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

- In continuous speech, phoneme identity is hard to dissociate from the acoustic signal
- 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 (Gaston and Marantz, 2017)
- We analyze MEG responses to phoneme properties in continuous, uninterrupted speech to determine when phonemes are processed as linguistically relevant stimuli

Predictor variables

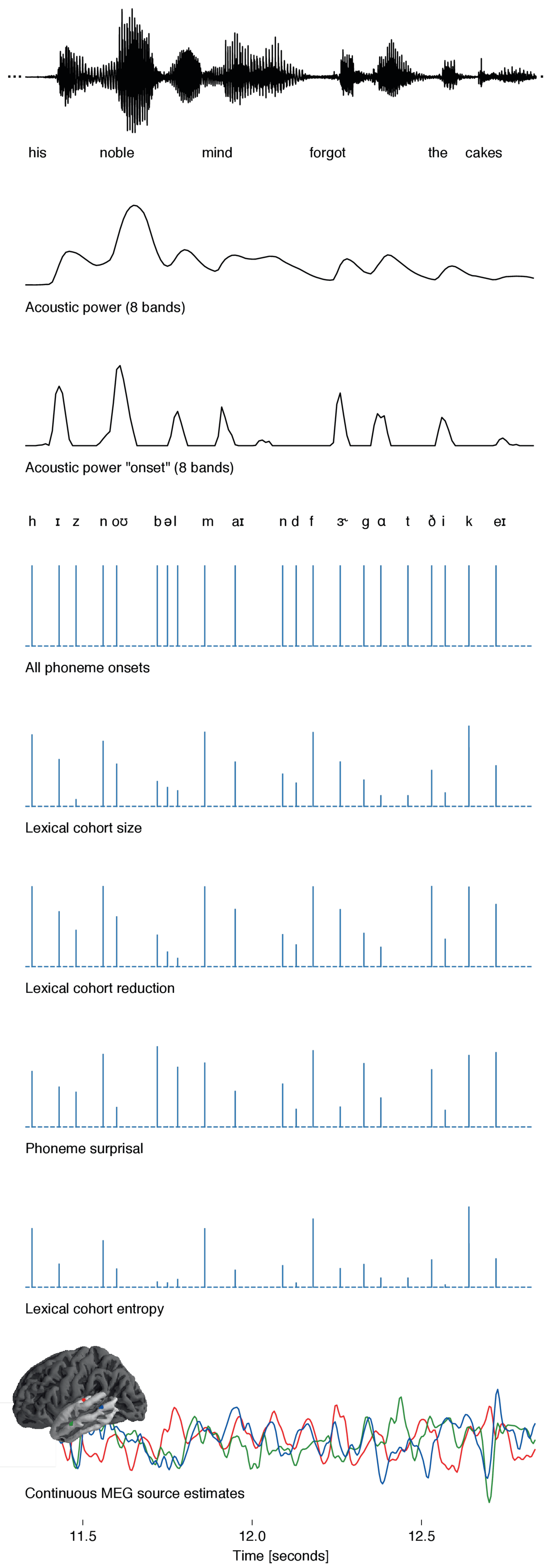
- **Acoustic envelope**: acoustic power
- **Acoustic "onset"**: rising slope of acoustic power
- **Cohort size**: number of word forms compatible with the current prefix
- **Cohort reduction**: number of words that the current phoneme excludes
- **Phoneme surprisal**: inverse of the conditional probability of the phoneme
- **Cohort entropy**: degree of uncertainty about the current word
- Two versions where applicable: **form** based, assuming a lexicon of word surface forms, and **segmented**, assuming sequential identification of morphemes (Balling and Baayen, 2012), but the two versions are highly correlated (r ≥ 0.97)

Stimuli

- **Solo**: one minute long audiobook segments
- **Two-speaker mix**: two audiobook segments, 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., 2017)
- Model evaluated using incremental prediction accuracy; least significant predictor excluded until all are significant (c.f. Balling and Baayen, 2012)



Results: single speaker

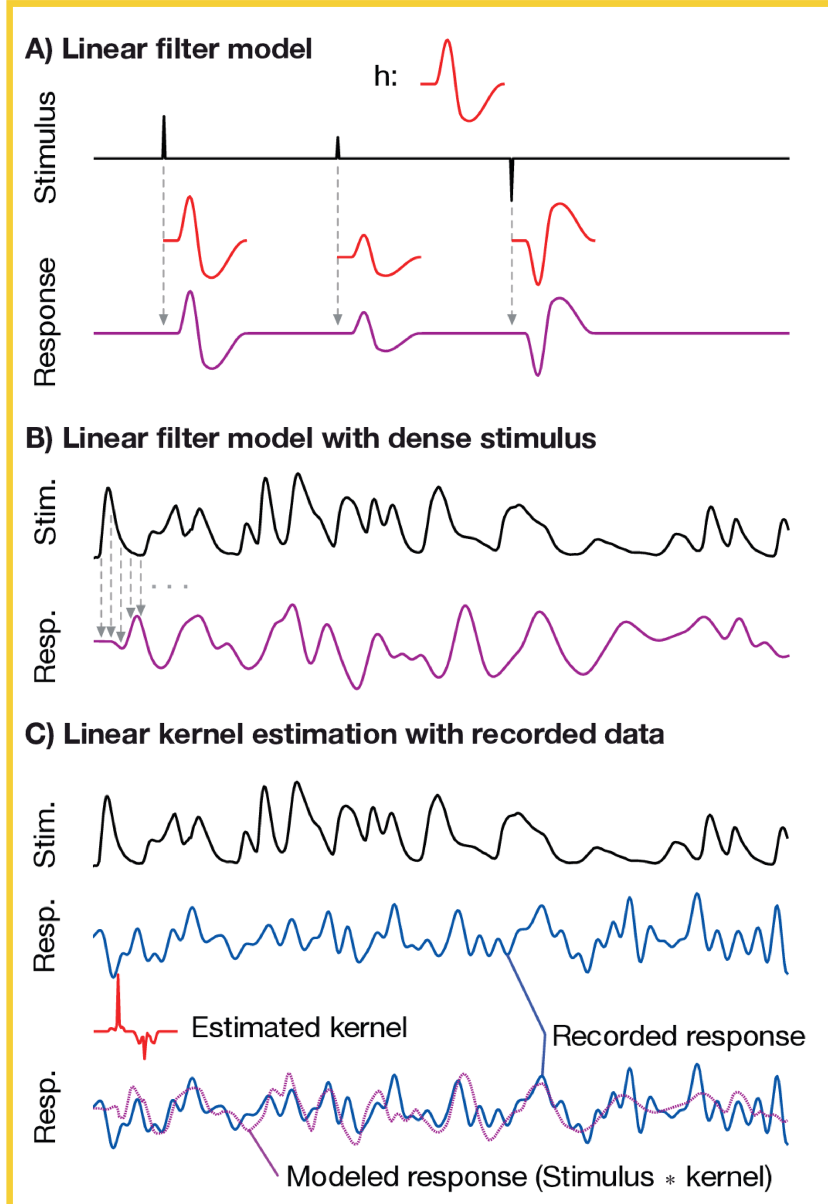
The set of phoneme-based variables was reduced to significant predictors:

Effect	P	
Acoustic envlope	< .001	***
Phone onset	< .001	***
Cohort (form)	.125	
Cohort (segmented)	< .001	***
Cohort reduction (form)	.376	
Cohort reduction (segmented)	.002	**
Surprisal	< .001	***
Entropy (form)	< .001	***
Entropy (segmented)	.161	

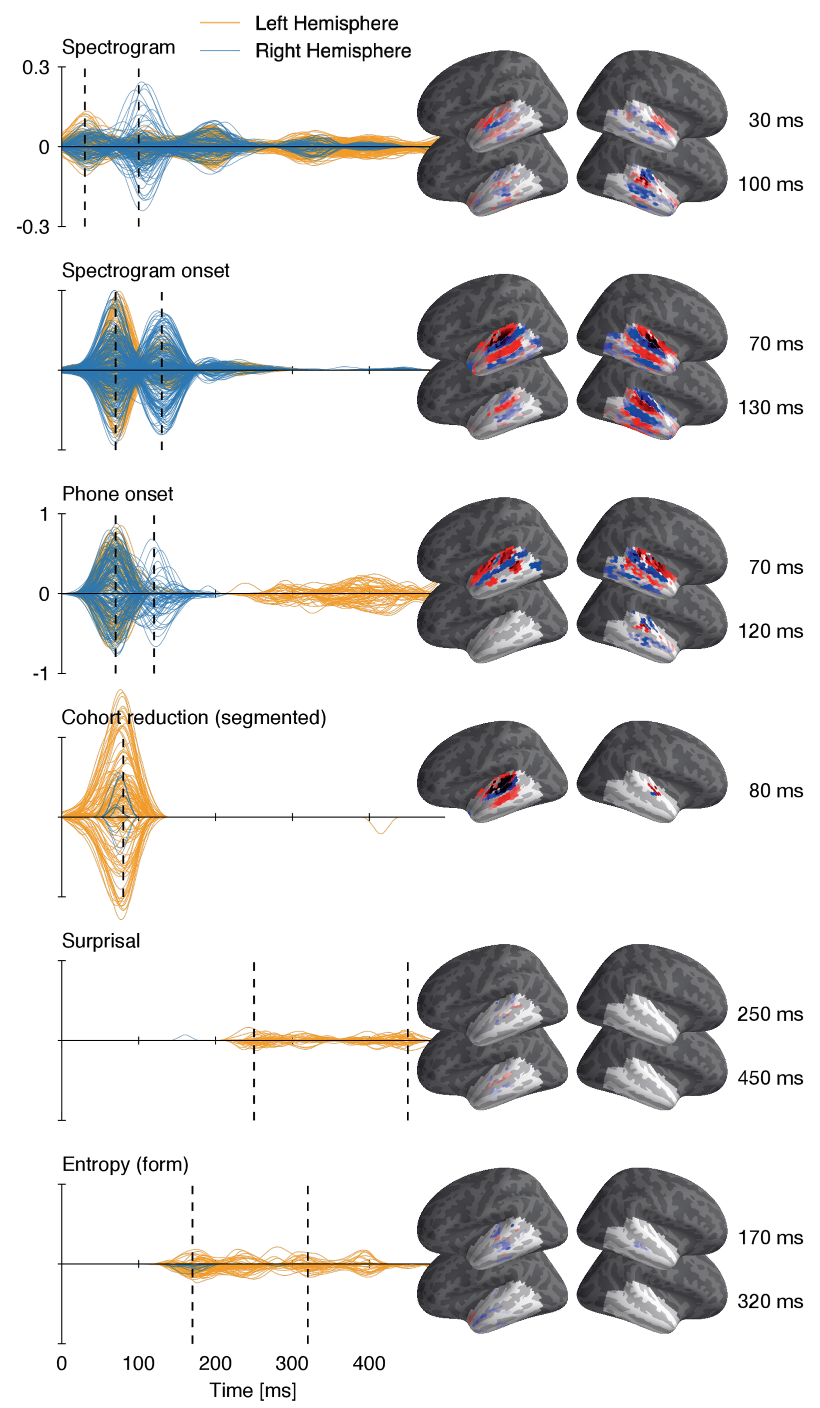
Significant predictors were further tested while controlling for the spectrogram (8 bands) and its onset representation:

Effect	P	
Spectrogram	< .001	***
Spectrogram onset	< .001	***
Phone onset	< .001	***
Cohort (segmented)	.791	
Cohort reduction (segmented)	< .001	***
Surprisal	< .001	***
Entropy (form)	.004	**

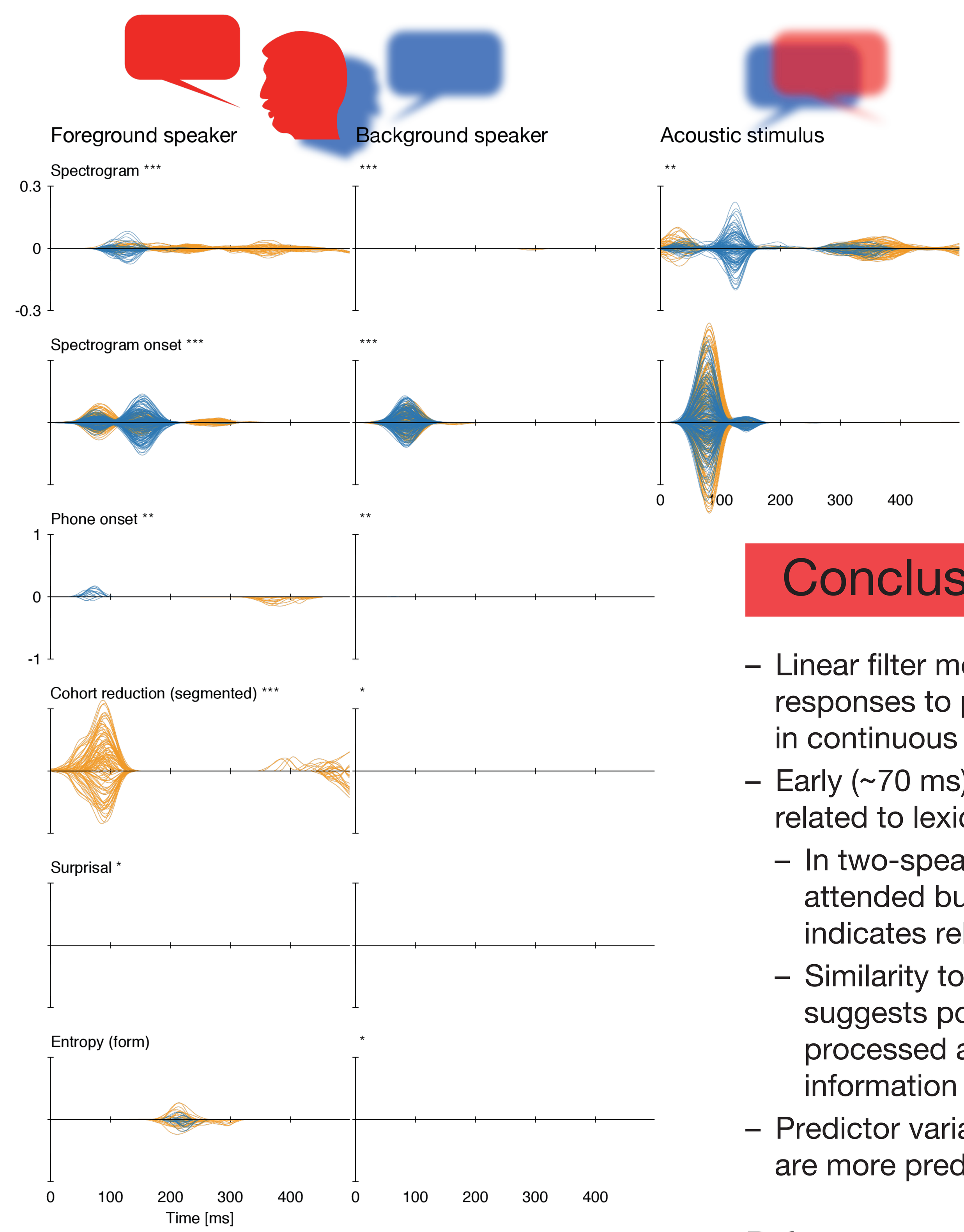
Method: linear kernel estimation



Response functions for the remaining predictors, masked to show only elements that are significant across participants:



Results: two speakers



- Model including all predictors that are significant in the single-talker model
- Responses to acoustic properties of unattended speech, but not to the same degree as attended speech
- Responses to phoneme-information of attended, but not unattended speech

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; only source estimates in the temporal lobes were retained for analysis (~315 source dipoles per hemisphere).

Predictor variables

- Acoustic envelope: average of all frequency channels of an auditory brainstem model (Yang et al., 1992). Spectrogram: average in 8 bands of the same model. Acoustic onset: positive slope of the spectrogram, 0 where the slope is negative. Phoneme predictor variables were constructed using pronunciations from the Carnegie Mellon University Pronunciation Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict), word frequencies from the SUBTLEX database (Brysbaert and New, 2009) and morphological decomposition from the CELEX database (Baayen et al., 1995).

Response functions

- Response functions were estimated separately for each virtual current source dipole using the boosting algorithm (David et al., 2007). Each predictor was tested by comparing prediction accuracy 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).

Conclusions

- Linear filter model can deconvolve brain responses to phoneme information properties in continuous speech
- Early (~70 ms) left auditory cortex response related to lexical cohort processing
  - In two-speaker stimuli, response to attended but not to ignored speaker indicates relevance to linguistic processing
  - Similarity to acoustic onset response suggests possibility that acoustic cues are processed as linguistically relevant information
- Predictor variables are correlated, but some are more predictive than others

References

Baayen, R.H., Piepenbrock, R., and Gulikers, L. (1995). CELEX2 LDC96L14 (Philadelphia: Linguistic Data Consortium).

Balling, L.W., and Baayen, R.H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. Cognition 125, 80–106.

Brodbeck, C., Presacco, A., and Simon, J.Z. (2017). Neural source dynamics of brain responses to continuous stimuli: speech processing from acoustics to comprehension. BioRxiv 182881.

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.

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