EEG can predict speech intelligibility

Ivan Iotzov, Lucas C. Parra
Biomedical Engineering, City College of New York

Abstract

Speech signals have a remarkable ability to entrain brain activity to the rapid fluctuations of speech sounds. For instance, one can readily measure a correlation of the sound amplitude with the evoked responses of the electroencephalogram (EEG), and the strength of this correlation is indicative of whether the listener is attending to the speech. In this study we asked whether this stimulus-response correlation is also predictive of speech intelligibility. We hypothesized that when a listener fails to understand the speech in adverse hearing conditions, attention wanes and stimulus-response correlation also drops. To test this, we measure a listener's ability to detect words in noisy speech while recording their brain activity using EEG. We alter intelligibility without changing the acoustic stimulus by pairing it with congruent and incongruent visual speech. For almost all subjects we found that an improvement in speech detection coincided with an increase in correlation between the noisy speech and the EEG measured over a period of 30 minutes. We conclude that simultaneous recordings of the perceived sound and the corresponding EEG response may be a practical tool to assess speech intelligibility in the context of hearing aids.

Introduction

When listening to sounds, brain activity follows the the fast fluctuations of the acoustic stimulus (<1s; Ding and Simon 2014, Haegens 2018). This is particularly true for speech, where both amplitude and spectral fluctuations of the sound have been shown to correlate with fluctuations in the EEG and MEG (Luo 2007, Nourski 2009, Ding 2015, Doelling 2014, Horton, 2014). This stimulus-related brain activity has been linked to attention (Ding 2012, Dmochowski 2017), in particular in a “cocktail party” scenario (Zion Golumbic 2013, O’Sullivan 2015, O’Sullivan 2017, Horton 2014, Power 2012), where the correlations are thought to reflect the ability of a listener to follow the attended speech. Correlation with the stimulus is sometimes referred to as ‘speech tracking’. Others refer to it as ‘neural entrainment’, suggesting that the stimulus entrains endogenous ongoing neural activity (Peele 2013, Ding 2015). Correlation of the neural signals has also been linked to an engagement with the stimulus (Dmochowski 2017) and the perception of phonemes (Kösem 2017). We attribute this phenomenon to an exogenous stimulus-driven process, which is illustrated by the consistent responses elicited across subjects by the same auditory stimulus (Ki 2016, Cohen 2017).

Interestingly, for speech in noise, brain responses remain correlated to the amplitude envelope of the clean speech (Ding and Simon 2013, Vanthornhout 2017), suggesting that neural processing is resistant to this acoustic degradation. This may be a reflection of the human ability to understand speech despite significant distortions. In fact, a number of studies have found a significant correlation between this neural correlate and the comprehension of noisy speech (Luo and Poeppel 2007, Ding and Simon 2013, Peele 2013, Kong 2015, Vanthornhout 2017). However, these studies generally changed speech intelligibility by altering various properties of the acoustic stimulus. Therefore, changes in speech-tracking could result from changes in low-level stimulus properties, rather than genuine changes in auditory processing of speech. In this study, we attempt to control for this confound by keeping the acoustic signal unchanged. Instead, we change intelligibility using congruent or incongruent audiovisual speech in noise (Park 2016).

Our experiment asks whether correlation of brain signals to the perceived sound is indicative of speech intelligibility in realistic scenarios. We hypothesize that when a listener fails to understand the speech in adverse hearing conditions, attention wanes and stimulus-response correlation drops, following the well established link between attention and speech tracking (Zion-Golumbic and Schroeder 2012, Ding and Simon 2012 O'Sullivan 2015). Here, intelligibility is measured objectively using a word detection task following Crosse (2015, 2016). To emulate challenging hearing conditions, we utilize speech in noise with a matching sound spectrum. Rather than correlating EEG responses to the original clean speech, as in previous speech tracking research, we correlate brain responses to the amplitude of the noisy speech. This enables us to predict intelligibility from the EEG in realistic scenarios where a clean version of a stimulus may not be available. This is particularly relevant in the context of hearing aids, where it is impossible to obtain clean, noise-free audio outside of a tightly controlled laboratory setting. Despite significant noise, we find that the stimulus-response correlations are predictive of whether subjects can detect words buried in that noise. This finding may enable next-generation hearing aids to be automatically optimized for speech intelligibility.

Methods

Subject Recruitment

Subjects were recruited from healthy volunteers using advertisements posted around the City College of New York campus, as well as online on the lab website. Subjects were screened for normal hearing based on self-report and were generally healthy. All recruitment materials as well as the screening questionnaire received approval from the CUNY Institutional Review Board. 20 subjects were recruited for this study (11 female, 16 right-handed, median age = 20).

Stimuli and Stimulus Presentation

The stimuli we use in this study were previously used in speech tracking experiments with EEG (Crosse 2015, Crosse 2016). Each condition is defined by a different level of added noise and whether the visual speech is congruent or incongruent with the auditory speech, i.e. the mouth movement of the speaker corresponds to the auditory speech, or does not. Stationary, colored noise was added to the signal with a signal-to-noise ratio of -9 dB and -6 dB. This noise was matched in spectrum to the speech signal for each 60 second segment (using 50 linear-prediction coefficients). In total there were four stimulus conditions combining -9 dB/-6 dB noise and congruent/incongruent visuals in a 2x2 design.

The stimuli consisted of 60 audiovisual talking head clips of President Barack Obama, each 60 seconds long. The experiment was divided into two sessions (separated by at least a week) with all four conditions included in each session. All 60 unique stimuli were presented in each session with 15 words per condition. Over the two sessions this yielded 120 presentations with 30 unique words per condition. Subjects were presented with a target word before each video. They were instructed to detect the target word while listening and watching the speaker, and to press a button as quickly as possible as soon as they heard the target word, which could appear more than once in the 60 second segment. Target words were selected so that each word was presented equally often in the four different conditions. Additionally, the stimuli were randomized over the two sessions to avoid repetition within session. Subjects were given feedback on their detection performance after each 60 second segment. We converged on this protocol after pilot experiments with a separate cohort of subjects in -9 dB, -6 dB, and -3 dB noise conditions. In these experiments we established that the behavioral detection performance at -3 dB was not significantly different between the congruent and incongruent conditions (N=24, p=0.093, See supplementary Figure S1). It was also found that behavioral word detection performance is somewhat erratic when no performance feedback is given during the task as subjects lose interest and lack motivation for the rather difficult task.

Behavioral detection performance

If the subject indicated that they heard the target word within 1.5 seconds of word presentation, this was coded as a correct detection. If the subject did not indicate they heard the target word, this was coded as a miss. Any errant response outside this 1.5 seconds window was coded as a false alarm. Correct detections can be reported relative to the total number of target words or relative to the total number of responses. In the literature on detection theory, the former is typically referred to as recall and the latter as precision. Because the term 'recall' may be confusing in the context of this paper, we will use the term 'detection' instead. Increasing the number of responses at random will tend to increase detection but reduce precision. A metric that captures both (as a their harmonic mean) is the F1 score, with a score of 1 indicating perfect precision and detection. The F1 score will be used here as the primary outcome measure because it captures the tradeoff between precision and detection. We also confirmed after data was collected that 1.5 seconds is a reasonable cutoff for detection based on the F1 score as a function of this time window (see Supplementary Figure S2).

EEG Recording

EEG was recorded using a BioSemi Active II amplifier with 64 electrodes arranged according to the International 10-20 System. Additionally, recordings were taken from six electrooculogram (EOG) electrodes, with three placed around each eye. The EEG was sampled at 512 Hz and saved using the BioSemi ActiView software. All recordings were performed inside an isolated chamber to prevent noise interference. Stimuli were presented to subjects using Sony MDR7506 headphones at 70 dB SPL, which were calibrated using pre-whitening with ten linear-prediction coefficients measured with a KEMAR (Knowles head & torso simulator) connected to a B&K type 2231 sound level meter. All stimuli were presented on a BenQ FP783 monitor and audiovisual playback was controlled by the PsychToolbox package (Kleiner 2007) for MATLAB (Mathworks, Natick, MA).

Stimulus Response Correlation (SRC)

The conventional approach in the speech tracking literature is to correlate the amplitude envelope of clean speech, $s(t)$, with the response in each EEG channel $r_i(t)$ (e.g. Zion Golumbic 2013). This models the brain responses as a linear "encoding" of the speech amplitude. Alternatively, EEG response is linearly filtered and combined across electrodes to best reconstruct the speech amplitude (e.g. O'Sullivan 2015). This "decoding" model of the stimulus is then correlated to the amplitude envelope of the clean speech. In both instances, model performance is measured as correlation, either with the stimulus $s(t)$ (decoding) or the response $r_i(t)$ (encoding). Here we used a hybrid encoding and decoding approach (Dmochowski 2017), by building a model that maximizes the correlation between the encoded stimulus $\hat{u}(t)$ and the decoded response $\hat{v}(t)$. These two signals are defined as:

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

$$\hat{v}(t) = \sum_i w_i r_i(t) \tag{1}$$

where $s(t)$ represents, in this case, the sound amplitude envelope at time *t*, $h(t)$ is the encoding filter being applied to the stimulus signal, $*$ represents a convolution, w_i are the weights applied to the neural response, and $r_i(t)$ is the neural response at time *t* in electrode *i*. We use canonical correlation analysis (CCA) to find the best model parameters $h(t)$ and w_i. CCA does this by maximizing the correlation between the encoded stimulus and decoded response. CCA computes several components, each capturing a portion of the correlated signal. The stimulus-response correlation (SRC) reported here is the sum of the correlation of $\hat{u}(t)$ and $\hat{v}(t)$ for the first three components. In practice, CCA is applied to two matrices, one for the stimulus feature (sound amplitude), the other for the brain response (EEG evoked response). Full discussion of the method and how to compute the spatial and temporal response functions of each component can be found in (Dmochowski, 2017).

Here, the decoding-encoding model is trained with CCA on all conditions from all subjects (corresponding matrices are simply concatenated in time). We use a version of CCA that is regularized with PCA, where we keep ten dimensions (out of 30 and 64 for stimulus and response respectively). These correspond to the 30 time lags used for the stimulus (corresponding to 1 second), and the 64 electrodes for the response. Note that CCA provides multiple dimensions (components) that are correlated in time between the two data matrices. The corresponding spatial and temporal response functions are shown in Figure 3a and 3b for the first three components. Using the resulting model, SRC is measured separately for each subject in each of the four stimulus conditions based on the correlation between $\hat{u}(t)$ and $\hat{v}(t)$ for the segments corresponding to each subject and condition (30 segments x 60 seconds for each condition). The result is a SRC measure for each subject in each of the four stimulus conditions.

Sound amplitude envelope computation

The sound amplitude $s(t)$ is calculated as the absolute value of the analytic signal after a Hilbert transform of the raw mono sound signal at its original sampling rate (48 kHz). The result is downsampled to match the frame rate of the video stimulus (30 Hz) and is then z-scored to standardize across stimuli. Finally, we construct a Toeplitz matrix from the z-scored sound envelope with 30 columns to capture up to 1 second delay, and add a column with a constant value (1) to model a potential offset. The first 29 rows of the stimulus are removed to avoid edge effects of the filtering.

EEG evoked response preprocessing

The EEG evoked response $r_i(t)$ is preprocessed as follows. The starting value is subtracted from the data in order to remove the DC offset and minimize transients from any filters that are applied. Then, a high-pass 5th order Butterworth filter with a cutoff frequency of 0.5 Hz is applied. Signal from the 6 EOG electrodes is then regressed out from the EEG channels with conventional least-squares, leaving 64 EEG channels for analysis. The data is visually inspected and any channels that are excessively noisy due to electrode or recording quality issues are set to zero. Additionally, any samples that were more than four standard deviations away from zero (in a 60 second segment) are set to 0 along with any samples preceding or following these outlier samples by less than 40 ms. Note that for a zero-mean signal setting samples to zero effectively removes those samples from the correlation measure, while discounting the correlation values by the fraction of samples removed. In total 3.18% of the data are set to zero with no meaningful difference across the 4 conditions (congruent -9 dB 3.15%, incongruent -9 dB: 2.15%, congruent -6 dB: 3.14%, incongruent -6 dB: 3.21%). The EEG is then downsampled to the framerate of the video stimuli (30 Hz).

Regression of visual feature

To reduce possible effects of the visual stimulus on the EEG response we regress out activity that correlates with frame-to-frame differences in luminance (squared and averaged over all pixels). We previously established that this feature correlates well with the EEG evoked response during movies (Poulsen 2017). In fact, for some movie stimuli it correlates with the EEG significantly better than the sound envelope (Dmochowski 2017). Prior to regression, we removed large outliers in frame-to-frame differences due to scene cuts by setting values that were over three standard deviations above the mean to zero. We then used linear regression to remove any activity in the EEG evoked response that correlates with this feature. To do so, and allow for delays and an offset, we built the same Toeplitz matrix as we did for the auditory feature, and regress this against each EEG electrode. We therefore build, following (Lalor and Foxe 2009, 2010), an encoding model for the visual feature $s(t)$ (the frame-to-frame differences)

$$\widehat{r}_i(t) = h_i * s(t) \tag{2}$$

and subtract that from the EEG response $r_i(t)$ prior to the CCA procedure to measure SRC. Fig. 3(c) shows the spatial distribution of the correlation values R for the encoding model (Pearson's correlation coefficient between estimated EEG response $\widehat{r}_i(t)$ and actual EEG response $r_i(t)$; R^2 is the more conventional figure of merit in linear regression but we use R as it lends itself better to visualization and statistical tests). To compare the explanatory power of this visual feature to that of the auditory feature, we repeat this regression analysis using the sound envelope of the noisy speech that is used in the main analysis. The resulting R values are shown in Fig. 3(d). In order to determine whether the R values of these two regression models were significantly different from one another (i.e. one feature was significantly more predictive than the other), we divided the available data into 20 equal segments and generated a visual and auditory regression model for each. We then evaluated whether the corresponding R of the 20 auditory and 20 visual regressions were significantly different from one another for each electrode using a t-test. Electrodes that are significantly better correlated with the auditory feature are indicated with a '+' symbol in figures 3(c) and in 3(d) if significantly better correlated with the visual feature.

For comparison, we also built an encoding-decoding CCA model (as in Eqs. 1) for the visual feature, after subtracting the auditory feature with an encoding model (as in Eq. 2). These results are shown in figure 3(b). We note that measuring SRC after linear subtraction of the nuisance feature is conceptually similar to partial coherence use previously in the context of ‘tracking’ audiovisual speech (Park 2016).

Statistical significance of SRC values

In order to estimate statistical significance of SRC values, 1000 sets of circularly shuffled EEG data were correlated with the stimuli using the same procedure as the normal EEG data. Shuffling was performed along the time dimension. Then, the correlation values of these shuffled sets were compared to the correlation values from the normal EEG data. For each of the three components, the normal SRC value was higher than all 1000 of the values generated from the circular shuffle procedure. Therefore, we concluded that $p < 0.001$ for all three components in the CCA model correlating noisy speech with the EEG responses.

Power analysis using bootstrapping

We calculate SRC for each 60 second segment for each subject in all four stimulus conditions (resulting in a total of 30 x 4 x 20 measured - stimuli x conditions x subjects). We then draw for each subject a sample of size N with replacement, separately for the congruent -9 dB and incongruent -6 dB conditions. We then perform a Wilcoxon rank-sum test to determine if the two conditions differ in SRC with a significance value of 0.05. We repeat the bootstrap sampling 1000 times and compute the fraction of these repeats where the test detected a significant difference. This fraction is the estimated statistical power.

Results

Study participants (N=20) were presented continuous audiovisual speech (of President Barack Obama giving an televised address about the Affordable Care Act). The auditory speech was presented with spectrally-matched acoustic noise at signal-to-noise ratios of -9 dB and -6 dB. To manipulate intelligibility the visual speech was either congruent or incongruent with the auditory speech (Park 2016). Participants were instructed to detect one or more occurences of a target word in each 60 second speech segment. Detection performance served as the objective indicator of speech intelligibility (Crosse 2016). We recorded EEG during stimulus presentation to assess stimulus-response correlation (SRC) between the envelope of the noisy sound and the EEG evoked response (see example in Figure 1). The goal was to test whether SRC is a viable neural predictor of speech intelligibility, in the absence of clean speech.

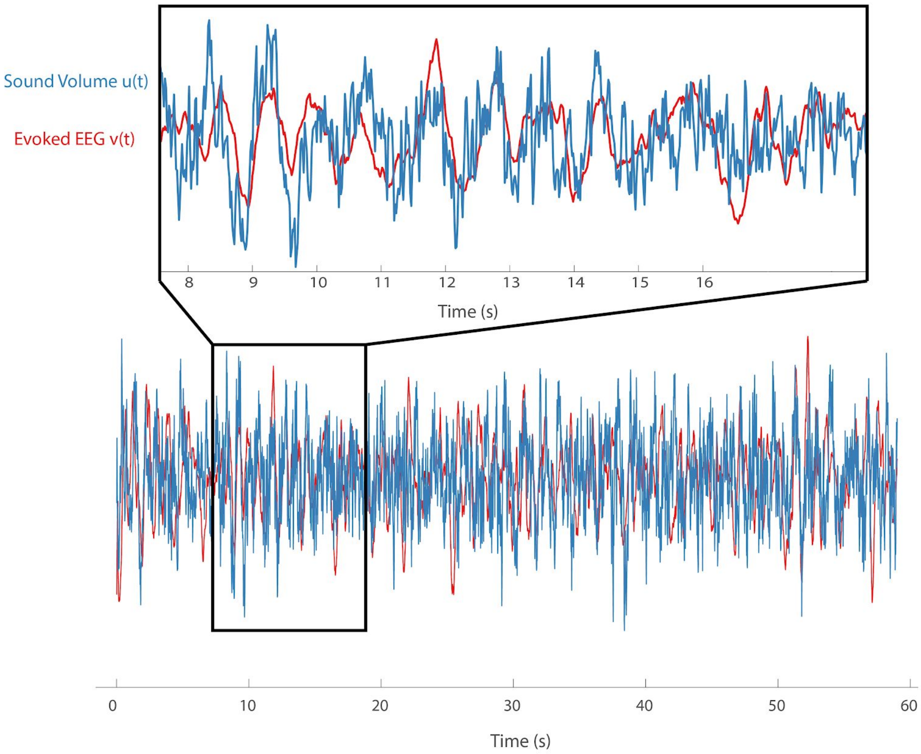

Figure 1 - Example of correlated stimulus and response in time. This represents 60s of the sound amplitude of the stimulus and the EEG evoked response. To be precise, here the filtered sound envelope $\hat{u}(t)$ is shown in blue and the projected EEG $\hat{v}(t)$ in red (see Eqs. 1).

Congruent visual speech improves word detection

Behavioral performance for detecting target words was quantified in terms of precision and detection rate. Both metrics improve in conditions with higher SNR and the congruent visual speech as compared to incongruent speech (Figure 1). In the following we use F1 score, which combines precision and detection rate, as the primary outcome measure of detection performance. We performed a two-way repeated measures ANOVA of the F1 score with main factors of noise level and congruency. We found a very large effect for congruency [F(1, 19) = 917.4, p = 1 x 10^{-17}], for noise level [F(1,19) = 284.9, p = 6 x10^{-13}], and also for the interaction between noise level and congruency [F(1,19) = 118.1, p = 1 x 10^{-9}]. Follow-up comparisons for each noise condition indicate that the effect of congruency is significant at both noise levels [-9 dB: t(38) = 13.9, p = 2 x 10^{-16} , -6 dB: t(38) = 3.91, p = 3 x10^{-4}]. Neither one of these effects are surprising as the relative benefits of visual speech at different noise levels are well established (Ross 2007). The results do however validate our approach of using audiovisual speech to manipulate intelligibility without changing the auditory stimulus.

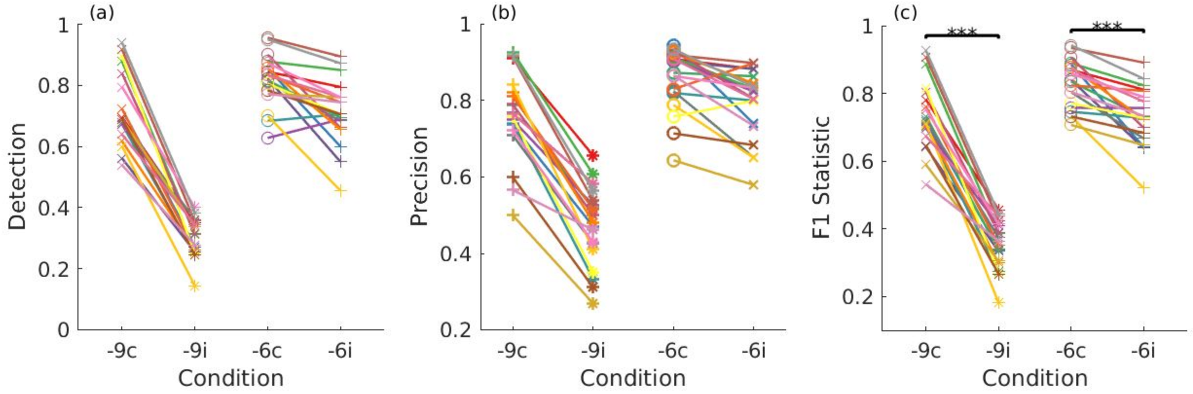

Figure 2. Behavioral word detection performance for congruent and incongruent audiovisual speech at -9 dB and -6 dB noise. Labels for the conditions are: -9c: -9 dB congruent, -9i: -9 dB incongruent, -6c: -6 dB congruent, -6i: -6 dB incongruent (a) The behavioral word detection performance of subjects. Each line corresponds to a different subject. (b) The precision of each subject over all trials in each condition. (c) F1 statistic for each subject.

Congruent visual speech enhances auditory stimulus-response correlation

We compute stimulus-response correlation (SRC) between the sound envelope of the noisy speech and the EEG evoked response. We use canonical correlation analysis (CCA) to relate the stimulus at various delays on one side, to the EEG response at multiple electrodes on the other side. The CCA model captures several components that correlate between the stimulus and the response (see Methods; Dmochowski 2017). The corresponding spatial and temporal response functions of the CCA model are shown in Figure 3 (a) for the three components with the strongest SRC. While correlation values are relatively small, they are nonetheless significant given the large amount of data available. [r = 0.039, 0.025, 0.009, respectively for each of the first 3 components, p < 0.001 using circular shuffle statistics]. We measure overall SRC by summing these correlation values over the first three components, but compute this separately for each subject and stimulus condition (Figure 4). SRC appears to improve with improved SNR and is larger for the congruent visual speech as compared to incongruent speech (Figure 4). A two-way repeated-measures ANOVA confirms a very significant effect for both congruency [F(1, 19) = 109.4, p = 2 x 10^{-9}] as well as noise level [F(1, 19) = 44.54, p = 2 x 10^{-6}], but reveals no interaction between the two [F(1,19) = 0.87, p = 0.363], suggesting that congruent visuals were equally effective at changing SRC at both noise levels. A follow-up pairwise comparison confirms that the congruency effect is present at both noise levels. [-9 dB: t(38) = 3.68, p = 7 x 10^{-4}, -6 dB: t(38) = 2.43, p = 0.02].

We suspected that these effects were the result of stronger evoked activity, not just in response to the speech, but also to the noise fluctuations. To test for this, we repeated the analysis now correlating the EEG to the envelope of the clean speech, and the envelope of only the noise (Figures S3). The SRC of EEG with the clean-speech envelope [r = 0.05, 0.03, 0.01, respectively for each of the first 3 components, p < 0.001 using circular shuffle statistics] increased numerically as compared to the noisy-speech envelope [from above: r = 0.039, 0.025, 0.009]. The SRC for the clean-speech envelope was modulated with SNR [F(1, 19) = 52.53, p = 7 x 10^{-7}] and congruence [F(1, 19) = 54.43, p = 5 x 10^{-7}], just as much as for the noisy-speech (see above). In contrast, the SRC with the noisy-only envelope was much weaker [r = 0.01, 0.0047, 0.0019, respectively for each of the first 3 components, p < 0.001, p = 0.001, p = 0.14 using circular shuffle statistics] and was not modulated by either noise level [F(1, 19) = 0.019, p = 0.89] or congruence [F(1, 19) = 0.82, p = 0.38]. This result conflicts with our initial hypothesis, and instead suggests that the listeners managed to extract speech fluctuations from the noise (the envelope of clean-speech and noisy-speech correlate weakly; -9 dB: r = 0.22 ± 0.092, -6 dB: r = 0.20 ± 0.14), and were more successful at this when the visual cues were congruent with the sound.

Visual fluctuations do not account for gain in auditory stimulus-response

An important confound in our experimental design is the visual information that is presented to the subject. In the congruent condition there may be visually evoked responses in the EEG that are correlated with the speech amplitude (elicited by lip or head movements), while in the incongruent condition this should not be the case. We made an effort to measure and remove evoked activity that can be explained from the visual stimulus. As a feature of the visual stimulus we used frame-to-frame differences in luminance, which has been previously established as a good correlate to the EEG (Poulsen 2017, Dmochowski 2017), and captures both head and lip movements as these are the dominant changes in these talking-head videos. First, we build a linear (encoding) model to predict the EEG from the visual and auditory features. The correlation of each EEG electrode with the stimulus-predicted EEG is shown in Figure 3c and 3d with auditory and visual features as predictors, respectively. We find that 40 electrodes are significantly better predicted by the sound envelope (these electrodes have higher R^2, indicated with a '+' in Figure 3c), whereas four electrodes are better predicted by the frame-to-frame differences (indicated with a '+' in Figure 3d). Thus, the auditory features dominates the EEG responses in most electrodes. Importantly, if we subtract the activity predicted by the visual feature from the EEG, the resulting SRC measure (between sound amplitude and EEG) are virtually unchanged from what is shown in Figure 4. The conclusions on the statistical comparisons for the effects also remain unaltered [congruency: F(1,19) = 82.71, p = 2×10^{-8}, noise: F(1,19) = 50.17, p = 9×10^{-7}, and the interaction of noise and congruency: F(1,19) = 0.85, p = 0.368].

To further test whether there are visual contributions to the audio-correlated EEG activity, we attempted to relate the EEG of the incongruent condition to the audio envelope matching the video, i.e. we try to relate the EEG to the audio that was not heard, but may have been inferred by subjects from the visuals. This should capture any nonlinear contributions the video might have on the EEG that is correlated to the auditory features. To this end we measured the SRC of the envelope of the clean unheard speech with the EEG in the incongruent video condition. A new CCA model is built for this analysis (neither the heard nor seen features were regressed out here). If the visuals alone allowed the subject to infer the speech, we would expect to find that there is a significant correlation between the seen visuals and the unheard audio. Instead, we find that the SRC generated by this analysis does not reach significance (neither individually nor in the sum of the first 3 components, r = 0.0075, 0.0057, 0.0047, measured for all four conditions combined, p>0.2 when compared against 1000 circular shuffled data).

Thus, in total, it is unlikely that the strong effects of congruency observed in Figure 4 are due to a direct contribution of correlated visual evoked activity. Instead, the results point to an interaction of the visual queues with the auditory processing.

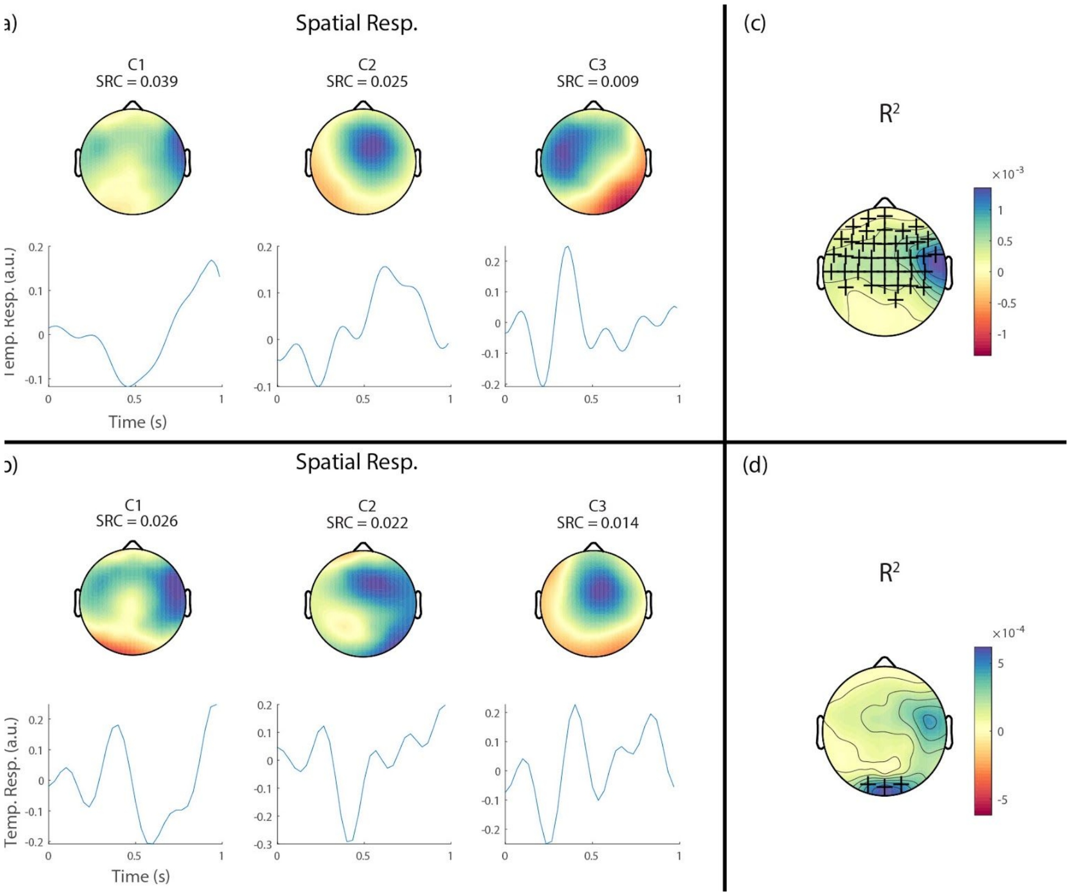

Figure 3 - Models for different features: (a, b) Models derived with canonical correlation analysis (CCA) for the auditory feature (a) and the visual feature (b). The auditory feature is the amplitude envelope of the noisy speech stimulus. The visual feature is the frame-to-frame difference of the video stimulus. The CCA model correlates the EEG response with each of these features. Stimulus-response correlations (SRC) values shown separately for each of the three CCA components indicate the average across all subjects and conditions. Scalp distribution (top) indicates the spatial EEG response of each component. Time courses indicate the temporal EEG response (bottom). (c, d) Prediction performance for the linear regression model for the auditory feature (c) and visual feature (d). Scalp distribution indicates the R^2 performance for predicting the EEG response from the features. '+' signs in panel (c) indicate that in those electrode locations the auditory feature predicted the EEG with higher R^2 than the visual feature, and vise versa in panel (d).

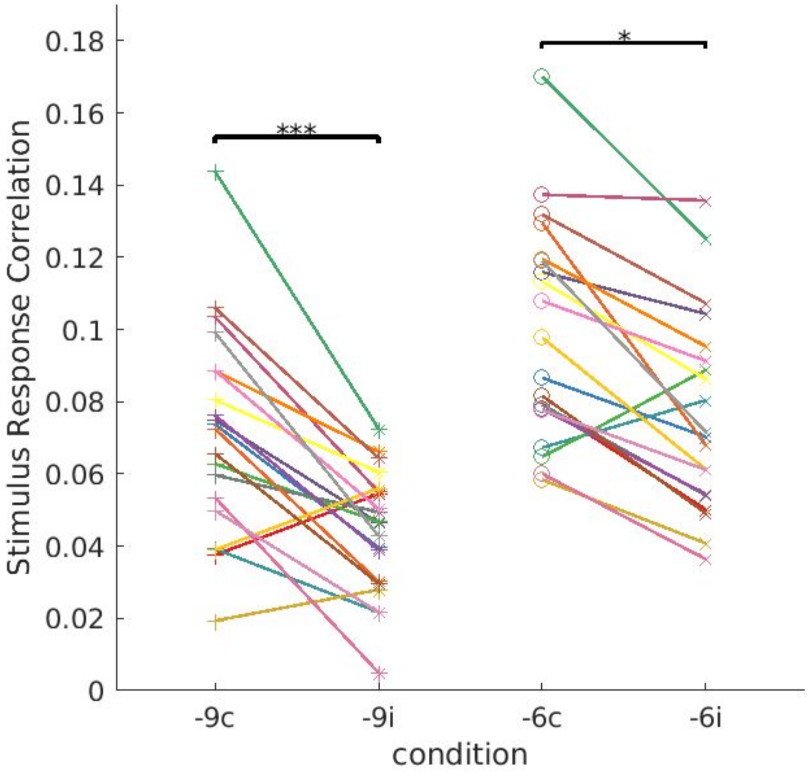

Figure 4 - SRC scores for each subject over all conditions. Each subject is shown in a different color and subject performance within each noise condition is connected with a line. Labels for the conditions are: -9c: -9 dB congruent, -9i: -9 dB incongruent, -6c: -6 dB congruent, -6i: -6 dB incongruent.

Behavior is Correlated with Stimulus Response Correlation

We have shown that both behavioral performance on the task as well as SRC increase in congruent conditions and in less noisy conditions. To summarize this, we display the results together in Figure 5(a). Evidently, for the majority of subjects, and in both noise conditions, an improvement in SRC coincides with an improvement in word detection performance (positive slope in Figure 5a). The same is evident in Figure 5(b) where most of the subjects tested appear in the 1st quadrant, meaning that the change in SRC and F1 have the same sign [p = 0.0026, p=0.0004 for -9 dB and -6 dB respectively, sign test]. These results indicate that the SRC is a significant predictor of behavioral performance within subjects.

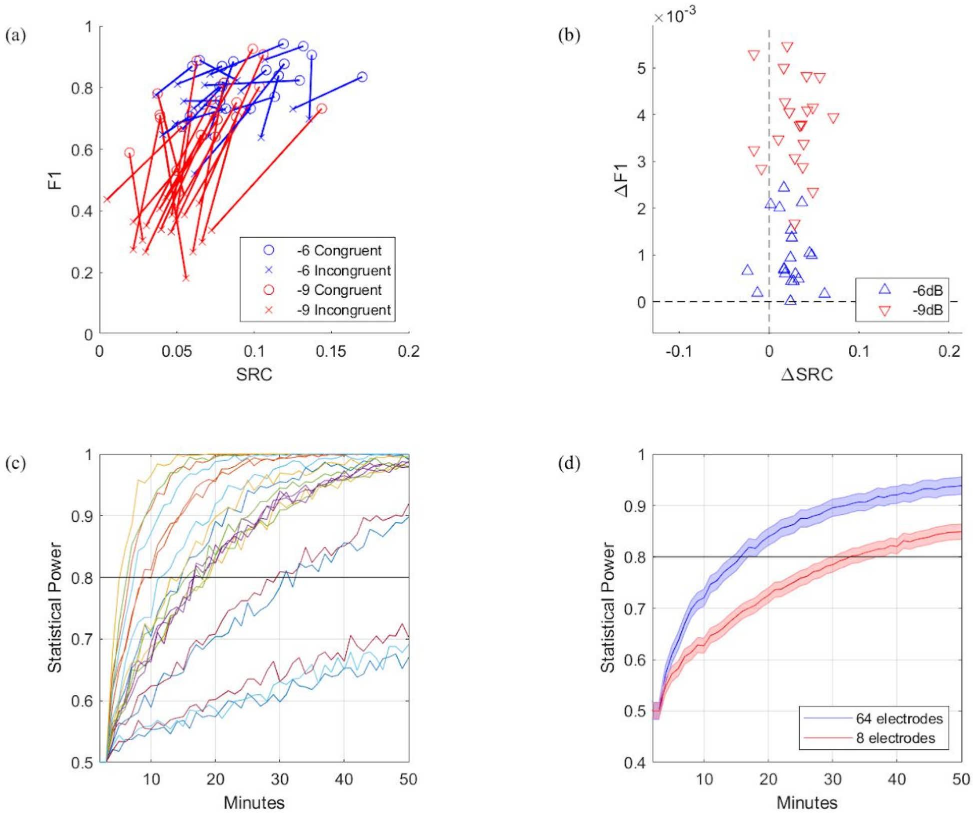

Figure 5 - (a) Comparison between behavioral word detection performance and SRC over conditions. Each line represents a subject. (b) Difference between congruent and incongruent conditions for each subject and both noise conditions. Each subject is represented as a point. Subjects that fall in first quadrant indicate that gains in one metric coincide with gains in the other metric for these subjects. (c) bootstrap estimate of statistics power, i.e. chance to detect a difference in SRC (between the -9 dB incongruent and -6 dB congruent conditions) with type 1 error rate of 0.05, as a function of EEG signal duration. Each curve represents one of the 20 subjects. (d) Bootstrap estimates of statistical power as a function of signal duration. Data was averaged over all 20 subjects to show contrast between predictive power of the full 64 electrode montage and a reduced 8 electrode montage. Shaded area represents standard error.

Detecting changes of stimulus-response correlation in single subjects

In order to use stimulus-response correlation in practice, say, to adjust a hearing aid for improved intelligibility, one needs to be able to assess this within single subjects after recording a limited amount of data (both perceived sound and EEG). Thus, we are interested in estimating how much EEG/sound data is needed to reliably measure a change in SRC for individual subjects. To this end we perform a statistical power analysis using bootstrapping. Specifically, we measure for each subject the odds of detecting a difference in SRC between two stimulus conditions for varying sample sizes. Each SRC value is calculated on 60 seconds of data, thus sample size corresponds to the number of minutes available to detect a change in SRC. In practice, changes in SRC may be due to changes in SNR as well as changes in neural processing (with an otherwise identical stimulus). To emulate this, we ask whether we can detect a difference between the -9 dB incongruent condition and the congruent -6 dB conditions, which correspond to robust gains in word detection performance. The results of Figure 5(c) suggest that for most subjects (15 out of 20) one can detect differences in SRC with 80% power with 5-20 minutes of data. In the context of hearing aids only a few electrode around the ears may be available (Mirkovic 2016, Debener 2012). To determine if our approach still works in this context, we repeat the analysis using only 4 electrodes immediately above the left and right ear (FT7, FT8, T7, T8, TP7, TP8, C5, C6). When using only these 8 electrodes we still find a strong dependence of SRC on signal-to-noise ratio [F(1,19) = 23.59, p = $1x10^{-4}$] and congruence [F(1,19) = 43.58, p = $2x10^{-4}$]. And with eight electrodes one now requires ~30-40 minutes of data to detect changes in SRC with 80% power. (Figure 5d).

Discussion

There is a large body of recent work on the phenomenon of “speech tracking” -- the correlation of brain activity to the rapid fluctuation of speech that a listener is attending to. Our main finding is that improved speech perception coincides with an increase in stimulus-response correlation, which is consistent with much of the previous literature (e.g. Ding and Simon 2013, Peele 2013, Vanthornhout 2017). Previous work has also shown that this correlation can be increased by adding the corresponding visual speech to the auditory speech (Crosse 2015), and that this gain correlates with gains in speech detection performance across subjects (Crosse 2016). We extend this earlier work by using incongruent speech as a control condition in noise. We demonstrated for the first time that this effect correlates with gains in behavioral word detection performance within individual subjects.

The amplitude fluctuations in noisy speech used in the present study at -9 dB are dominated by noise (they are only weakly correlated with the original clean speech). The fact that correlation of brain responses to these mostly-random noise fluctuations are predictive of speech intelligibility is, in our view, quite remarkable. Previous speech tracking work assumed that the correlation to the clean speech is an indication of speech processing. One compelling theory is that the correlation of EEG to speech reflects the parsing of hierarchical semantic features of language (Ding 2015), i.e. endogenous cortical rhythms are entrained to the rhythmic structure of words, phrases, and sentences that occur in natural speech. Others claim that this is not necessary and that the correlation can be explained simply as a response to the acoustic features of words presented in the stimuli (Frank 2018). The present results suggest that when a listener successfully identifies speech embedded in noise, this speech also evokes correlated EEG responses, and that EEG responses are weaker when the identical speech sound is not detected by the listener. Note that we make no causal claim here in terms of word detection causing the evoked responses or vise versa. It is just as possible that when a listener fails to understand the speech in adverse hearing conditions, attention to the speech sound wanes, responses evoked by speech diminish, and stimulus-response correlation drops.

Our work was motivated by the desire to objectively assess intelligibility in the context of hearing aids. There are several important novel contributions over prior work on speech tracking in this regard. First, we have shown reliable prediction of intelligibility within individual subjects. This is a key requirement to tuning a hearing aid for intelligibility in each individual. Second, we have maintained the auditory stimulus unchanged, thus ruling out the possibility that changes in auditory stimulus characteristics altered the estimate of stimulus-response correlation. Thus changes in SRC observed here are likely due to the differences in how the auditory stimulus itself is processed. In contrast, most previous literature on speech-tracking modified the sound (Kösem 2017, Luo 2007, Nourski 2009) and thus changes in SRC may have resulted from altered characteristics of the sound itself. One exception is the work of Crosse (2016), who did maintain the auditory stimulus constant, but did not include a visual control condition as we have done here with the incongruent speech. Third, our measure of stimulus-response correlation did not require access to the original clean speech, which is not available in practical scenarios. To our knowledge all prior work on speech tracking has relied on access to the original clean speech.

Given our pragmatic interest in hearing aids, we also tested how much data is required to reliably predict the modulation of SRC to allow inferring changes in intelligibility. We find that 5 to 30 minutes of data are sufficient to detect changes in SRC (with 80% power in the majority of subjects). In this study, the auditory stimulus was kept constant in order to explore purely cognitive effects. In practice, the stimulus conditions do continually change. In this regard it is interesting to note that when the SNR improves, SRC also increases. Thus, SRC may be a useful metric for speech intelligibility within and across signal conditions.

One caveat to our results is that the congruent visual stimulus (e.g. lip movements) could have elicited brain responses that are correlated to the auditory stimulus (Crosse 2016), whereas in the incongruent condition additive evoked response would be uncorrelated (Park 2016). Thus, increased stimulus-response correlation with the sound envelope could be a reflection of this additive evoked activity and not a reflection of altered auditory processing. To mitigate this confound, we remove any EEG activity that could be linearly explained by features of the visual stimulus (similarly to Park 2016). Here we used frame-to-frame differences of the video stimulus as visual features because it was previously found to capture a significant fraction of the EEG evoked response in video (Poulsen 2017, Dmochowski 2017), and because it captures lip and head movements, which can both aid speech recognition in noise (Sumby and Pollack 1954, Thomas and Jordan 2004, Munhall 2004). While there may be other visual features that contributed to such a confound, we note that visually-related EEG activity captured almost as much of the EEG variance as the sound envelope, and that removing this activity had only a marginal effect on the link between SRC and intelligibility. Most importantly, we demonstrated that the EEG activity elicited by the incongruent video does not significantly correlate with the corresponding unheard audio. Thus the video is unlikely to contribute directly to the evoked response that correlates with the sound in the incongruent condition. Instead, the enhanced evoked responses in the congruent condition are likely the result of an interaction between vision and auditory processing, consistent with the results of Crosse et al. 2016. Ultimately, however, this caveat about the visual confound remains.

There are a number of open questions worth addressing in future research. For instance, there have been attempts to narrow the range of endogenous neural rhythms that are the source of neural entrainment with speech. Some researchers claim that delta range oscillations (1-4 Hz) are primarily responsible (Molinaro 2018). Others argue that theta range oscillations (4-8 Hz) are the primary driver (Ghitza 2012, Bosker 2018), and still others claim that it is both, or that oscillatory activity in other frequency bands, such as gamma activity (30-50 Hz) plays a large role as well (Nourski 2009). Future work could focus on specific frequency bands to determine which is most predictive of intelligibility as we have measured here. The present paradigm may be particularly useful in this context as the auditory stimulus is kept constant, thus ruling out spectral differences that are due to differing stimulus properties.

Here we have measured stimulus-response correlation after filtering the stimulus in time and filtering the EEG response in space. We refer to these operations as ‘encoding’ and ‘decoding’ respectively (Dmochowski 2017). Correlation between stimulus and response is then assessed on these filtered signals, which may contribute to the robustness of the approach because the filters are optimized to remove unrelated activity. The approach differs from the conventional approaches used in the speech-tracking literature, which measure correlation with the stimulus after only ‘decoding’ the EEG activity (e.g. de Taillez 2018, Mesgarani 2012, O’Sullivan 2015, Vanthornhout 2017), or only ‘encoding’ the stimulus (e.g. Ding 2014, Hyafil 2015, Power 2012, Zion Golumbic 2013). Future work could compare the relative merits of these various methods for the purpose of predicting speech intelligibility.

Here we have used EEG recorded over the entire scalp, which is not realistic in the context of hearing aids. An important practical question is whether it suffices to record EEG in the ear (Kidmose 2012, Looney 2014, Goverdovsky 2016) or around the ear (Mirkovic 2016) in order to reliably measure SRC that is predictive of intelligibility. In this context it may be worth noting that attentional modulation of SRC to continuous speech have been measured with auditory brainstem responses with just two electrodes (mastoid and Cz), although the signals are admittedly difficult to detect (Forte 2017).

In total, we have provided a first proof of principle that variations in intelligibility, as assessed objectively in terms of word detection, can be predicted from a subject's EEG without access to the clean speech. Future work outlined here are only the initial steps required to develop neurally-adaptive speech enhancement for hearing aids as we have envisioned it here.

Acknowledgements

We thank Mick Crosse and collaborators for sharing the audiovisual speech stimuli used in this study, and the code to adjust SNR. We also thank Ed Lalor for suggestions on how to address the issue of visual confounds. We thank the reviewers of this paper for a number of helpful suggestions on how to improve the analysis presented here. We also want to thank Brian Aguilar Avila for careful proofreading. We would like to acknowledge the generous financial support of GN Resound. This work was also supported in part by the Army Research Laboratory under Cooperative Agreement Number W911NF-10-2-0022.

References

Arnal, L. H., & Giraud, A. L. (2012). Cortical oscillations and sensory predictions. Trends in Cognitive Sciences, 16(7), 390–398.

Aru, J., Aru, J., Priesemann, V., Wibral, M., Lana, L., Pipa, G., … Vicente, R. (2015). Untangling cross-frequency coupling in neuroscience. Current Opinion in Neurobiology, 31(September 2014), 51–61.

Biesmans, W., Das, N., Francart, T., & Bertrand, A. (2017). Auditory-inspired speech envelope extraction methods for improved EEG-based auditory attention detection in a cocktail party scenario. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 25(5), 402–412.

Bosker, H. R., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent speech: behavioural evidence from rate normalisation. Language, Cognition and Neuroscience, 0(0), 1–13.

Canolty, R. T., & Knight, R. T. (2010). The functional role of cross-frequency coupling. Trends in Cognitive Sciences, 14(11), 506–515.

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.

de Taillez, T., Kollmeier, B., & Meyer, B. T. (2018). Machine learning for decoding listeners’ attention from electroencephalography evoked by continuous speech. European Journal of Neuroscience, (June), 1–8.

Debener, S., Minow, F., Emkes, R., Gandras, K., & de Vos, M. (2012). How about taking a low-cost, small, and wireless EEG for a walk? Psychophysiology, 49(11), 1617–1621.

Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Cortical Measures of Phoneme-Level Speech Encoding Correlate with the Perceived Clarity of Natural Speech. Eneuro, ENEURO.0084-18.2018.

Ding, N., Simon, J. Z. (2012). Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. Journal of Neurophysiology, 107(1), 78–89.

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., Chatterjee, M., & Simon, J. Z. (2014). Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. NeuroImage, 88, 41–46.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2015). Cortical tracking of hierarchical linguistic structures in connected speech. Nature Neuroscience, 19(1), 158–164.

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.

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. NeuroImage, 85, 761–768.

Forte, A. E., Etard, O., & Reichenbach, T. (2017). The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. ELife, 6, 1–12.

Frank, S. L., & Yang, J. (2018). Lexical representation explains cortical entrainment during speech comprehension. PLoS ONE, 13(5), 1–11.

Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech: Intelligibility of speech with a manipulated modulation spectrum. Frontiers in Psychology, 3(JUL), 1–12.

Goverdovsky, V., Looney, D., Kidmose, P., & Mandic, D. P. (2016). In-Ear EEG From Viscoelastic Generic Earpieces: Robust and Unobtrusive 24/7 Monitoring. IEEE Sensors Journal, 16(1), 271–277.

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.

Hotelling, H. (1936). Relations Between Two Sets of Variates. Biometrika, 28(3/4), 321.

Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., & Giraud, A.-L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. ELife, 4.

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.

Kidmose, P., Looney, D., & Mandic, D. P. (2012). Auditory evoked responses from Ear-EEG recordings. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 586–589.

Kleiner, M., Brainard, D. H., Pelli, D. G., Broussard, C., Wolf, T., & Niehorster, D. (2007). What's new in Psychtoolbox-3? Perception, 36, S14.

Kösem, A., Bosker, H. R., Takashima, A., Meyer, A., Jensen, O., & Hagoort, P. (2017). Neural entrainment determines the words we hear.

Kong Y.Y., Somarowthu A., & Ding N. Effects of Spectral Degradation on Attentional Modulation of Cortical Auditory Responses to Continuous Speech. Journal of the Association for Research in Otolaryngology 16: 783–796, 2015.

Lalor E.C., & Foxe J.J.. Visual evoked spread spectrum analysis (VESPA) responses to stimuli biased towards magnocellular and parvocellular pathways. Vision research. 2009 Jan 1;49(1):127-33.

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. European Journal of Neuroscience, 31(1), 189–193.

Looney, D., Kidmose, P., & Mandic, D. P. (2014). Ear-EEG: User-Centered and Wearable BCI. In C. Guger, B. Allison, & E. C. Leuthardt (Eds.), Brain-Computer Interface Research: A State-of-the-Art Summary -2 (pp. 41–50). Berlin, Heidelberg: Springer Berlin Heidelberg.

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

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature, 485(7397), 233–236.

Mirkovic, B., Bleichner, M. G., De Vos, M., & Debener, S. (2016). Target speaker detection with concealed EEG around the ear. Frontiers in Neuroscience, 10(JUL), 1–11.

Molinaro, N., & Lizarazu, M. (2018). Delta(but not theta)-band cortical entrainment involves speech-specific processing. European Journal of Neuroscience.

Munhall K.G., Jones J.A., Callan D.E., Kuratate T., & Vatikiotis-Bateson E. Visual prosody and speech intelligibility: Head movement improves auditory speech perception. Psychological science. 2004 Feb;15(2):133-7.

Nourski, K. V., Reale, R. A., Oya, H., Kawasaki, H., Kovach, C. K., Chen, H., … Brugge, J. F. (2009). Temporal Envelope of Time-Compressed Speech Represented in the Human Auditory Cortex. Journal of Neuroscience, 29(49), 15564–15574.

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.

O'Sullivan, J., Chen, Z., Sheth, S. A., McKhann, G., Mehta, A. D., & Mesgarani, N. (2017). Neural decoding of attentional selection in multi-speaker environments without access to separated sources. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 1644–1647.

Park H., Kayser C., Thut G., & Gross J. Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. Elife. 2016 May 5;5:e14521.

Peelle J.E., Gross J., & Davis M.H. Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. Cerebral Cortex 23: 1378–1387, 2013.

Petroni, A., Cohen, S. S., Ai, L., Langer, N., Henin, S., Vanderwal, T., … Parra, L. C. (2017). Age and sex modulate the variability of neural responses to naturalistic videos. BioRxiv.

Poulsen, A. T., Kamronn, S., Dmochowski, J., Parra, L. C., & Hansen, L. K. (2017). EEG in the classroom: Synchronised neural recordings during video presentation. Scientific Reports, 7.

Power, A. J., Foxe, J. J., Forde, E. J., Reilly, R. B., & Lalor, E. C. (2012). At what time is the cocktail party? A late locus of selective attention to natural speech. European Journal of Neuroscience, 35(9), 1497–1503.

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.

Sumby WH, Pollack I. Visual contribution to speech intelligibility in noise. The journal of the acoustical society of america. 1954 Mar;26(2):212-5.

Thomas SM, Jordan TR (2004) Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. J Exp Psychol Hum Percept Perform 30: 873–888.

Van Eyndhoven, S., Francart, T., & Bertrand, A. (2017). EEG-Informed Attended Speaker Extraction from Recorded Speech Mixtures with Application in Neuro-Steered Hearing Prostheses. IEEE Transactions on Biomedical Engineering, 64(5), 1045–1056.

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

Wang, Z.-Q., Le Roux, J., & R. Hershey, J. (2018). Multi-Channel Deep Clustering: Discriminative Spectral and Spatial Embeddings for Speaker-Independent Speech Separation. IEEE International Conference on Acoustics, Speech and Signal Processing.

Wong, D. D. E., Fuglsang, S. A., Hjortkjær, J., & Ceolini, E. (2018). A Comparison of Temporal Response Function Estimation Methods for EEG-based Auditory Attention Decoding. Biorxiv, 1–22.

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.

Zion-Golumbic E, Schroeder CE. Attention modulates ‘speech-tracking’ at a cocktail party. Trends in cognitive sciences. 2012 Jul 1;16(7):363-4.

Zoefel, B., Archer-Boyd, A., & Davis, M. H. (2018). Phase Entrainment of Brain Oscillations Causally Modulates Neural Responses to Intelligible Speech. Current Biology, 28(3), 401–408.e5.

Supplement

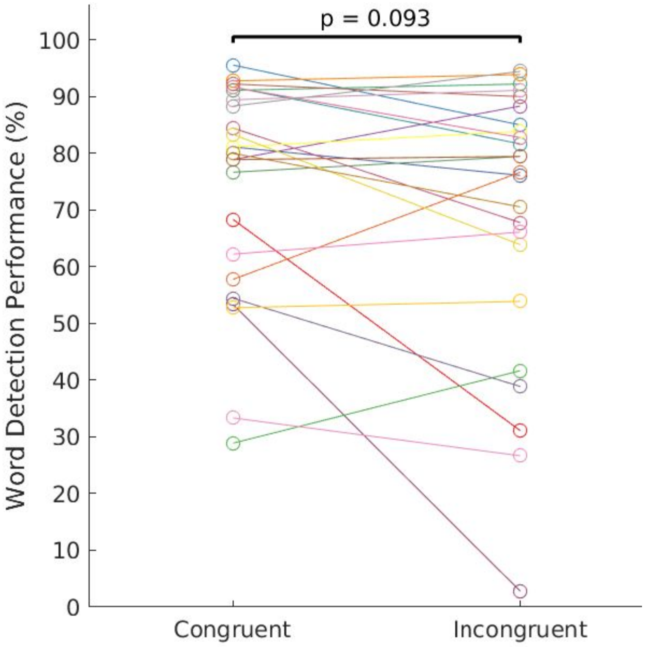

Figure S1 - Performance on the behavioral task at -3 dB SNR in 24 subjects. We did not find a significant difference in the performance. We note that in this pilot experiment subjects did not receive performance feedback during the experiment. Providing this feedback was useful to keep subjects motivated. We ascribe the low performance of a few subjects in this relatively easy task a lack of motivation. Additionally, it is known that at -3 dB, the difference of congruent and incongruent speech is relatively small (Ross 2007)

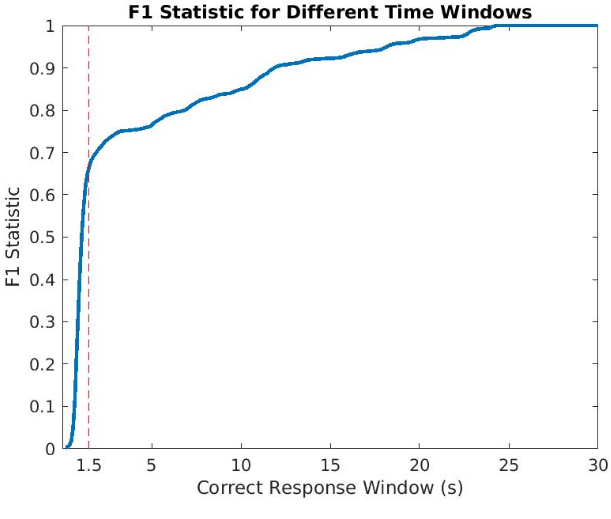

Figure S2 - Results for the F1 score over all subjects and conditions using different size windows for categorizing hist, miss and false alarms. The window used in the main analysis (1.5 s) is marked in red. At this specific value, F1 score no longer improves substantially.

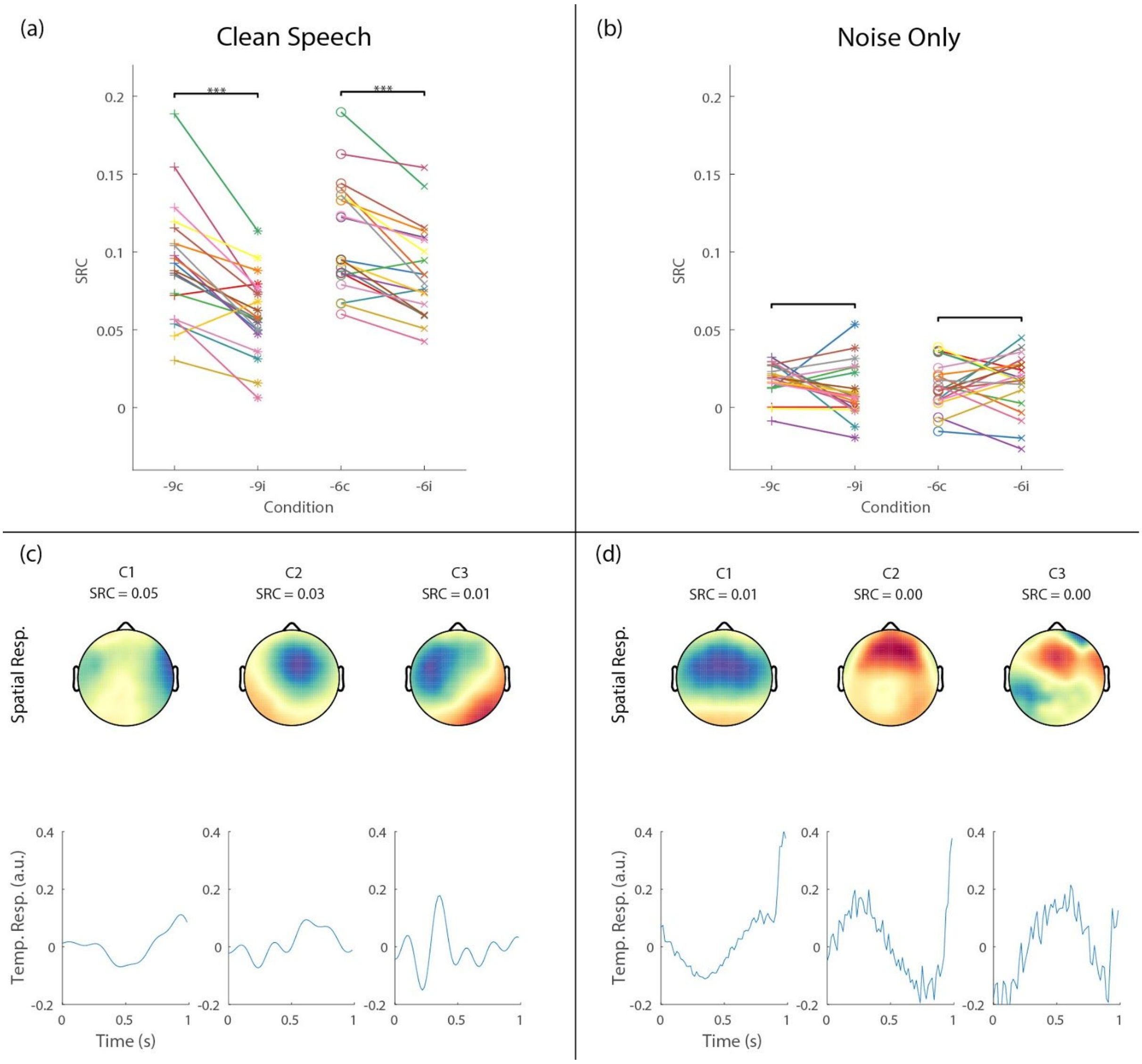

Figure S3 - (a-b) SRC scores for CCA model trained, respectively, with clean speech envelope (a) and noise-only audio envelope (b). The noise-only signal is extracted from the noisy speech heard by subjects after subtracting the clean-speech signal. (c-d) Spatial and temporal response for the clean-speech model and the noise-only model respectively.

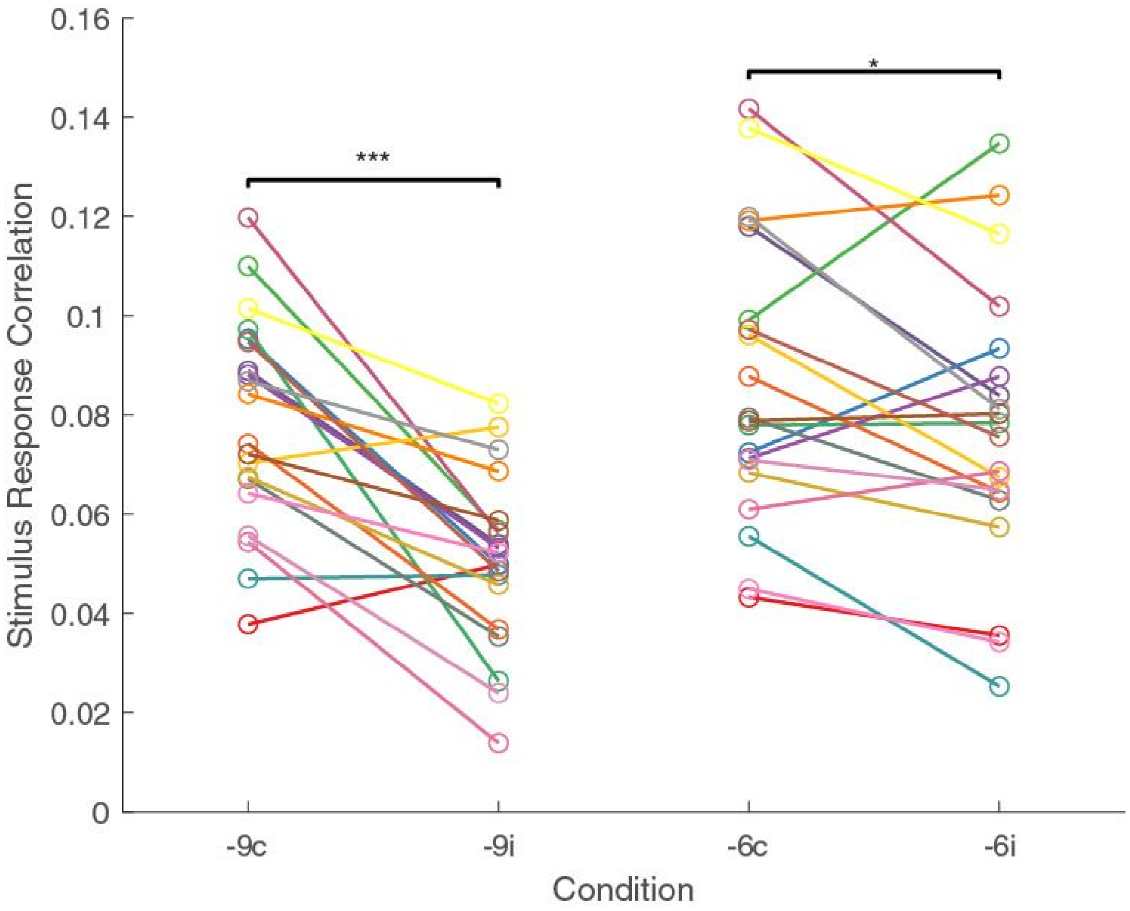

Figure S4 - SRC scores for all subjects using a CCA model trained using clean speech and an 8 electrode subset of the full 64.