Tracking phoneme processing during continuous speech perception with MEG

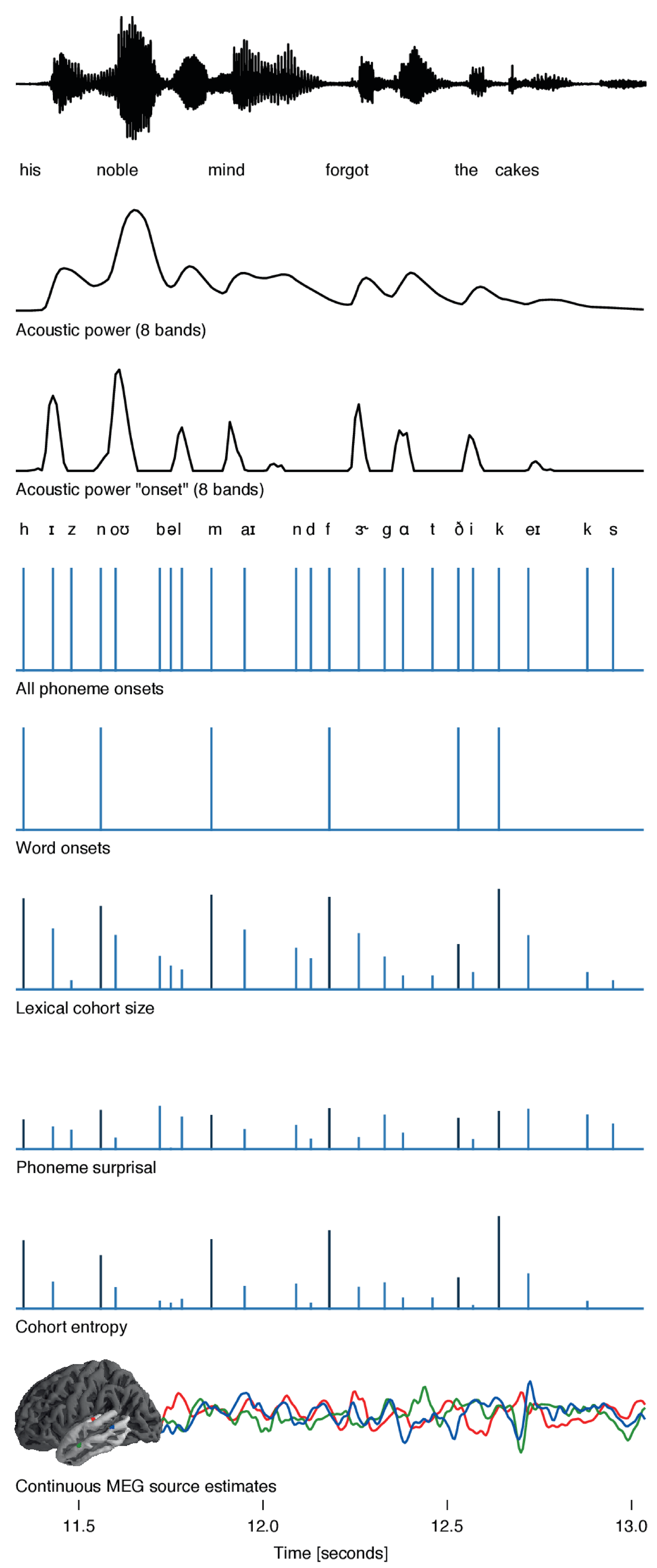
Christian Brodbeck*[1] & Jonathan Z. Simon[1,2,3]

[1]Institute for Systems Research, [2]Department of Electrical and Computer Engineering, [3]Department of Biology
University of Maryland, College Park, Maryland; *brodbeck@umd.edu

Computational Sensorimotor Systems Lab

Introduction

- Aim: characterizing how information from phonemes is integrated for word perception in continuous speech comprehension
- Phonemes represent the continuous acoustic speech signal with discrete linguistic categories. However, brain responses to phoneme identity (/a/, /ε/, /θ/, ...) are hard to dissociate from acoustic responses because each phoneme is associated with a characteristic acoustic pattern
- Phonemes incrementally provide information about spoken words (e.g. Norris and McQueen, 2008); information theoretic measures like phoneme surprisal and lexical cohort entropy influence behavioral and MEG responses to isolated word stimuli (e.g. Gaston and Marantz, 2017)
- Here we analyze MEG responses to phoneme information properties in continuous, uninterrupted speech to determine how phonemes are processed as linguistically relevant stimuli



Predictor variables

- **Acoustic spectrogram**: acoustic power in 8 logarithmically spaced bands
- **Acoustic "onset"**: rising slope of acoustic power in the same bands
- **Cohort size**: number of word forms compatible with the current prefix
- **Cohort reduction**: number of words that the current phoneme excludes
- **Phoneme surprisal**: inverse of the conditional probability of the phoneme
- **Cohort entropy**: degree of uncertainty about the current word
- To account for the possibility that the first phoneme of each word is processed differently (Marslen-Wilson, 1987), word onset was modeled separately from the subsequent phonemes for each variable

Stimuli

- **Solo**: one minute long audiobook segments
- **Two-speaker mix**: two audiobook segments mixed at equal loudness, task to attend to one while ignoring the other

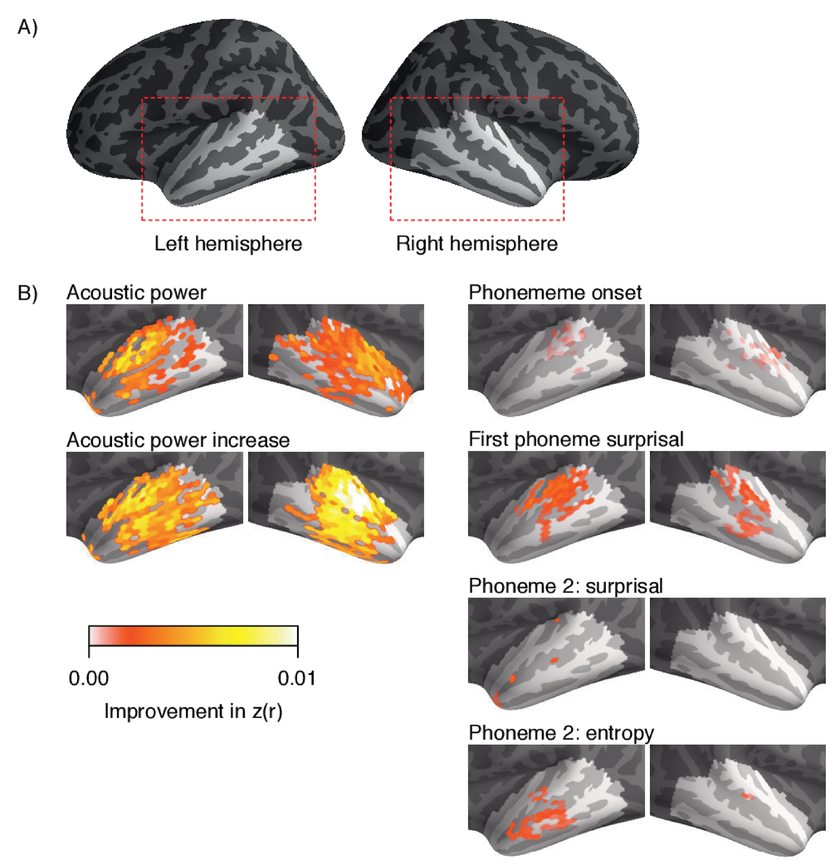
Analysis method

Linear kernel estimation predicts source localized continuous MEG responses from multiple concurrent predictor variables; predictors compete to explain variance (Brodbeck et al., 2018).

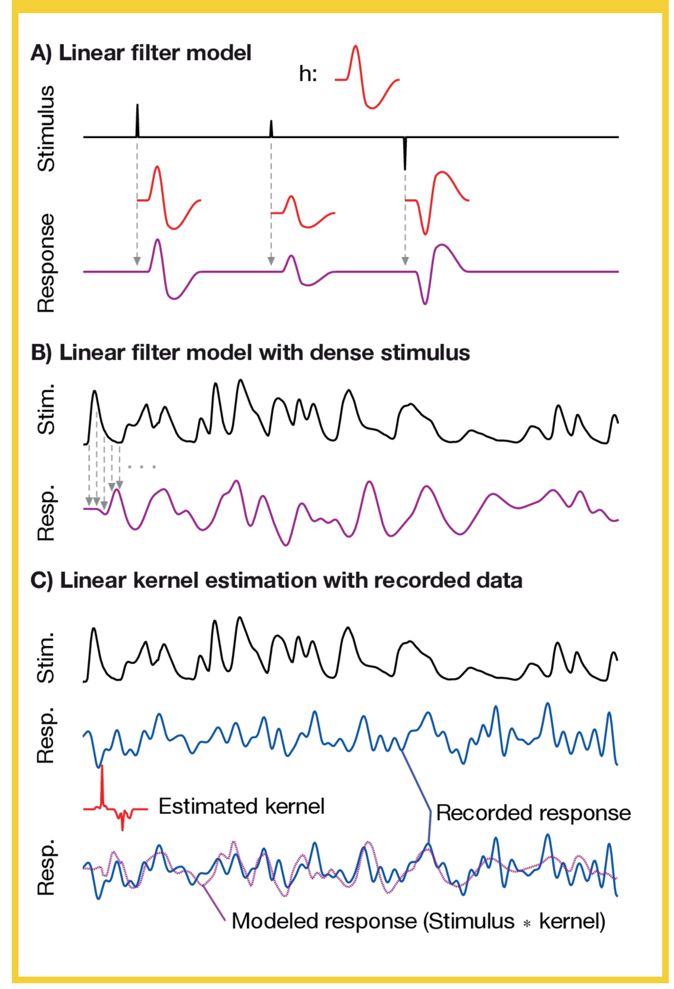
Results: single speaker

Responses to single speaker stimuli were modeled by iteratively excluding the least significant predictor until all remaining predictors were significant:

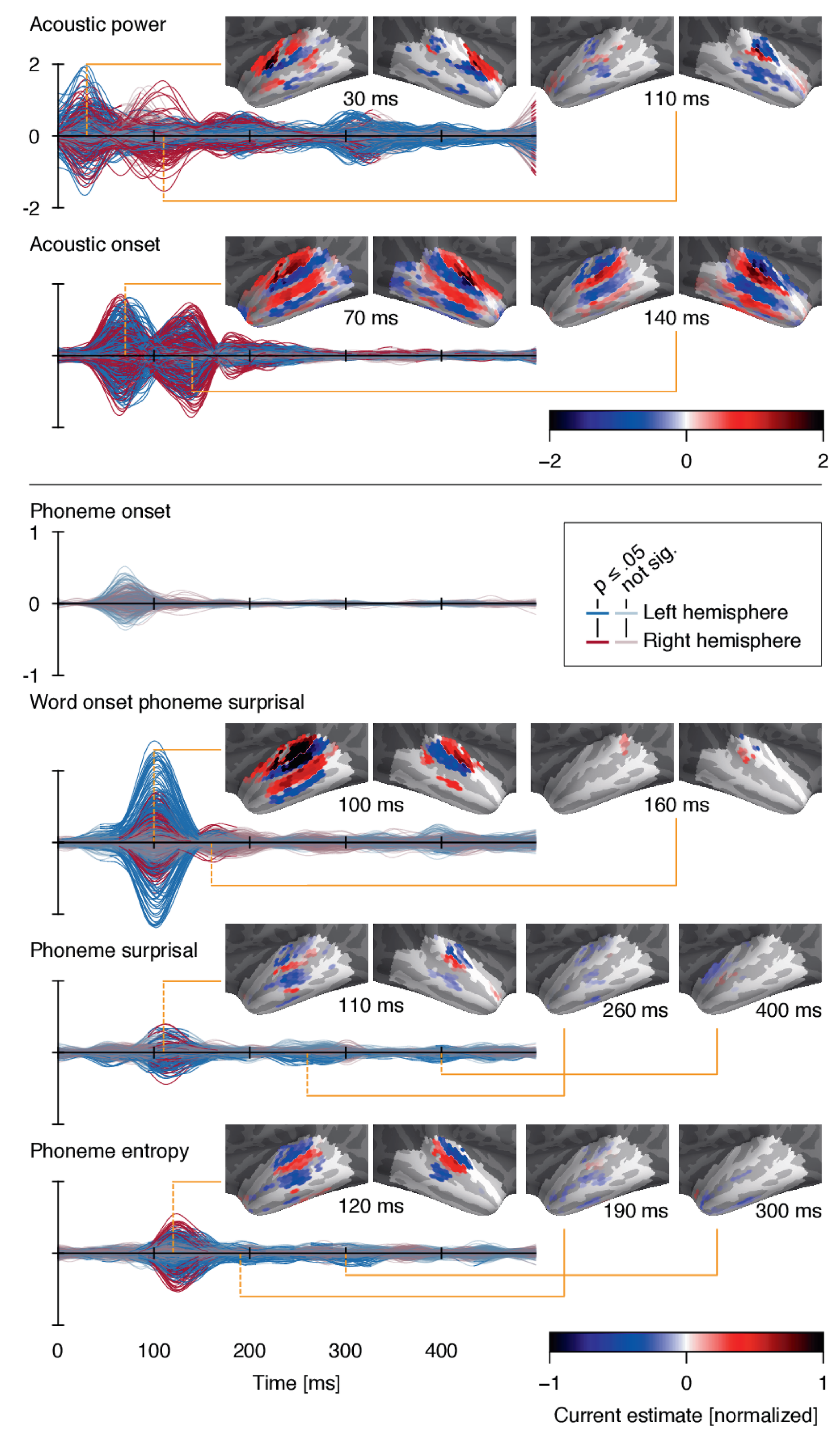
- Bilateral acoustic responses
- Responses to phoneme information more dominant in the left hemisphere



Method: kernel estimation



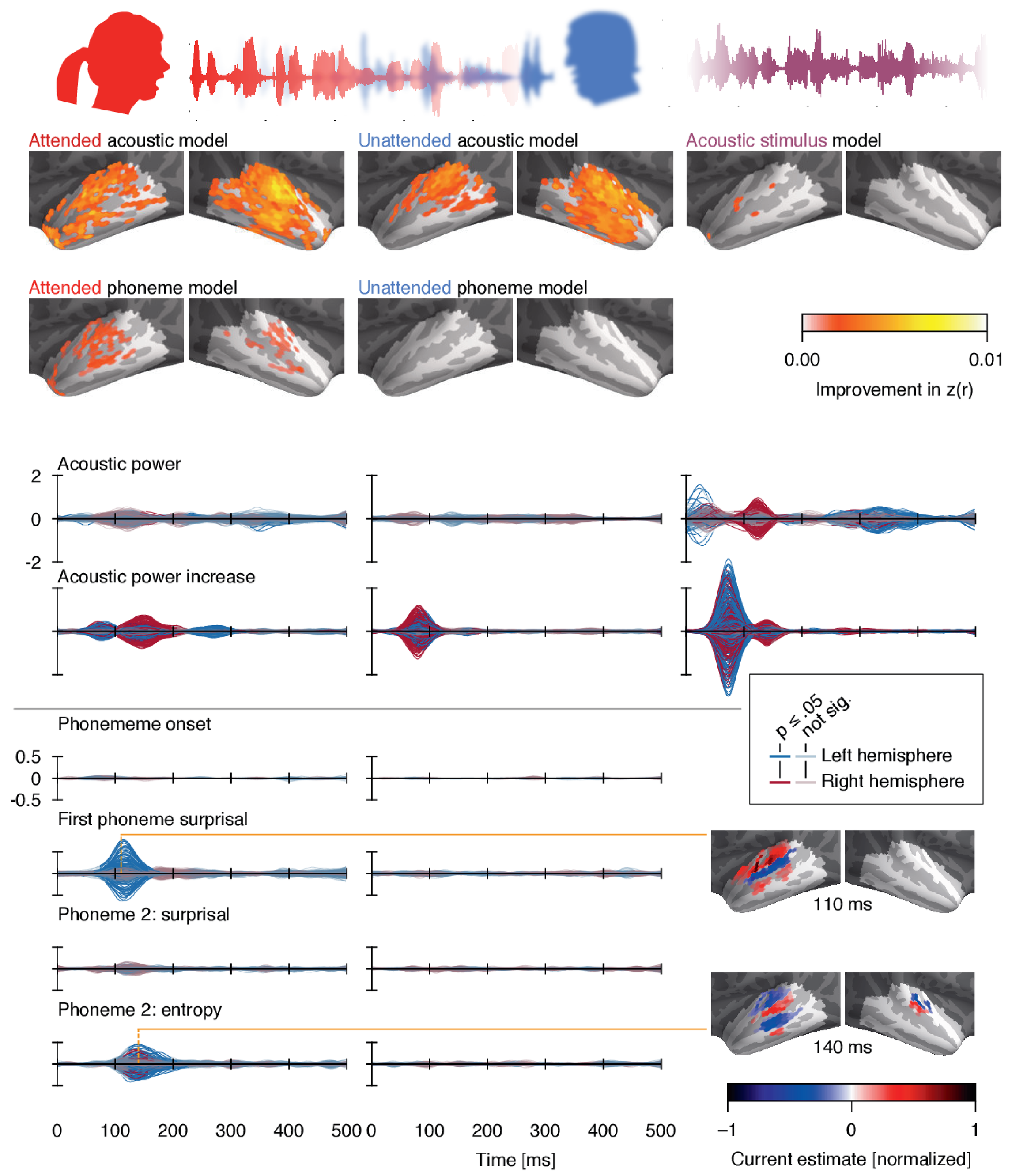
- Extended responses to acoustic features
- Monophasic early responses to phonemes



Results: two speakers

Responses to two speakers modeled using variables significant in single speaker model:

- Responses reflect acoustic information from attended and unattended speech
- Significant respones to phoneme information in attended but not unattended speech
- Attended response peaks similar to single speaker case



Conclusions

- Responses to phonemes can be disentangled from responses to underlying acoustic features
- Responses to phoneme surprisal and entropy suggest information from phonemes is used to constrain the lexical cohort within ~110-120 ms of phoneme onset
- Response to word onset suggests fast real-time lexical segmentation of continuous speech
- In two-speaker stimuli, only attended speech is processed lexically (cf. Broderick et al., 2017)
- Source localization suggests that lexical processing of phonetic information takes place in the lateral temporal lobe in or near auditory cortex

Methods

MEG data

- 26 participants listened to one-minute long segments from A Child's History of England by Charles Dickens. In each of 4 blocks, subjects heard 4 repetitions of a mix of two segments, one spoken by a female and one by a male speaker. They were instructed to focus on one speaker while ignoring the other (counter-balanced across subjects). Then, each of the two segments was presented in isolation. After each presentation, subjects answered a comprehension question.
- An average brain model ("fsaverage", FreeSurfer) was scaled and coregistered to each subject's head shape. MEG data were projected to source space with a distributed minimum norm inverse solution. Source dipoles were constrained to be orthogonal to the cortical surface. Only source estimates in the temporal lobes were retained for analysis (~315 source dipoles per hemisphere).

Predictor variables

- Acoustic spectrogram predictors were based on an auditory brainstem model (Yang et al., 1992). Acoustic power representation: the spectrogram was averaged across frequency in 8 bands. Acoustic onset representation: positive slope of the acoustic power, 0 where the slope is negative.
- Phonemes were labeled using the Gentle forced aligner (https://lowerquality.com/gentle/) and hand corrected
- Phoneme predictor variables were constructed using pronunciations from the Carnegie Mellon University Pronunciation Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict) and word frequencies from the SUBTLEX database (Brysbaert and New, 2009).

Response functions

- Response functions were estimated separately for each virtual current source dipole using the boosting algorithm (David et al., 2007), using a response shape prior of Hamming windows of 50 ms width. Each predictor was tested by comparing prediction accuracy (correlation between predicted and measured response) of the full model to a model in which the predictor was temporally permuted. Model improvements and response functions were assessed using permutation tests based on threshold-free cluster enhancement (Smith and Nichols, 2009).

References

Brodbeck, C., Presacco, A., & Simon, J. Z. (2018). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. NeuroImage, 172, 162–174.

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2017). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech.

Brysbaert, M., and New, B. (2009). Moving beyond Kucera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. Behav Res Methods 41, 977–990.

David, S.V., Mesgarani, N., and Shamma, S.A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. Netw. Comput. Neural Syst. 18, 191–212.

Gaston, P., and Marantz, A. (2017). The time course of contextual cohort effects in auditory processing of category-ambiguous words: MEG evidence for a single “clash” as noun or verb. Lang. Cogn. Neurosci. 0, 1–22.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. Cognition, 25(1), 71–102.

Norris, D., and McQueen, J.M. (2008). Shortlist B: A Bayesian model of continuous speech recognition. Psychol. Rev. 115, 357–395.

Smith, S.M., and Nichols, T.E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage 44, 83–98.

Yang, X., Wang, K., and Shamma, S.A. (1992). Auditory representations of acoustic signals. IEEE Trans. Inf. Theory 38, 824–839.

Supported by National Institutes of Health (NIH) grant R01-DC-014085

Poster PDF available at http://ter.ps/simonpubs