2pPP13

Decoding speech envelopes from EEGs: A comparison of regularized linear regression and long short-term memory deep neural network

Zhe-chen Guo[1], Kevin Pangottil[2], Bharath Chandrasekaran[3], and Fernando Llanos[1]

[1]*Department of Linguistics and* [2]*Department of Computer Science, University of Texas at Austin*

[3]*Department of Communication Science and Disorders, University of Pittsburgh*

1. Introduction

- ❑ Slow temporal modulation (speech envelope) is critical for perceiving vowels and consonants [1].
- ❑ Neurons in the auditory cortex phase-lock to changes in the speech envelope [2–4].
- ❑ The neural encoding of speech sounds can be assessed by comparing neural (e.g., EEG) and envelope oscillations [5–8].
  - Often using linear decoders such as **multivariate temporal response functions (mTRF)** [7].
  - However, the correlation is usually low (< 0.1).
- ❑ Recent work has shown success in using non-linear decoders such as **long short-term memory deep neural networks (LSTMs)**
  - For classification (e.g., whether the EEG matches the envelope: [9]) and regression (e.g., decoding envelopes of short, isolated sentences: [10]) problems.

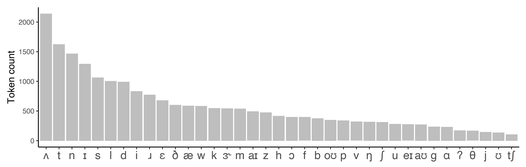
Research questions

1. Does an LSTM model outperform the mTRF in decoding speech envelopes from natural running speech?
2. To what extent is the decoding performance of the LSTM and mTRF models consistent across subjects (listeners) and across phoneme categories?

2. Brain-to-speech decoder

EEG Dataset

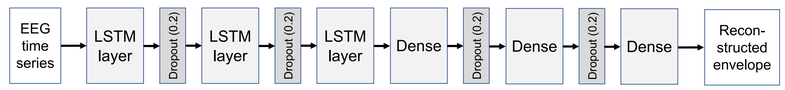
- ❑ Mastoid-referenced EEG recorded at 62 electrodes from Reetzke et al. [11].
- ❑ 15 native English speakers listening attentively to an audiobook.
- ❑ 300-ms segments of EEG and speech envelope from each phoneme onset
- ❑ 37 phonemes (21,547 tokens in total)



- 323,205 data points (21,547 × 15 subjects)

Envelope Reconstruction

- ❑ Performed subject-dependent decoding for each phoneme using 10-fold cross-validation.
- ❑ **mTRF**
  - Backward model using MATLAB mTRF toolbox [7]
  - Lags: 0 to 300 ms; tuned ridge parameter with 10% of training data
- ❑ **LSTM**



  - Trained for 150 epochs with LSTM hidden size = 256, learning rate = 10^{-4}, batch size = 16, Adam optimizer, and MSE loss
- ❑ **Pearson correlation (*r*)** between target and reconstructed envelopes served as the metric of reconstruction accuracy.

3. Results

Fig 1. Average Pearson correlations (*r*) between target and reconstructed envelopes by phoneme and model (LSTM vs. mTRF) with bars representing 95% confidence intervals.

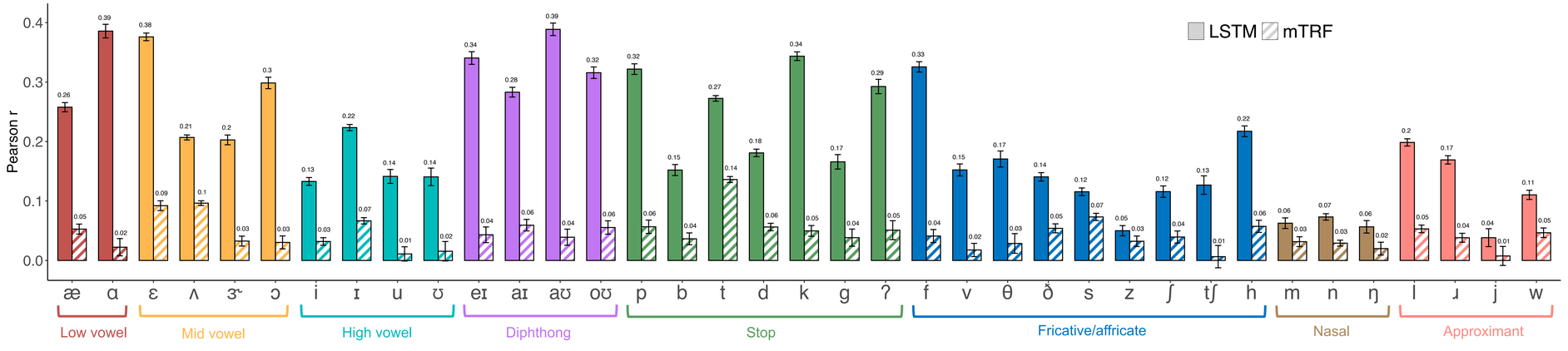


Fig 2. Average *r* values of the three most frequent vowels (/ʌ, i, ɪ/) and consonants (/t, s, n/) by subject and model.

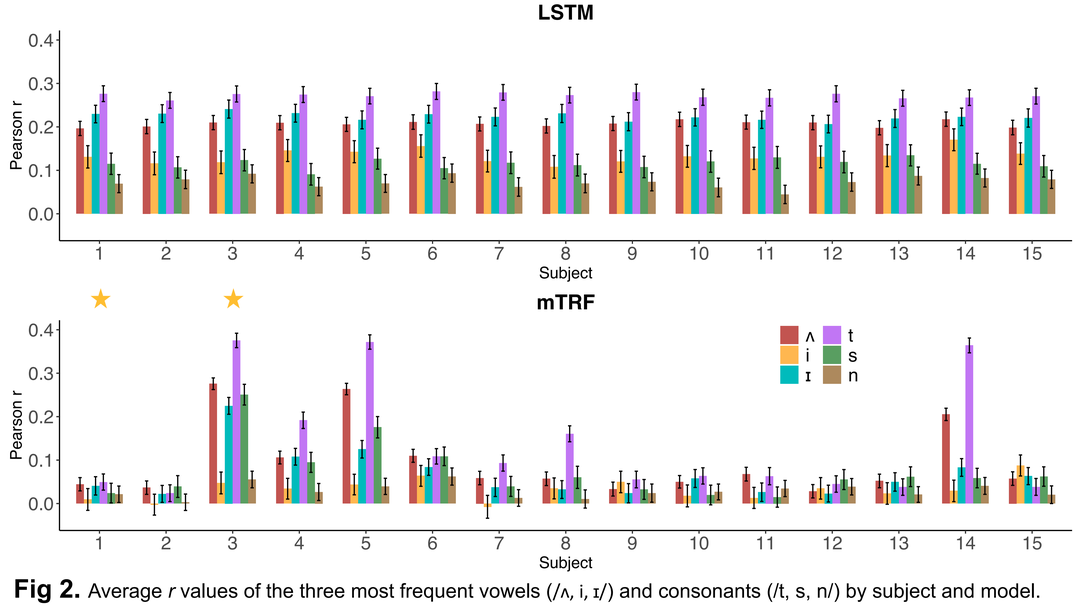
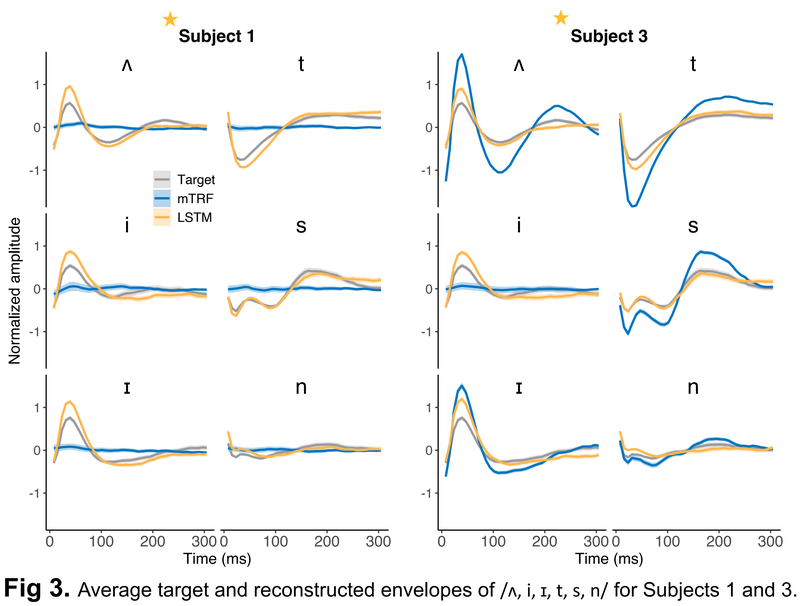


Fig 3. Average target and reconstructed envelopes of /ʌ, i, ɪ, t, s, n/ for Subjects 1 and 3.



4. Discussion

- ❑ For all phonemes, the LSTM model (mean r = 0.20) significantly (p < 0.05) outperformed the mTRF (mean r = 0.06).
- ❑ Reconstruction accuracy was equally high and much less variable across subjects for the LSTM model.
  - The mTRF could reconstruct envelopes with accuracy similar to that of the LSTM, but only for certain subjects (e.g., Subject 3).
  - Decoding performance of linear approaches can be highly subject-dependent.
- ❑ Decoding accuracy of the LSTM model was lower for high vowels compared with other vowels and nasals compared with other consonants.
  - Possibly due to their lower amplitude.
- ❑ The findings demonstrate the potential of non-linear approaches to investigating the neural representation of speech envelopes and developing brain-computer interfaces.
- ❑ Future research will experiment with frequency-domain features (e.g., EEG spectrograms) and different model architectures.

References

➤**[1]** Rosen, S. (1992). Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *336*(1278), 367–373. ➤**[2]** Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, *54*(6), 1001–1010. ➤**[3]** Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences*, *98*(23), 13367–13372. ➤**[4]** 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. ➤**[5]** 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*(19), 2457–2465. ➤**[6]** Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859. ➤**[7]** 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. ➤**[8]** Vanthornhout, J., Decruy, L., Wouters, J., Simon, J. Z., & Francart, T. (2018). Speech intelligibility predicted from neural entrainment of the speech envelope. *Journal of the Association for Research in Otolaryngology*, *19*, 181–191. ➤**[9]** Monesi, M. J., Accou, B., Montoya-Martinez, J., Francart, T., & Van Hamme, H. (2020). An LSTM based architecture to relate speech stimulus to EEG. In *IEEE International Conference on Acoustics, Speech and Signal Processing* (ICASSP) (pp. 941-945). IEEE. ➤**[10]** Dash, D., Ferrari, P., Berstis, K., & Wang, J. (2021). Imagined, intended, and spoken speech envelope synthesis from neuromagnetic signals. In *Proceedings of Speech and Computer: 23rd International Conference (SPECOM 2021* (pp. 134-145). Springer International Publishing. ➤**[11]** Reetzke, R., Gnanateja, G. N., & Chandrasekaran, B. (2021). Neural tracking of the speech envelope is differentially modulated by attention and language experience. *Brain and Language*, *213*, 104891.