Linear-nonlinear Bernoulli modeling for quantifying temporal coding of phonemes in brain responses to continuous speech

Nathaniel J. Zuk (zukn@tcd.ie)
Department of Electronic & Electrical Engineering, Trinity College Dublin
Dublin 2, Ireland

Giovanni M. Di Liberto (diliberg@tcd.ie)
Laboratoire des Systèmes Perceptifs, UMR 8248, Département d'Etudes Cognitives, École Normale Supérierue
PSL University, France 750075

Edmund C. Lalor (edmund.lalor@urmc.rochester.edu)
Department of Biomedical Engineering, University of Rochester
Rochester, New York, USA 14627

Abstract:

The electroencephalographic (EEG) response to a sound of interest is often quantified by averaging time-locked signals over many repetitions in order to get an event-related potential (ERP). While this technique can identify an average response, it does not easily allow one to validate the robustness of that response nor variation of the response over repetitions of the sound. Here, we extend the ERP technique as a linear-nonlinear Bernoulli (LNB) model, inspired by neural models, in order to develop a framework for decoding the timing of stimulus events. We use this technique to analyze EEG recordings during presentations of continuous speech and examine neural responses to phonemes, which have been shown to have characteristic EEG responses. Pattern analysis of the confusion between phonemes separates phonemes into vowel and constants, indicating separate ERPs that can robustly predict these phoneme classes. We also find that vowels are decoded more accurately than consonants, and the time course of vowel predictability tracks the rhythm of vowels, while consonant predictability does not track the rhythm of consonants. Overall, we demonstrate a specific instance in which a linear-nonlinear Bernoulli modeling framework can be used to compare ERPs and quantify the ability to decode stimulus events from EEG.

Keywords: EEG, ERP, auditory, neural decoding, speech

Introduction

To understand how the brain responds to sound, a common technique used for electroencephalography (EEG) is to present a sound numerous times and average the evoked response over many repetitions (Sur & Sinha, 2009). This technique to get an event related potential (ERP) presumes that the neural responses are consistent across each presentation of the sound, but the response could vary over time due to changes in cognitive state or due to adaptation (Näätänen, 1982). Quantifying this change over time, however, is difficult using current ERP-based methods.

More recently, linear modeling has been used to identify a neural response to multiple types of events during the presentation of a continuous stimulus, such as phonemes or words in speech (Di Liberto et al, 2015; Brodbeck et al, 2018). This framework has a benefit of allowing researchers to evaluate the model by quantifying its ability to predict EEG. Still, it becomes difficult to evaluate the contribution of individual components or identify redundant information in these models, and it is even similarly difficult to determine how these contributions might change over time. Additionally, regularization, which is often necessary to prevent overfitting, could normalize the relative contribution of different events in the model, making cross-component comparisons difficult.

Advances in modeling spikes from single-cell neural recordings could resolve some of these issues in EEG (Schwartz et al, 2006; Meyer et al, 2017). In one basic form of analysis, the input signal is averaged over spike times in order to get a template for the input signal that evokes a spike. This template is then used to map the linear dot-product between the input and the template into a probability of spiking using a nonlinear transformation that depends upon the event probability distribution. If the event probability is assumed to be Bernoulli-distributed, together this model known as a linear-nonlinear Bernoulli (LNB) model.

Here, we turn the ERP into an LNB model that can be used to quantify the temporal encoding of events. Unlike spike models, the ERP is used as a template to quantify the probability of stimulus events based on the

correlation between the template and the ongoing EEG. We then apply this technique to construct a separate LNB model for each phoneme in continuous speech, and then use the model to reconstruct the timing of the trained phoneme as well as other untrained phonemes that may share a similar neural response. We show that pattern analysis of cross-phoneme reconstruction accuracies allows us to identify phoneme characteristics that evoke unique responses in the brain.

Methods

EEG collection and preprocessing

During EEG recording, 10 subjects listened to an audiobook over headphones, where the audiobook was split into 29 trials, each about 155 s in length (Di Liberto et al, 2015). The EEG was referenced to the mastoids, downsampled to 128 Hz, and then filtered between 1-15 Hz.

Phoneme decoding

First, the ERP was computed for each phoneme, and a predicted EEG signal was created by placing the phoneme-specific ERP at each phoneme onset. ERPs were calculated at delays 0-300 ms post-phoneme. Then, to reduce the dimensionality of the EEG data, canonical correlation analysis (CCA) was used to identify two canonical components in the recorded EEG that maximize the correlation between the EEG and the predicted signal, in order to optimize decoding accuracy. Lastly, the LNB model was computed with ridge regression and optimized with 10-fold cross validation to maximize the likelihood of the observed timing of phonemes given the fitted model. A logistic function was used to map the dot-product between the ERP and the ongoing EEG into a Bernoulli event probability, using *glmfit* in Matlab.

Speech phonemes were identified identically to Di Liberto et al, 2015. The onset times of the phonemes were used to label event times for the phoneme-specific ERPs. The LNB model was created separately for each phoneme, and then tested on all phonemes for each left out trial. The ability to reconstruct the timing of each phoneme was quantified using an "adjusted log-likelihood" which was equal to the log-likelihood of the model's prediction minus the log-likelihood of a model that only represents the average event probability of the phoneme. This adjustment was necessary to make phoneme decoding accuracy comparable across phonemes with different frequencies.

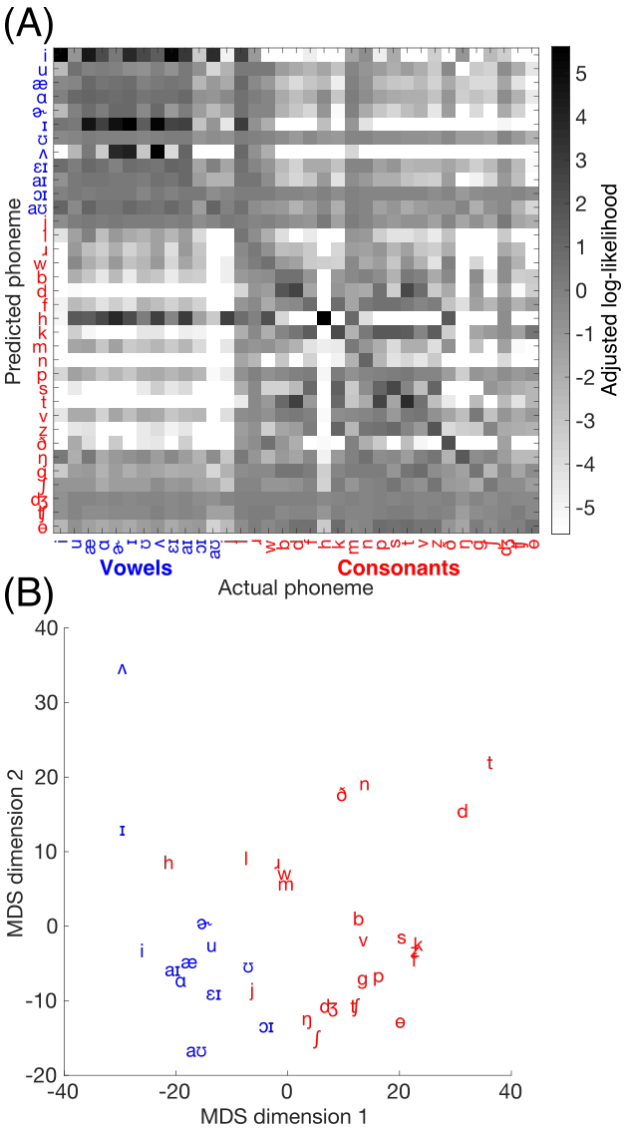

Figure 1: (A) Adjusted log-likelihoods of the "Predicted Phoneme" trained on the "Actual Phonemes" for an example subject. The log-likelihoods were adjusted by subtracting the log-likelihood of a model that represents the average probability of the phoneme with no information about the EEG. (B) Multidimensional scaling (MDS) of the log-likelihood matrix (as in A) averaged across all subjects. Vowels are labeled in blue, and consonants are labeled in red. MDS highlights a separation of vowels and consonants, indicating the confusion within each class, and a vowel-consonant model may more optimally capture the neural responses to phonemes.

Results

Phoneme-specific LNB models could reconstruct onset times for its target phoneme as well as the onset times of other phonemes. Specifically, **Figure 1A** shows that models fit for specific vowels and diphthongs are more predictive of other phonemes within the same categories than the consonants. Multi-dimensional scaling of the averaged reconstruction matrix across subjects indicated that vowels and consonants were fairly well separated (F-test of phoneme class separation: F = 22.96, p << 0.001) **(Figure 1B)**. This suggests that a model that predicts consonants and vowels rather than individual phonemes may more optimally capture the neural responses to phonemes in the brain.

We then created LNB models that predicted vowels and consonants separately. Vowel and consonant reconstructions were significantly better than chance (Wilcoxon signed-rank, vowels: z = 14.00, p << 0.001; consonants: z = 12.50, p << 0.001). Additionally, vowels were reconstructed significantly better than consonants (Wilcoxon signed-rank: z = 13.70, p << 0.001).

A closer examination of the time-varying event probabilities reconstructed by the two models show signals that fluctuate with opposite polarities: increases

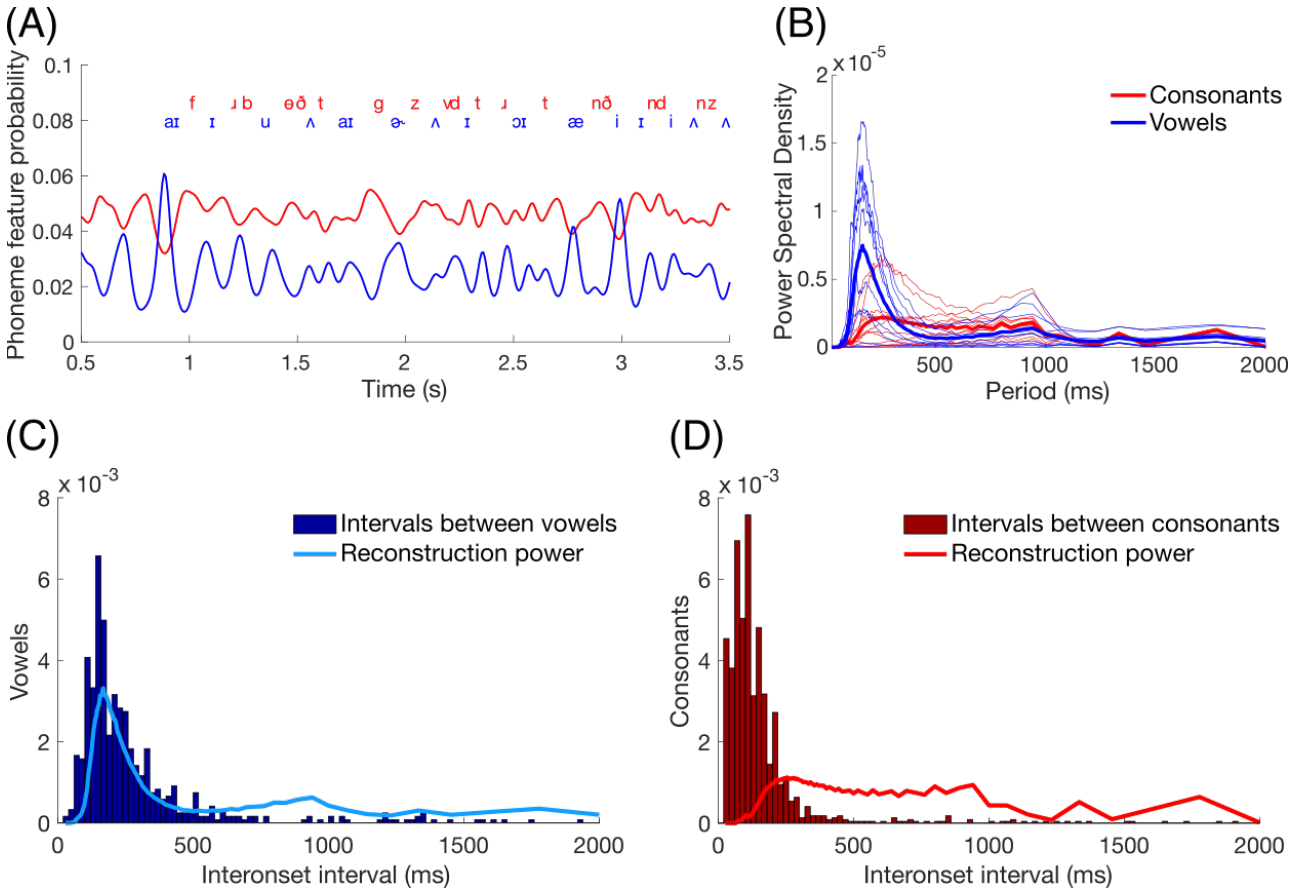

Figure 2: (A) Reconstruction of vowel (blue) and consonant (red) probabilities for one example trial in one subject. The actual phonemes are indicated above in blue for vowels and red for consonants. Note that the vowel reconstruction appears to fluctuate at a regular frequency, suggesting an optimal frequency at which the brain is tracking vowels. (B) Power spectral density of the reconstructions for vowels and consonants, using the colors indicated in A. The power is plotted as a function of the period of the oscillation. Each thin line is the power averaged across trials for one subject. The thick lines indicate the average across subjects. (C) Interonset interval histogram of vowels across all trials (dark blue), overlaid with the average power spectral density of the reconstruction across subjects, as in B (light blue). Both have been normalized by their areas between 30 and 2000 ms. The power of the reconstructions captures the peak interonset interval for the vowels. (D) Interonset interval histogram (dark red) and average power across subjects (light red) for consonants. Unlike for vowels, the consonants reconstructions do not capture the regularity in consonant intervals.

in the probability of a vowel occur when the probability of a consonant decreases **(Figure 2A)**. Note that there were no model constraints linking consonant and vowel probabilities because the two models were fit separately. It is also apparent that the fluctuations are larger for vowels than consonants, which could relate to the improved reconstruction accuracy for vowels compared to consonants.

By analyzing the spectrum of the time-varying probabilities, we found that the probability of vowels fluctuate around 6 Hz **(Figure 2B)**. In contrast, the spectrum for the consonants is much flatter, indicating less regularity. Moreover, the peak of the spectrum for the vowel reconstructions matches the peak in the interonset interval distribution for vowels, while consonants have less overlap despite having a similar, albeit shorter, peak interonset interval **(Figure 2C,D)**. This suggests that the primary signal being captured by the phoneme ERP model may also be a signal relevant for capturing syllables and speech rhythm (Oganian & Chang, 2018; Anumanchipalli et al, 2019).

Conclusion

With an LNB framework for representing evoked responses in EEG, we have shown that neural responses to multiple event types can be compared and reduced to event-related classes. Furthermore, analyzing the time-varying probability of phonemes revealed the stronger and more regular encoding of vowels in continuous speech than consonants. This framework can be extended further in the future by quantifying nonlinear effects of event history on evoked responses, or by using non-monotonic nonlinearities, in order to capture more information about the time course of evoked responses in the EEG that would not be readily captured with the typical ERP approach.

Acknowledgments

This work was supported by an SFI Career Development Award (CDA/15/3316) and by the Del Monte Institute for Neuroscience at the University of Rochester. GDL was supported by the EU H2020-ICT grant 644732 (COCOHA). The authors would also like to acknowledge the Neuromorphic Engineering Workshop, where early stages of the LNB model for EEG were first developed.

References

Anumanchipalli, G.K., Chartier, J., & Chang, E.F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature*, 568, 493-498.

Brodbeck, C, Hong, L.E., & Simon, J.Z. (2018). Rapid transformation of auditory to linguistic representations of continuous speech. *Current Biology*, 28, 3976-3983.e5.

Di Liberto, G.M., O'Sullivan, J.A., & Lalor, E.C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25, 2457-2465.

Oganian, Y. & Chang, E.F. (2018). A speech envelope landmark for syllable encoding in human superior temporal gyrus. *bioRxiv*. doi: 10.1101/388280

Meyer, A.F., Williamson, R.S., Linden, J.F., & Sahani, M. (2017). Models of stimulus-response functions: elaboration, estimation, and evaluation. *Frontiers in Systems Neuroscience*, 10:109. doi: 10.3389/fnsys.2016.00109

Näätänen, R. (1982). Processing negativity: an evoked-potential reflection of selective attention. *Psychological Bulletin*, 92, 605-640.

Schwartz, O., Pillow, J.W., Rust, N.C., & Simoncelli, E.P. (2006). Spike-triggered neural characterization. *Journal of Vision*, 6, 484-507.

Sur, S. & Sinha, V.K. (2009). Event-related potential: an overview. *Industrial Psychiatry Journal*, 18, 70-73.