squared error (MSE) of the estimated input by using known statistical structures and stimulus correlations to improve reconstruction accuracy, even if the information is not explicitly encoded in neural responses. This is in contrast to the *flat prior* method which first determines a minimum MSE forward mapping from stimulus to neural data, and then uses the inverse of this mapping to perform stimulus reconstruction given new data. As such, the optimal prior and flat prior methods are the complementary backward and forward predictions in the linear regression framework. See Mesgarani et al. [10] for a detailed description of the mathematics involved.

Previous research using stimulus reconstruction has attempted to reconstruct the full spectrogram of an input stimulus using either neurophysiological recordings in animals [10] or human electrocorticogram data [19]. Given the poorer signal to noise ratio of EEG, particularly at frequencies above ≈30 Hz, we limited ourselves to reconstructing the slow (<20 Hz) amplitude envelope of the speech stream.

C. Attention Study

1) Subjects

Forty subjects took part (mean ± standard deviation age, 27.3 ± 3.2 years; 32 male; seven left-handed). The experiment was undertaken in accordance with the Declaration of Helsinki. The Ethics Committees of the Nathan Kline Institute and 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. These data have been published previously using a different analysis approach [9].

2) Stimuli and Experimental Procedure

Each subject undertook 30 trials, each of approximately one minute in length, where they were presented with two classic works of fiction; one in the left ear and the other in the right ear, using headphones. Subjects were divided into two groups of 20 with each group being instructed to attend to the story in either the left or right ear throughout all 30 trials (i.e. approximately 1800 s of data per subject). After each trial, subjects were required to answer between four and six multiple choice questions on both stories. The questions had four possible answers. For both stories, each trial began where the story ended on the previous trial, with no repeat stimulus presentations. Stimulus amplitudes in each stream within each trial were normalized to have the same root mean squared (RMS) intensity.

3) Decoder Fitting

We wished to estimate to which of the two speakers a subject was attending based on a single trial (≈60 s). To do this, we needed to reconstruct the causal input stimulus from their recorded neural data. This was then correlated with the actual attended and unattended speech streams, and whichever had a greater correlation was chosen as our estimate of the attended stream. This reconstruction was achieved using an optimal prior linear mapping function which we will refer to hereafter as a 'decoder'. When fitting these decoders we chose to quantify the mapping from the training data to the corresponding attended speech. In addition, we wished to test our decoding ability on data that were not used to fit the decoder. As such, for each subject, we fit thirty decoders (one for each trial) using the other 29 trials as training data (leave-one-out cross validation).

D. Audiovisual Study

1) Subjects

Eight right-handed subjects (4 females; mean age, 24.5 years; range, 20–30 years) took part in this experiment, all of whom were free of neurological diseases and had normal hearing and normal or corrected-to-normal vision.

2) Stimuli and Experimental Procedure

Video recordings of natural speech were used to preserve ecological validity. The videos were drawn from a collection of YouTube movies featuring Barack Obama (uploaded by barackobama.com). Seven videos were each truncated to 120s using VideoPad Video Editor and rendered into 640 × 360 pixel movies with a digitization rate of 29.97 frames/s. The stereo soundtracks were digitized at 44.1 kHz with 16-bit resolution. Stimulus presentation was conducted using Presentation software and delivered using a 19" CRT monitor and Sennheiser HD650 headphones. Each video was presented three times, once for each of the following conditions; audio-only (A), visual-only (V) and audiovisual (AV). The order in which the three conditions were presented was randomized. Subjects were instructed to maintain visual fixation for the duration of each run on the speaker's mouth for V and AV, and on a grey crosshair for A. Before each movie began, the first frame was displayed for three seconds. Subjects were positioned 70 cm from the monitor and instructed to keep eye-blinking and all other motor activity to a minimum.

3) Quantification of Encoding

We wanted to compare the cortical representation of natural speech during A and AV speech. To quantify how well this information was encoded during each condition, we reconstructed the amplitude envelope of the input signal from the EEG data and correlated it with the envelope of the actual speech signal. Given that we had seven runs per condition we trained and tested our decoders using leave-one-out cross validation. The decoders were fitted using EEG data from only the channels anterior of the midline as we wanted to assess the effect of visual speech on auditory processing without the confound of including extra visual cortical activity. To examine if there was any significant unimodal contribution from visual speech at these electrodes, the performance values (Pearson's *r*) of the V decoder were compared to zero. The performance values of the A and AV decoders were then compared against each other on a matched trial basis.

III. Results

A. Effects of Attention

1) Behavior

As was reported previously, our behavioral results clearly showed that subjects were compliant in terms of the attention task [9]. On average, subjects correctly answered 80.4 ± 7.3 % of questions on the attended story and 27.1 ± 7.0 % on the unattended story, which, consistent with previous reports on dichotic listening behavior, was not statistically greater than chance ($P = 0.77$).

2) Decoding Attention

We wanted to avoid introducing any attentional, stimulus-specific or subject-specific bias in terms of the choice of fitted decoder. As such, for each subject and each trial, we estimated the causal speech stimulus using the attended decoders from every other subject. For example, for trial 1 of subject 1, we used every other subject's attended decoder (trained on trials 2 to 30) to reconstruct the stimulus from subject 1's EEG data. This resulted in 39 reconstructed stimulus envelopes, which were summed together and then correlated with both the attended and unattended streams. Whichever speech stream had a greater correlation (Pearson's r) with our estimate was chosen as the attended stream. Importantly, this approach minimizes any stimulus bias in terms of the decoder fitting because 19 subjects were attending one stream, while 20 were attending the other. Using this approach, we were able to significantly determine which speaker was being attended to for 35 of our 40 subjects (mean 77 %, min 43 %, max 100 %). Decoding accuracy was deemed significant at 63.33% based on a binomial test at the 5% significance level. (Fig. 1).

B. Effects of Visual Input

To analyze the contribution from visual speech, we took the Pearson correlation coefficients from the V decoder and compared the mean of their distribution to zero using a one-sampled t-test. The difference was not significant ($P = 0.73$), indicating that there was no significant unimodal contribution from visual speech at the electrodes on the anterior half of the head.

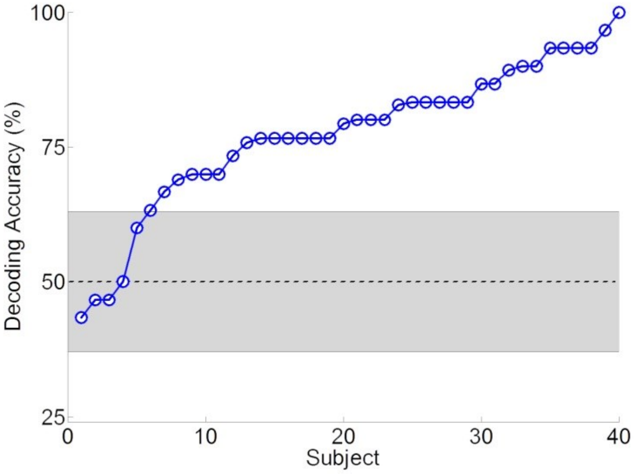

Figure 1. Accuracy in determining which speech stream a subject was attending to. The grey area indicates the significant level of decoding accuracy (63%) based on a binomial test at the 5% significance level.

Figure 2 is a scatter plot of the single trial performance values of the A and AV decoders. The correlation values from each decoder are matched according to subject and trial number, and plotted against each other. The different colors are representative of whether the correlation between the reconstruction and the original envelope was significant for an alpha level of 0.05 (Blue = both trials significant; red = only AV trial significant; green = only A trial significant; cyan = neither trials significant). A total of 56 trials are plotted in total (8 subjects, 7 runs). Out of these, 94.6% of the AV trials were found to be significant whereas only 78.6% of the A trials were significant. Furthermore, the AV correlations (r values) scored higher than the A values in 66.1% of the 56 trials.

To test the significance of this result, we used a paired t-test to compare mean of the A and AV r-distributions. This difference was found to be significant ($P < 0.01$), suggesting that the AV decoder performed better than the A decoder. We interpret this as evidence that the amplitude envelope of auditory speech was represented more robustly in the cortex during audiovisual speech than it was during audio-only speech.

IV. Discussion

In the first study, we show for the first time the ability to decode EEG data on a single trial basis using the method of stimulus reconstruction. We were able to determine to which of two speakers a subject was attending to for 35 out of our 40 subjects. However, the percentage of trials accurately predicted was not as high as hoped, with a mean of 77 %. Much of this could be attributed to the fact that the data were originally collected for a different study [9], and as such, we were unable to tailor the experiment for our specific needs. Most crucial is the fact that none of the subjects listened to single speaker speech stimuli. Ideally, we would train our decoders on this instead, akin to stimulus reconstruction studies carried out previously [10, 19]. Instead, our data were acquired during the dichotic listening task, thus biasing our decoders to the attended stream. If we had used each subject's decoder to reconstruct their own stimulus for each trial, we would have

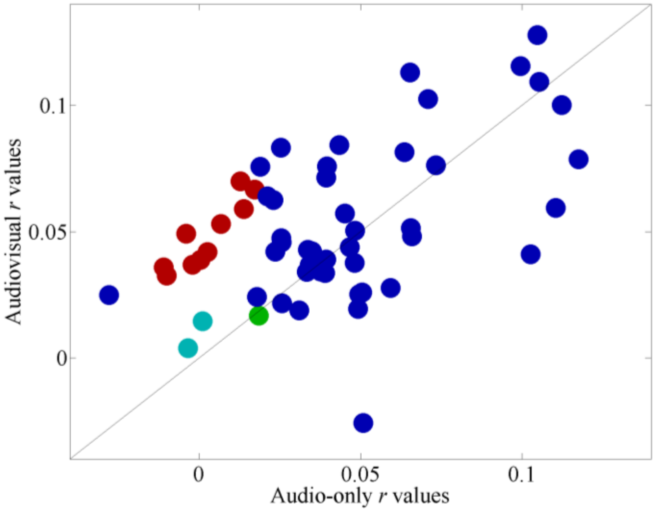

Figure 2. Comparison of the A and AV decoder performance at accurately reconstructing the input stimulus. Out of the 56 trials, the AV decoder performed better on 66.1 % of the trials.

significantly increased our percentage accuracy in predicting attention. However, as mentioned previously, this would have been biased. The fact that we were able to decode attention at all using the average decoder technique is therefore encouraging, indicating that with better data, a much greater accuracy might be achieved.

In terms of multimodal inputs, we used this method to quantify how well auditory speech was represented in the cortex for different modes of speech. From just 8 subjects, each with 14 minutes of data, we found a significant improvement in the encoding of audiovisual speech compared with audio-only speech. By restricting our analysis to frontal electrodes only, we were able to limit contributions from visual cortex, which would contribute more to audiovisual EEG data than during the audio-only condition.

The advantage of the stimulus reconstruction method over other approaches is in its ability to analyze all electrodes simultaneously, thus using all of the available information that is spread across the scalp at each instant in time. It does this by finding a multivariate linear filter that incorporates the channel covariance structure in the estimation of the impulse response, resulting in a significant quantitative improvement in the input-output mapping.

The speed with which this can be done suggests a possible role for this method in the future design of Brain Computer Interfaces (BCIs). With sufficient data and training, it is feasible that a subject-specific decoder could be created, which would have enough sensitivity to accurately decode attention at latencies far shorter than 60 seconds.

The disadvantage of this method however, is the inability to use more traditional methods to localize the generators of scalp recorded activity, and the lack of an illustrative impulse response function in which to analyze the amplitude and latencies of various brain functions. Therefore, it would be advantageous to accompany this approach with a forward mapping method, such as the AESPA, in order to obtain a fuller understanding of the acquired data.

Acknowledgment

The authors acknowledge the organizers of the 2012 Neuromorphic Cognition Engineering Workshop in Telluride, Colorado, USA, where much of this work was conceived.

References

[1] A. Bidet-Caulet, C. Fischer, J. Besle, P. E. Aguera, M. H. Giard, and O. Bertrand, "Effects of selective attention on the electrophysiological representation of concurrent sounds in the human auditory cortex," *Journal of Neuroscience,* vol. 27, no. 35, pp. 9252-9261, Aug, 2007.

[2] T. W. Picton, and S. A. Hillyard, "Human Auditory Evoked-Potentials .2. Effects of Attention," *Electroencephalography and Clinical Neurophysiology,* vol. 36, no. 2, pp. 191-199, 1974.

[3] T. W. Picton, S. A. Hillyard, H. I. Krausz, and R. Galambos, "Human Auditorry Evoked-Potentials .1. Evaluation of Components," *Electroencephalography and Clinical Neurophysiology,* vol. 36, no. 2, pp. 179-190, 1974.

[4] J. S. Snyder, C. Alain, and T. W. Picton, "Effects of attention on neuroelectric correlates of auditory stream segregation," *Journal of Cognitive Neuroscience,* vol. 18, no. 1, pp. 1-13, Jan, 2006.

[5] K. V. Nourski, R. A. Reale, H. Oya, H. Kawasaki, C. K. Kovach, H. M. Chen, M. A. Howard, and J. F. Brugge, "Temporal Envelope of Time-Compressed Speech Represented in the Human Auditory Cortex," *Journal of Neuroscience,* vol. 29, no. 49, pp. 15564-15574, Dec, 2009.

[6] R. E. Millman, G. Prendergast, M. Hymers, and G. G. R. Green, "Representations of the temporal envelope of sounds in human auditory cortex: Can the results from invasive intracortical "depth" electrode recordings be replicated using non-invasive MEG "virtual electrodes"?," *NeuroImage,* vol. 64, no. 0, pp. 185-196, 1/1/, 2013.

[7] S. J. Aiken, and T. W. Picton, "Human cortical responses to the speech envelope," *Ear and Hearing,* vol. 29, no. 2, pp. 139-157, Apr, 2008.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[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, no. 18, pp. 8036-8042, Sep, 1999.

[13] 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.

[14] M. Hämäläinen, R. Hari, R. J. Ilmoniemi, J. Knuutila, and O. V. Lounasmaa, "Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain," *Reviews of modern Physics,* vol. 65, no. 2, pp. 413, 1993.

[15] R. Näätänen, *Attention and brain function*: Lawrence Erlbaum Associates, Inc, 1992.

[16] S. Kahkonen, J. Ahveninen, I. P. Jaaskelainen, S. Kaakkola, R. Naatanen, J. Huttunen, and E. Pekkonen, "Effects of Haloperidol on Selective Attention A Combined Whole-Head MEG and High-Resolution EEG Study," *Neuropsychopharmacology,* vol. 25, no. 4, pp. 498-504, 10//print, 2001.

[17] R. Drullman, J. M. Festen, and R. Plomp, "Effect of Reducing Slow Temporal Modulations on Speech Reception," *Journal of the Acoustical Society of America,* vol. 95, no. 5, pp. 2670-2680, May, 1994.

[18] R. van der Horst, A. R. Leeuw, and W. A. Dreschler, "Importance of temporal-envelope cues in consonant recognition," *Journal of the Acoustical Society of America,* vol. 105, no. 3, pp. 1801-1809, Mar, 1999.

[19] N. Mesgarani, and E. F. Chang, "Selective cortical representation of attended speaker in multi-talker speech perception," *Nature,* vol. 485, no. 7397, pp. 233-U118, May, 2012.

[20] F. Rieke, D. A. Bodnar, and W. Bialek, "Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents," *Proceedings of the Royal Society B-Biological Sciences,* vol. 262, no. 1365, pp. 259-265, Dec, 1995.

[21] W. Bialek, F. Rieke, R. R. D. Vansteveninck, and D. Warland, "Reading a Neural Code," *Science,* vol. 252, no. 5014, pp. 1854-1857, Jun, 1991.