Spectral Encoding Profiles for Multilevel Linguistic Predictions

Clément Sauvage, Pierre Guilleminot, Benjamin Morillon
(clement.sauvage@univ-amu.fr)

Institut de Neurosciences des Systèmes,
Dynamic of Communication and Auditory Processes team,
Hôpital de la Timone,
Institut National de la Santé et de la Recherche Médicale,
Marseille 13010 France

Abstract

Speech processing is believed to rely on two types of information. First, the predictions, which are endogenous and flowing down the cortical hierarchy in a top-down manner, and second, the prediction errors, computed as the difference between the effective and the predicted inputs at each stage of the said hierarchy, in a bottom-up flow of information. The putative role of neural oscillations to mediate those signals is still a question of debate. Here we recorded intracranial EEG activity of 45 epileptic patients, while they listened to ecological speech. We used a Large Language Model to extract proxies of both prediction uncertainty and prediction errors at three linguistic levels, phonemes, syllables and words. We found that encoding of prediction errors and prediction uncertainty peaks respectively in high-gamma and beta bands, only in Primary Auditory Cortex and Superior Temporal Gyrus.

Keywords: humans; neuroimaging; speech; audition; predictive coding; large language models

Introduction

Language processing involves predicting content at multiple timescales, encompassing phonemes, words, up to sentences (Donhauser & Baillet, 2020; Heilbron et al., 2022). Hierarchical Predictive Coding suggests that the brain generates predictions and estimates prediction errors at each level of the processing hierarchy. Advances in electrophysiology and Artificial Neural Networks have provided detailed insights into neural representations, from low-level acoustic features (Yi et al., 2019) to semantic representations (Gallant et al., 2016; Caucheteux et al., 2023).

Despite these advancements, the mechanisms by which predictions and prediction errors are mediated throughout the cortical hierarchy during speech processing, and how they are encoded within brain activity, remain unresolved. In macaque visual processing, studies have attributed gamma (>40Hz) and alpha-beta (12-35Hz) frequency bands to prediction errors (feedforward signals) and predictions (feedback signals), respectively (Bastos et al., 2015). We aim to investigate whether this framework applies to ecological speech processing in humans.

Methods and Results

Acquisition. Intracortical EEG data were acquired in Hôpital de La Timone's department of epileptology. The locations of the electrode implantations were determined solely on clinical grounds. They listened to a 10' story in French. The pool of N=45 subjects totalized 6654 contacts-pairs, whose activity was downsampled to 100Hz.

Feature engineering. Textual transcription of the audio stimulus has been fed into CamemBERT (Martin et al., 2019), a LLM trained on French, and probability distributions over tokens were extracted. We first approximated the prediction errors as the average surprisal of tokens composing each word, and prediction uncertainty as the mean entropy of the probability distributions. Then, phonemic and syllabic surprisal and entropy were derived using a combination of Lexique (New et al., 2001) (repertoire of french words) and Cohort model (Marslen-Wilson & Welsh, 1978). This constituted our set of 6 linguistic predictive features : (word, phoneme, syllable) x (surprisal, entropy).

Analysis. We used Temporal Response Function (TRF) to estimate the variance of the neural data explained by the features of interest via a linear ridge regression (Crosse et al., 2016). Control acoustic features encompassed the envelope, the absolute value of its derivative, the spectral flux, the f0 envelope, and the word, syllable and phoneme onsets.

Statistics. We first selected responsive channels by computing the absolute performance of a first model with control + interest features, and a second model with control features only. This yielded a relative performance (Δr^2), that was compared with the chance distribution of relative scores estimated through permutations of the features of interest. Only the channels whose Δr^2 was above the 95th percentile in the chance distribution were kept. To statistically compare the strength of encoding between features of interest, we used a Wilcoxon-rank signed test on the model's relative scores at each frequency, p-values were corrected against false discoveries (FDR correction).

Broadband encoding. Permutation tests allowed us to isolate channels whose broadband activity encodes features of interest. We separated two 'clusters' of channels based on the set of features that drives the reconstruction accuracy: the first, named *'Acoustic',* is mostly encoding control acoustic features, and is located in bilateral Heschl's gyri and in the left Superior Temporal Gyrus. The second cluster, named *'Linguistic',* whose response is relatively more driven by the response to features of interest, is distributed, and encompasses notably the left Medial Temporal Gyrus/Sulcus, Temporal Base and Inferior Frontal Gyrus.

Frequency-resolved encoding. As in the broadband procedure encoding, we used a TRF model to reconstruct neural activity. In this case, the analysis was performed 40 times, in order to estimate the amplitude of filtered neural data within 40 frequency bands ranging from 0.1 up until 150Hz. Amplitude was extracted using Hilbert transform. This procedure led to 1 encoding score by tested frequency band and by set of regressors. By subtracting the frequency-resolved performance of the base model, we obtained the relative reconstruction accuracies : the Spectral Encoding Profiles. The SEP of either the three linguistic levels of speech (phoneme, syllable, word), or the two types of linguistic predictive features (entropy, surprise) were compared. We observed that: 1. Phoneme is the more strongly encoded linguistic level, and 2. The SEPs are similar across linguistic levels, with encoding peaking at low (<12 Hz) and high-gamma (~60–85 Hz) frequencies. In the *'Acoustic'* cluster, the distinction proposed in (Bastos et al., 2015) seems to hold, as we observed stronger encoding for surprise in the high-gamma (~60–85 Hz; and low-delta, ~1Hz) band) and stronger encoding of entropy in the alpha-beta band (~13–16 Hz; p < .05, FDR-corrected). In the *'Linguistic'* cluster, no relevant distinction appears to be encoded in the power of the neural data.

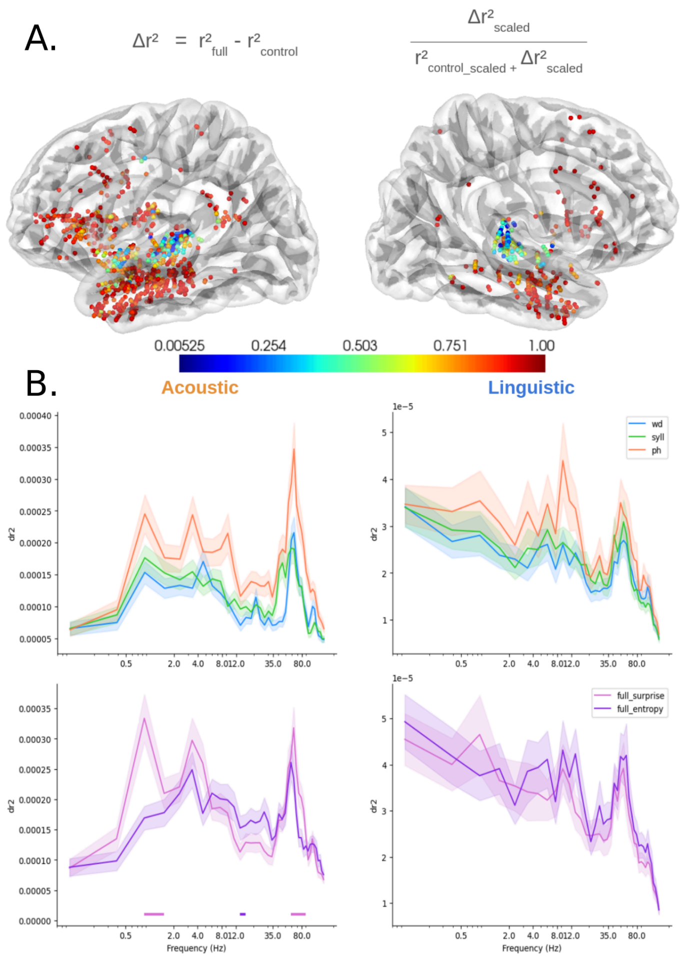

Figure 1: Channels encoding linguistic predictive features at any level, and their SEPs split between *'Acoustic'* (left) and *'Linguistic'* (right) clusters. A. Repartition of proportions of scaled variance explained by predictive linguistic features. All represented channels significantly responsive to the aforementioned features. B. Top : SEPs separated for each linguistic level (regressor set = surprisal + entropy). Bottom : SEPs separated between surprisal and entropy (regressor set = word + syllable + phoneme).

Discussion

We first observed that the encoding pattern was similar across the spectrum for each linguistic level, the phoneme being the most strongly encoded. Second, the distinction between a high-level prediction uncertainty and a lower-level prediction error encoded in distinct frequency bands, as proposed by (Bastos et al., 2015) was observed only in the *'Acoustic'* cluster of electrodes.

Acknowledgements

Funded by the European Union (ERC, SPEEDY, ERC-CoG-101043344) and Fondation Pour l'Audition (FPA RD-2022-09). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them.

References

Donhauser, P. W., & Baillet, S. (2020). Two distinct neural timescales for predictive speech processing. *Neuron*, *105*(2), 385-393.

Yi, H. G., Leonard, M. K., & Chang, E. F. (2019). The encoding of speech sounds in the superior temporal gyrus. *Neuron*, *102*(6), 1096-1110.

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & De Lange, F. P. (2022). A hierarchy of linguistic predictions during natural language comprehension. *Proceedings of the National Academy of Sciences*, *119*(32), e2201968119.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453-458.

Gwilliams, L., Marantz, A., Poeppel, D., & King, J. R. (2024). Top-down information shapes lexical processing when listening to continuous speech. *Language, Cognition and Neuroscience*, *39*(8), 1045-1058.

Caucheteux, C., Gramfort, A., & King, J. R. (2023). Evidence of a predictive coding hierarchy in the human brain listening to speech. *Nature human behaviour*, *7*(3), 430-441.

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J. M., Oostenveld, R., Dowdall, J. R., ... & Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*, *85*(2), 390-401.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de La Clergerie, É. V., ... & Sagot, B. (2019). CamemBERT: a tasty French language model. *arXiv preprint arXiv:1911.03894*.

New, B., Pallier, C., Ferrand, L., & Matos, R. (2001). Une base de données lexicales du français contemporain sur internet: LEXIQUE™//A lexical database for contemporary french: LEXIQUE™. *L'année psychologique*, *101*(3), 447-462.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive psychology*, *10*(1), 29-63.

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Frontiers in human neuroscience*, *10*, 604.