Using EEG to Predict Speech Intelligibility

Ivan Iotzov (iiotzov@gradcenter.cuny.edu)
Department of Biomedical Engineering, 160 Convent Avenue
New York, NY 10030, U.S.A.

Lucas C. Parra (parra@ccny.cuny.edu)
Department of Biomedical Engineering, 160 Convent Avenue
New York, NY 10030, U.S.A.

Abstract

Speech signals have the ability to entrain brain activity to the rapid fluctuations found in speech sounds. This entrainment can be measured using electroencephalographic (EEG) recordings and is strong enough to allow discrimination between attended and unattended speech sources. In this study, we investigated whether these entrainment responses can be used to measure how intelligible a speech signal is to a subject. We hypothesized that when intelligibility is degraded, attention wanes and the stimulus-response correlation will decrease. To test this, we measured a listener's ability to detect words in noisy, natural speech while recording brain activity using EEG. We altered intelligibility by presenting congruent or incongruent video of the speaker along with their speech. For almost all subjects, word detection performance improved in the congruent condition and this improvement coincided with an increase in stimulus-response correlation. We conclude that simultaneous recordings of perceived sound and EEG activity may represent a practical tool to assess speech intelligibility, specifically in the context of hearing aid devices.

Keywords: EEG; stimulus-response correlation; speech comprehension; speech intelligibility

Introduction

While listening to sounds, brain activity follows the fast fluctuations of the acoustic stimulus (Ding & Simon, 2014; Haegens & Zion Golumbic, 2018). This effect can also be observed in subjects listening to speech signals, where EEG and MEG signals have been shown to correlate with fluctuations in signal amplitude and spectral content (Luo & Poeppel, 2007; Horton, Srinivasan, & D'Zmura, 2014). This stimulus-driven brain activity has been linked to attention (Zion Golumbic et al., 2013; O'Sullivan et al., 2015), in particular in multi-speaker scenarios where these correlations are thought to reflect the listener's ability to follow the attended speech stream. We see similar findings in studies of noisy speech, where the clean speech envelope still shows correlation with the brain response (Ding & Simon, 2013; Vanthornhout, Decruy, Wouters, Simon, & Francart, 2017). We attribute this speech tracking phenomenon to an exogenous stimulus-driven process due to the consistent responses elicited by a speech stimulus across subjects (Ki, Kelly, & Parra, 2016; Cohen, Henin, & Parra, 2017).

A consistent confound in previous research has been that speech intelligibility is modulated by altering various properties of the stimulus. This makes it difficult to determine whether changes in speech tracking are caused by genuine changes in auditory processing or merely a reflection of the altered stimulus. We have attempted to mitigate this confound by keeping the stimulus unchanged and modulating speech intelligibility through visual cues. The audio presented to subjects in our two experimental conditions is identical and we modulate intelligibility by presenting visuals that are either congruent (i.e. the mouth of the speaker and the heard audio align) or incongruent.

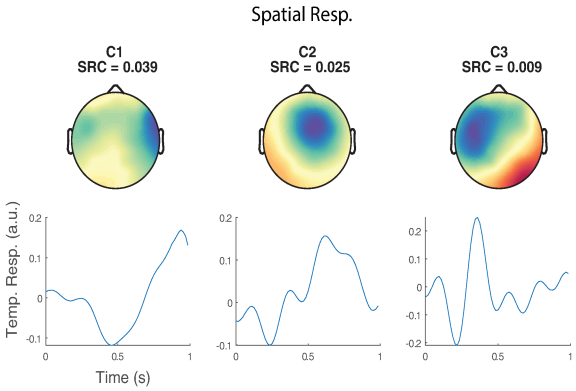

Figure 1: Visualization of the CCA model showing spatial (top) and temporal (bottom) EEG response functions. SRC values shown are an average over all subjects and conditions.

Methods

Stimulus Presentation and Behavioral Measures

The stimuli used in this experiment were previously used in other speech tracking experiments with EEG (Crosse, Butler, & Lalor, 2015; Crosse, Di Liberto, & Lalor, 2016). The stimuli consist of 120 audiovisual talking head clips of President Barack Obama discussing the Affordable Care Act, each 60s long. In total, there were four stimulus conditions combining -9 dB/ -6 dB noise and congruent/incongruent visuals in a 2 x 2 design. Subjects were presented with 30 stimuli in each of the four conditions, for a total of 120 trials that took place over 2 experimental sessions.

Before each 60s trial, subjects were presented with a 'target word' and were instructed to press a button whenever they heard the target. Target words were selected so that each word occurs the same number of times in all four conditions. Responses within 1.5s of word presentation were coded as correct detection and any responses outside of this window were coded as false alarms. Correct detections can be reported either relative to the total number of target words (detection) or relative to the total number of responses (precision). In order to capture both of these properties, we report behavioral performance using the F1 score, which is the harmonic mean of detection and precision.

EEG and Stimulus Processing

EEG was recorded from 20 healthy subjects with normal hearing using a BioSemi Active II amplifier with 64 electrodes, in addition to 6 electrooculogram (EOG) electrodes. The EEG was sampled at 512 Hz and later downsampled to the framerate of the video presentation (30 Hz). Before correlation to the stimulus, the EEG signal was preprocessed as follows: the initial value was subtracted from the data to remove the DC offset. A high-pass 5th order Butterworth filter with a cutoff of 0.5 Hz was applied. The signal from the 6 EOG electrodes was regressed out using a least-squares algorithm. Finally, artifacts and channels with recording quality issues were removed.

The steps used to calculate the stimulus amplitude envelope are as follows: the sound amplitude envelope is calculated as the absolute value of the analytic signal after a Hilbert transform fo the raw mono sound signal at its original sampling rate of 48 kHz. The result is then downsampled to the framerate of the video (30 Hz) and z-scored. Then, a Toeplitz matrix with 30 columns is constructed to capture up to a 1s delay.

Stimulus-Response Correlation (SRC)

The models used in much of the speech tracking literature are typically encoding or decoding models. The encoding approach uses various features of the stimulus to predict the brain response, while the decoding approach works from the brain response and attempts to reconstruct the stimulus. Here, we used a hybrid encoding and decoding approach (Dmochowski, Ki, DeGuzman, Sajda, & Parra, 2017). The model attempts to maximize the correlation between the encoded stimulus $\hat{u}(t)$ and the decoded response $\hat{v}(t)$. The two signals are defined as:

$$\hat{u}(t) = h(t) * s(t)$$

$$\hat{v}(t) = \sum_i w_i r_i(t)$$

where $s(t)$ is the sound amplitude envelope at time t, $h(t)$ is the encoding filter applied to the stimulus signal, * represents convolution, w_i are the weights applied to the neural response, and $r_i(t)$ is the EEG signal value at time t at electrode i. The model parameters $h(t)$ and w_i are found using canonical correlation analysis (CCA). CCA looks to maximize the correlation between the encoded stimulus and the decoded brain response by computing several components that each capture some portion of the correlated signal. The stimulus-response correlation (SRC) that we report is the sum of the correlation of $\hat{u}(t)$ and $\hat{v}(t)$ for the first three components. For a more full discussion of the method please see (Dmochowski et al., 2017).

The model is trained using data from all subjects in all conditions. The resulting spatial and temporal response functions can be seen in Fig. 1. The model is then used to separately compute SRC for each subject in each of the four conditions. Statistical significance of SRC values is estimated by comparing values to those produced by correlating 1000 sets of circularly shuffled EEG data with the stimuli using the same procedure as the normal EEG data.

Results

Congruent Visual Speech Increases Word Detection

Subject performance on the behavioral task was quantified in terms of the F1 score (described in Methods above). We performed a two-way repeated measures ANOVA on these scores, considering factors of noise level and congruency. We found very large effects on behavior of both noise level $[F(1,19) = 284.9, p = 6x10^{-13}]$ and congruency $[F(1,19) = 917.4, p = 1x10^{-17}]$, as well as for the interaction between noise level and congruency $[F(1,19) = 118.1, p = 1x10^{-9}]$.

These results are in line with our expectations, as the effects of visual information on the perception of noisy speech as well characterized (Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007). Additionally, the results indicate that we successfully manipulated the intelligibility of the audiovisual speech without altering the auditory portion of the stimulus. For a visual representation of these results, please see Fig. 2.

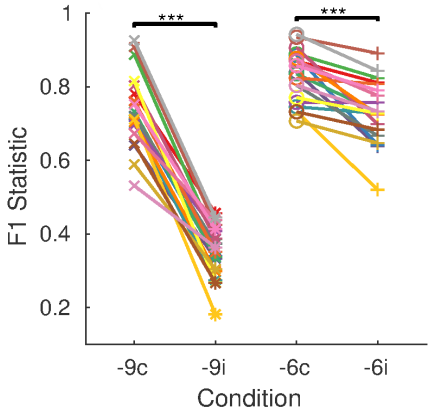

Figure 2: Performance on the behavioral word detection task for each subject in each condition (-6dB congruent, -6dB incongruent, -9dB congruent, -9dB incongruent) reported as F1 statistic.

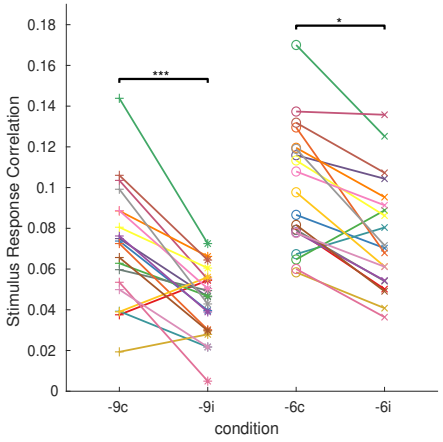

Figure 3: SRC scores for each subject in all conditions. Performance within each noise condition is connected with a line.

Congruent Visual Speech Increases Stimulus-Response Correlation

Using the hybrid approach described above, we used CCA to relate the auditory amplitude envelope at various time delays to the EEG signal across various electrodes. This process results in a number of correlated components, of which we only examine the three most highly correlated. The resulting spatial and temporal response functions can be seen in Fig. 1. The correlation values found are small, but are significant given the large amount of data collected *(r = 0.039, 0.025, 0.009, respectively for first 3 SRC components, p ¡ 0.001 using shuffle statistics)*. Overall SRC values are reported as the sum of these first three components, calculated separately for each subject and condition (See Fig. 3).

We performed a two-way repeated measures ANOVA on these SRC values and found a very significant effect for both noise level $[F(1,19) = 44.54, p = 2x10^{-6}]$ and congruency $[F(1,19) = 109.4, p = 2x10^{-9}]$. In contrast with the behavioral results, we did not find a significant interaction between the two effects, suggesting that the effect of congruency on SRC is equally strong at both noise levels. To test the efficacy of correlating the noisy speech signal to the EEG responses, we repeated the analysis using the envelope of the clean speech and the envelope of the noise alone, instead of the mixed noisy speech as above.

The results of this secondary analysis showed increased SRC values for the clean speech *(r = 0.05, 0.03, 0.01, respectively for first 3 components, p ¡ 0.001 using shuffle statistics)* compared to the noisy speech *(r = 0.039, 0.025, 0.009, respectively for first 3 SRC components)*, as well as strong effects of noise level $[F(1,19) = 52.53, p = 7x10^{-7}]$ and congruency $[F(1,19) = 54.43, p = 5x10^{-7}]$. The SRC for the noise-only envelope was much weaker *(r = 0.01, 0.0047, 0.0019, p ¡ 0.001, p = 0.001, p = 0.14, respectively for first 3 components, p-values calculated through shuffle statistics)*, and we did not find a significant effect for noise level, nor for congruency. This result conflicts with our initial hypothesis, and suggests that listeners were able to extract speech from the noisy environment and were not 'tracking' the noise in the same way as the speech. However, considering the small differences between the noisy speech and clean speech results, this result does validate our approach using the noisy speech, which is the only input available in real-world scenarios.

We visually relate behavioral task performance to SRC in Fig. 4a. For the majority of subjects, an increase in SRC coincides with increased performance on the behavioral task (indicated by positive slope in Fig. 4a). Fig. 4b is a different view of the same result, showing that the change in behavioral performance and SRC have the same sign for most subjects *(p = 0.0026, p = 0.0004 for -9dB and -6dB respectively, sign test)*.

Conclusion

Our main finding, that improved speech perception coincides with an increase in stimulus-response correlation, is consistent with much of the previous research into speech tracking (Ding & Simon, 2013; Peelle, Gross, & Davis, 2013; Vanthornhout et al., 2017; Crosse et al., 2015). We extend this previous work by using incongruent audiovisual speech as a control condition in noise, and demonstrated that this effect is correlated with gains in behavioral performance measures within individual subjects.

The motivation for our work was to find an objective assessment for the intelligibility of speech in the context of hearing aids. There are several novel contributions that our work offers in this context. First, we have shown the reliable ability to predict intelligibility within individual subjects, which is critical in tuning a hearing aid for an individual. Second, because we did not alter the auditory stimulus, confounds surrounding the alteration of the auditory stimulus have been eliminated. The changes in SRC observed here are thus likely due to the processing of the auditory stimulus by the subject, and not a result of changes to the stimulus itself. Finally, in contrast to previous work, our SRC measure does not require access to the clean speech in order to make predictions about intelligibility. In practical scenarios, access to the clean speech is impossible, thus making an approach that can work within noisy contexts essential.

References

Cohen, S. S., Henin, S., & Parra, L. C. (2017). Engaging narratives evoke similar neural activity and lead to similar time perception. *Scientific Reports*, *7*(1), 1–10.

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. *Journal of Neuroscience*, *35*(42), 14195–14204.

Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration. *Journal of Neuroscience*, *36*(38), 9888–9895.

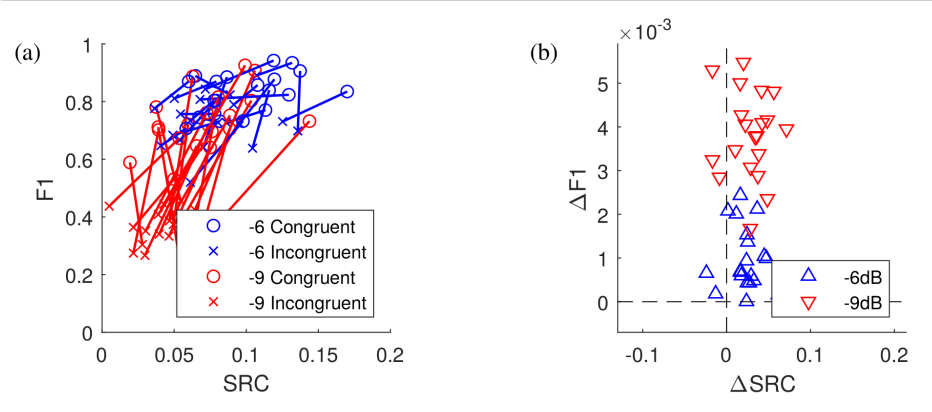

Figure 4: (a) Comparison between behavioral word detection performance and SRC for each subject in all conditions. (b) Difference between congruent and incongruent conditions within each noise level for each subject. Points in the first quadrant indicate that gains in one measure coincide with gains in the other.

Ding, N., & Simon, J. Z. (2013). Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical Representation of Speech. *Journal of Neuroscience*, *33*(13), 5728–5735.

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*.

Dmochowski, J. P., Ki, J. J., DeGuzman, P., Sajda, P., & Parra, L. C. (2017). Extracting multidimensional stimulus-response correlations using hybrid encoding-decoding of neural activity. *NeuroImage*(May), 1–13.

Haegens, S., & Zion Golumbic, E. (2018). Rhythmic facilitation of sensory processing: A critical review. *Neuroscience and Biobehavioral Reviews*, *86*(December 2017), 150–165.

Horton, C., Srinivasan, R., & D'Zmura, M. (2014). Envelope responses in single-trial EEG indicate attended speaker in a cocktail party'. *Journal of Neural Engineering*, *11*(4), 046015.

Ki, J. J., Kelly, S. P., & Parra, L. C. (2016). Attention Strongly Modulates Reliability of Neural Responses to Naturalistic Narrative Stimuli. *Journal of Neuroscience*, *36*(10), 3092–3101.

Luo, H., & Poeppel, D. (2007). Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. *Neuron*, *54*(6), 1001–1010.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., ... Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706.

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, *23*(6), 1378–1387.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147–1153.

Vanthornhout, J., Decruy, L., Wouters, J., Simon, J., & Francart, T. (2017). Speech intelligibility predicted from neural entrainment of the speech envelope. *Journal of the Association for Research in Otolaryngology*(637424).

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., ... Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, *77*(5), 980–991.