Brain representation of acoustic features during goal-oriented dialogues

Joaquin Gonzalez[1,2], Nicolás Nieto[3], Mauro Veneziano[4], Pablo Brusco[1,5], Agustin Gravano[6,7], and Juan E Kamienkowski[1,4,*]

[1] Laboratorio de Inteligencia Artificial Aplicada, Instituto de Cs. de la Computación, Fac. de Cs. Exactas y Naturales, Univ. de Buenos Aires - CONICET
[2] Depto. de Física, Fac. de Cs. Exactas y Naturales, Universidad de Buenos Aires
[3] Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), Univ. Nacional del Litoral - CONICET
[4] Maestria de Explotación de Datos y Descubrimiento del Conocimiento, Fac. de Cs. Exactas y Naturales, Univ. de Buenos Aires
[5] Depto. de Computación, Fac. de Cs. Exactas y Naturales, Univ. de Buenos Aires
[6] Escuela de Negocios, Universidad Torcuato Di Tella
[7] CONICET
joaquin.gonzalez6693@gmail.com

Abstract. During dialogue, speakers have to rapidly represent different aspects of speech in order to respond and continue engaged. In particular, acoustic aspects indicate different intentions or emotions. Understanding how the brain represents those characteristics, both spatially and temporally, is important for developing better interactive systems. In this work, we analyse electroencephalography (EEG) data and audio recordings from ten participants engaged in a dialogue during a collaborative task. We aim to better understand how the envelope of the previous instants of the speech is encoded in the brain. Our results showed a good performance in predicting EEG signals, in particular in frontal electrodes. Moreover, our findings suggest that the brain encodes the speech envelope information with a latency between 0 and 100ms. These results are widely consistent between participants and also with previous work. The present work paves the way for studying brain representations of speech in natural scenarios with EEG.

Keywords: Electroencephalography · Natural Speech · Encoding

1 Introduction

Cognitive neuroscience had a great boost in recent years, mainly driven by the amount and complexity of the available data, together with methods and computational power to analyse it. Experimental protocols moved from highly controlled stimuli to more natural scenarios [14]. This helps to understand the complex brain mechanisms behind everyday behaviours, such as communicating through spoken language [4,14,19,20,27]. Moreover, on one hand, this knowledge could improve the naturalness of current spoken dialog systems by exploring the exact timing of the turn-taking cues in the brain [2]. On the other hand, it will open new paths for exploring brain-computer interfaces [23,28]. Those scenarios require new analysis methods capable of dealing with larger, more complex and often incomplete or unbalanced data [14,17]. Encoding models are multivariable predictive models that take different features of the stimuli to predict some measurement of the brain activity. Among them, ridge regression is widely used for predicting EEG signals in state-of-the-art research [6,8,10,15,24].

Previous work has studied how certain attributes of the speech signal, such as sound envelope, phonemes or pitch level, are represented in the brain [7,18,21,3,16]. These studies used mainly intracranial electrocorticogram (ECoG) recordings and prerecorded stimuli such as isolated sentences or radio monologues. In this work, we aimed to explore the possibility of detecting markers of these representations in non-invasive electroencephalography (EEG) signals and in more natural scenarios, such as participants engaged in a natural dialogue.

2 Materials and Methods

2.1 Data description

Tasks and Participants. The experimental task involves two participants engaged in dialogue while playing a series of instances of the Objects Games [12]. Each session contains 15-30 of such instances. A total of 10 native speakers of Argentine Spanish (4 female, 25.3±5.7 y.o.) participated in the study and agreed to join by signing a consent form. This is a subset of a larger study designed to study turn-taking cues in brain activity.

EEG recording and preprocessing. EEG activity was recorded at 128 positions on a standard 10–20 montage at 1024 Hz, along with the electrooculogram (EOG) and the linked mastoids reference, using a BioSemi Active-Two system. The signals were band-pass filtered between 0.1 and 100 Hz, and a notch filter was applied between 49 and 51 Hz. The intervals between trials were removed and an Independent Component Analysis (ICA) was applied to the remaining data, mainly for removing ocular and muscular artifacts, and noisy channels. Artifactual components were selected using EyeCatch [1], ADJUST plugins [22] and supervised by an expert. The preprocessing of the EEG signal was performed using EEGLAB toolbox in Matlab [5]. Finally, the EEG signal was bandpass filtered between 0 - 40 Hz using a simmetric linear-phase FIR non-casual filter from MNE python library, then scaled to $[0, 1]$ and downsampled to 128 Hz for the present analysis [11].

Audio registration and preprocessing. The audio was recorded on separate channels by a TASCAM DR-100 digital recorder (at 44.1 kHz, 16 bits), using a Rode HS-1 head-mounted close-talking microphone for each participant. The recorded audio signal was later downsampled to 16 KHz and automatically synchronised with the EEG signal. This procedure was carried out by recording a low-resolution copy of the audio signal in the EEG recordings, as analogue inputs, and then finding the time offset that maximizes the cross-correlation between the two audio copies. For the present analysis, we focus on the speech envelope which was computed using scipy's implementation of the Hilbert transform followed by a 3rd order Butterworth low-pass filter of 25 Hz [26,6,24,7]. It was then downsampled taking the average in non-overlapping windows of 125 samples and scaled between $[0, 1]$; resulting in a 128 Hz signal, whose time-points matched those of the EEG. Audio and EEG intervals were semi-automatically labeled according to who was speaking/listening.

2.2 Analysis

Ridge regression. The rationale of our experiment is to predict the values of the EEG signal (*time-points*) of one participant, given the envelope of the preceding speech signal of the other. Ridge regression is particularly useful in this scenario because the independent variables in our model are highly correlated between each other. By adding a degree of bias to the regression estimates ($L2$ regularization parameter), ridge regression will shrink the model weights, therefore preventing multicollinearity. We generate and fit a model for each EEG channel independently (Fig. 1A) using sklearn implementation of Ridge regression [25,24,15,6]. As the input matrix is dense, the algorithm first applies Cholesky decomposition and then computes the regression using scipy [26].

The EEG time-points were selected in the periods where the participants were listening and not speaking (model targets). On average, we obtained $N_{timepoints} =$ 54277 per participant (range: [19403 to 74347]). For each EEG time-point, the previous 530 ms of the envelope signal (*time delays*) were taken as inputs based on previous results [15,13,7]. We excluded the 3 ms that the sound wave takes to reach the 1 m apart interlocutor by dropping the most recent envelope sample. As the sample rate from both the audio and EEG is 128 Hz, this results in an input segment of 523 ms long ($N_{delays} = 68$). These *time delays* represent the inputs of our model (X), a $N_{delays} \times N_{timpoints}$ matrix. The model weights w are obtained by minimizing the eq. 1.

$$|Y - Xw|^2 + \alpha|w|^2 \tag{1}$$

where Y is a $1 \times N_{timepoints}$ corresponding to the EEG amplitude for each time-point (Fig. 1A). The obtained w represent the importance of the information from the envelope at each time delay. The regularization parameter α was set to 1000 for all channels and participants.

The training and test set were taken from each participant separately, and the model was used in such a way that the results were first obtained for each individual, and then averaged across participants. A 5-fold cross-validation was performed without shuffling the data. Furthermore, as the only hyper-parameter in the model was set, no validation set was used.

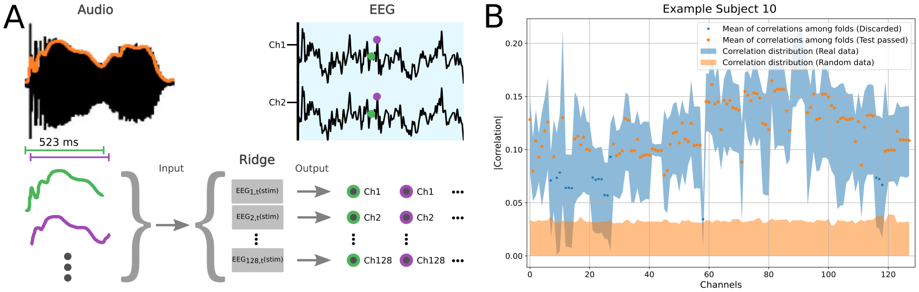

Fig. 1. A. Data analysis pipeline. **B**. Pearson correlation values per electrode for one participant. Absolute values of the correlation obtained for all folds with randomised (orange) and real (blue) stimuli. The dots show the mean correlations between folds for channels that did (orange) and did not (blue) present significant differences in all 5 folds.

Statistical tests. To validate that the information obtained from the envelope signal is relevant to predict the EEG signal, a permutation test was performed by training the model on 3000 random permutations of the input matrix. The Pearson correlation between the EEG signal and the one predicted by the model trained on real stimuli was compared with the empirical null distribution of correlations obtained with the randomized stimuli. The process was repeated for each electrode in each fold of the cross-validation procedure. The threshold p-value was set at 5% and corrected using the Bonferroni criteria ($0.05/128 \approx 0.0004$). Figure 1B shows the range of the absolute values of the correlation obtained with real stimuli for the 5 folds (blue), and the range of the absolute values of the empirical distributions for all the permutations in the 5 folds (orange), for each channel of one participant as an example. The dots correspond to the average of the real absolute values of the correlation for every channel. Only channels that were significant (p-value < 0.0004) in all folds were considered significant for each participant, and are marked as orange dots.

3 Results

Briefly, we aimed to predict the brain activity of participants engaged in dialogue from the envelope of the preceding half second of speech signal. We estimate the importance of the information from the envelope at a certain time delay as the weights of a regression fit to each channel. The weights are obtained for each participant and then averaged across participants. They presented a ramp towards 0ms with a small negative peak at 101 ms (Fig.2A). The spatial distribution of the weights at 0 and 101 ms had a similar wide frontal distribution. Importantly, the weight's curve patterns held when artificially adding delays from after 0ms. These curves were smooth due to the correlation of contiguous time-samples of the envelope, but rapidly decreased after 0ms.

The absolute values of the correlation between the real and the estimated signal presented a broad spatial distribution, with the higher values distributed

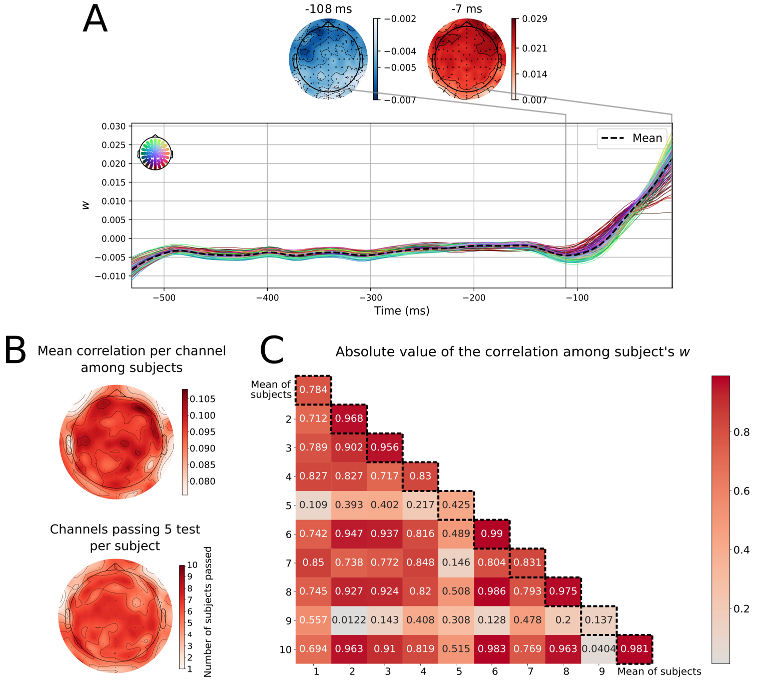

Fig. 2. A. Time-course of the weights (w) averaged across participants for all electrodes (in colours) and across electrodes (black dashed line). Spatial distributions of w at the positive and the negative peaks. **B**. Participants average of the absolute values of the correlation between the predicted and the original EEG signal (top) and the spatial distribution of the number of participants that had significant results for a given channel (bottom). **C**. Correlations between participant's average weights across channels with each other (lower triangle) and with the global average across participants (black dashed line in **A**) (diagonal).

in anterior regions (Fig. 2B, top panel). The correlation varies from 0.075 to 0.105, which is comparable with previous literature [6,7,13,15,24]. Moreover, the maximum value for a single participant resulted as high as 0.2 (averaged across folds). We calculated the number of participants in which each channel was significant (Fig. 2B, bottom). This also presented a broad distribution, with a maximum of 7 in central locations. Furthermore, ws presented quite similar patterns among participants, as shown by the high correlation values of the time-courses between them (Fig. 2C). This clearly indicates that results across participants are highly robust.

4 Conclusions

Altogether, the analysis here presented shows consistent results, supported by different measures such as the similarity across participants and the conservative criteria applied to the results, as is the Bonferroni correction among channels, where a significant number of electrodes passed the test in all folds.

Correlation values between the predicted and the real EEG signals (0.094 ± 0.019) were comparable with the values obtained in other studies for the envelope with EEG (between 0.01 and 0.1) [6,7] and ECoG (between 0.01 and 0.25) [24,15,13], which is a technique that is expected to be more sensitive and less prone to muscular artifacts. Moreover, we presented evidence that the envelope of the speech signal is encoded within a maximum delay of 100ms in the frontal regions of the brain, suggesting that its processing occurs with a latency of less than 100 ms.

These results are of great importance for understanding the precise timing in which the envelope, or other turn-taking cues, are processed in the brain in the context of a natural dialogue, which might help improve Spoken Dialog Systems, as it provides useful information concerning processing times that could be used by the system. Furthermore, it would be possible to carry out a similar analysis considering speech instead of listening time intervals, aiming to identify EEG signals when pronouncing speech with different intentions or emotions, which could find application in speech oriented Brain Computer Interfaces.

The present work paves the way for studying brain representations of speech in ecological scenarios with EEG. Future work will focus on expanding the speech feature space explored, including acoustic –as the pitch, the jitter, or the shimmer–, phonological, or pragmatic characteristics of the speech. These last set of characteristics, such as the speech act, are largely unexplored with AI methods [2,9]. Furthermore, we will aim to narrow the brain representations of these features not only in space and time, but also in the frequency domain.

Acknowledgements

We thank Victoria Peterson for her thoughtful early contributions to the project. The authors were supported by the National Science and Technology Research Council (CONICET) and the University of Buenos Aires (UBA). The research was supported by the University of Buenos Aires (UBA) (20020190100054BA), The National Agency of Promotion of Science and Technology (PICT 2018-2699) and the National Scientific and Technical Research Council (PIP 11220150100787CO).

Code availability

Encouraging reproductive and collaborative science, the code and sample data used for the present analysis is publicly available in `https://github.com/jegonza66/Speech-encoding`.

References

1. Bigdely-Shamlo, N., Kreutz-Delgado, K., Kothe, C., Makeig, S.: Eyecatch: Datamining over half a million eeg independent components to construct a fully-automated eye-component detector. In: 2013 35th Annual International Conference

of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 5845–5848. IEEE (2013)

2. Bögels, S., Levinson, S.C.: The brain behind the response: Insights into turn-taking in conversation from neuroimaging. Research on Language and Social Interaction **50**(1), 71–89 (2017)
3. Cheung, C., Hamilton, L.S., Johnson, K., Chang, E.F.: The auditory representation of speech sounds in human motor cortex. Elife **5**, e12577 (2016)
4. David, S.V., Vinje, W.E., Gallant, J.L.: Natural stimulus statistics alter the receptive field structure of v1 neurons. Journal of Neuroscience **24**(31), 6991–7006 (2004)
5. Delorme, A., Makeig, S.: Eeglab: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. Journal of neuroscience methods **134**(1), 9–21 (2004)
6. Desai, M., Holder, J., Villarreal, C., Clark, N., Hamilton, L.S.: Generalizable eeg encoding models with naturalistic audiovisual stimuli. bioRxiv (2021)
7. Di Liberto, G.M., O'Sullivan, J.A., Lalor, E.C.: Low-frequency cortical entrainment to speech reflects phoneme-level processing. Current Biology **25**(19), 2457–2465 (2015)
8. Etard, O., Reichenbach, T.: Neural speech tracking in the theta and in the delta frequency band differentially encode clarity and comprehension of speech in noise. Journal of Neuroscience **39**(29), 5750–5759 (2019)
9. Gisladottir, R.S., Bögels, S., Levinson, S.C.: Oscillatory brain responses reflect anticipation during comprehension of speech acts in spoken dialog. Frontiers in human neuroscience **12**, 34 (2018)
10. Golumbic, E.M.Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., et al.: Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". Neuron **77**(5), 980–991 (2013)
11. Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., Hämäläinen, M.S.: MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience **7**(267), 1–13 (2013). https://doi.org/10.3389/fnins.2013.00267
12. Gravano, A., Hirschberg, J.: Turn-taking cues in task-oriented dialogue. Computer Speech & Language **25**(3), 601–634 (2011)
13. Hamilton, L.S., Edwards, E., Chang, E.F.: A spatial map of onset and sustained responses to speech in the human superior temporal gyrus. Current Biology **28**(12), 1860–1871 (2018)
14. Hamilton, L.S., Huth, A.G.: The revolution will not be controlled: natural stimuli in speech neuroscience. Language, Cognition and Neuroscience **35**(5), 573–582 (2020)
15. Hamilton, L.S., Oganian, Y., Chang, E.F.: Topography of speech-related acoustic and phonological feature encoding throughout the human core and parabelt auditory cortex. BioRxiv (2020)
16. de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., Theunissen, F.E.: The hierarchical cortical organization of human speech processing. Journal of Neuroscience **37**(27), 6539–6557 (2017)
17. Holdgraf, C.R., Rieger, J.W., Micheli, C., Martin, S., Knight, R.T., Theunissen, F.E.: Encoding and decoding models in cognitive electrophysiology. Frontiers in systems neuroscience **11**, 61 (2017)
18. Hullett, P.W., Hamilton, L.S., Mesgarani, N., Schreiner, C.E., Chang, E.F.: Human superior temporal gyrus organization of spectrotemporal modulation tuning derived from speech stimuli. Journal of Neuroscience **36**(6), 2014–2026 (2016)

19. Huth, A.G., De Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L.: Natural speech reveals the semantic maps that tile human cerebral cortex. Nature **532**(7600), 453–458 (2016)
20. Lerner, Y., Honey, C.J., Silbert, L.J., Hasson, U.: Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. Journal of Neuroscience **31**(8), 2906–2915 (2011)
21. Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F.: Phonetic feature encoding in human superior temporal gyrus. Science **343**(6174), 1006–1010 (2014)
22. Mognon, A., Jovicich, J., Bruzzone, L., Buiatti, M.: Adjust: An automatic eeg artifact detector based on the joint use of spatial and temporal features. Psychophysiology **48**(2), 229–240 (2011)
23. Moses, D.A., Metzger, S.L., Liu, J.R., Anumanchipalli, G.K., Makin, J.G., Sun, P.F., Chartier, J., Dougherty, M.E., Liu, P.M., Abrams, G.M., et al.: Neuroprosthesis for decoding speech in a paralyzed person with anarthria. New England Journal of Medicine **385**(3), 217–227 (2021)
24. Oganian, Y., Chang, E.F.: A speech envelope landmark for syllable encoding in human superior temporal gyrus. Science advances **5**(11), eaay6279 (2019)
25. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
26. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C.J., Polat, İ., Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A.H., Pedregosa, F., van Mulbregt, P., SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. Nature Methods **17**, 261–272 (2020). https://doi.org/10.1038/s41592-019-0686-2
27. Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., Mitchell, T.: Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. PloS one **9**(11), e112575 (2014)
28. Wolpaw, J.R., Birbaumer, N., McFarland, D.J., Pfurtscheller, G., Vaughan, T.M.: Brain–computer interfaces for communication and control. Clinical neurophysiology **113**(6), 767–791 (2002)