Report

Current Biology

Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech

Highlights

- MEG responses to continuous speech analyzed at both acoustic and lexical level
- Responses indicate phoneme level predictive coding and lexical competition
- Phonetic information is processed lexically as early as ~114 ms after phoneme onset
- Lexical responses are restricted to attended speech in a selective attention paradigm

Authors

Christian Brodbeck, L. Elliot Hong, Jonathan Z. Simon

Correspondence

brodbeck@umd.edu (C.B.), jzsimon@umd.edu (J.Z.S.)

In Brief

Analyzing MEG responses to continuous narrative speech, Brodbeck et al. find evidence of early lexical processing, involving both phoneme-level predictive coding and lexical competition. In a selective attention paradigm, involving two concurrent speakers, responses indicate that lexical processing is restricted to the attended speech.

Brodbeck et al., 2018, Current Biology *28*, 3976–3983
December 17, 2018
https://doi.org/10.1016/j.cub.2018.10.042

CellPress

CellPress

Current Biology
Report

Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech

Christian Brodbeck,[1,5,*] L. Elliot Hong,[2] and Jonathan Z. Simon[1,3,4,*]
[1]Institute for Systems Research, University of Maryland, College Park, MD 20742, USA
[2]Department of Psychiatry, Maryland Psychiatric Research Center, University of Maryland School of Medicine, Baltimore, MD 21201, USA
[3]Department of Electrical and Computer Engineering, University of Maryland, College Park, MD 20742, USA
[4]Department of Biology, University of Maryland, College Park, MD 20742, USA
[5]Lead Contact
*Correspondence: brodbeck@umd.edu (C.B.), jzsimon@umd.edu (J.Z.S.)
https://doi.org/10.1016/j.cub.2018.10.042

SUMMARY

During speech perception, a central task of the auditory cortex is to analyze complex acoustic patterns to allow detection of the words that encode a linguistic message [1]. It is generally thought that this process includes at least one intermediate, phonetic, level of representations [2–6], localized bilaterally in the superior temporal lobe [7–9]. Phonetic representations reflect a transition from acoustic to linguistic information, classifying acoustic patterns into linguistically meaningful units, which can serve as input to mechanisms that access abstract word representations [10, 11]. While recent research has identified neural signals arising from successful recognition of individual words in continuous speech [12–15], no explicit neurophysiological signal has been found demonstrating the transition from acoustic and/or phonetic to symbolic, lexical representations. Here, we report a response reflecting the incremental integration of phonetic information for word identification, dominantly localized to the left temporal lobe. The short response latency, approximately 114 ms relative to phoneme onset, suggests that phonetic information is used for lexical processing as soon as it becomes available. Responses also tracked word boundaries, confirming previous reports of immediate lexical segmentation [16, 17]. These new results were further investigated using a cocktail-party paradigm [18, 19] in which participants listened to a mix of two talkers, attending to one and ignoring the other. Analysis indicates neural lexical processing of only the attended, but not the unattended, speech stream. Thus, while responses to acoustic features reflect attention through selective amplification of attended speech, responses consistent with a lexical processing model reveal categorically selective processing.

RESULTS AND DISCUSSION

Magnetoencephalography (MEG) responses to continuous narrative speech were analyzed with a framework designed to measure acoustic and lexical processing simultaneously. Source-localized brain responses were modeled as linear responses to multiple predictor variables that reflect acoustic and lexical properties of continuous speech (see Figure 1). Each source's response time course was modeled as a sum of responses to all predictors, such that the predictors competed for explaining variance in the response [13]. Initially, a range of predictor variables were generated to cover a variety of neural processes plausibly involved in lexical processing. Statistical model comparison was then used to determine which of those variables significantly improved neural response predictions.

Acoustic properties were modeled through an 8-band auditory spectrogram and its half wave rectified derivative to model both the continuously varying "acoustic envelope" and "acoustic onsets" [20].

Responses to phonemes, regardless of their informational value, were modeled with a binary "phoneme onset" predictor variable. Modulation of phoneme responses due to lexical processing was modeled with impulses at phoneme onsets of variable size. All variables were based on the premise, widely supported by behavioral experiments, that phonetic information is used to incrementally constrain possibilities for the word that is currently being processed [10, 11, 21]; this entails initial activation of multiple candidate lexical items, which compete for recognition until they become incompatible with the input. For example, after hearing the phoneme sequence /noʊ/, both *noble* and *notable* might be activated as potential candidates, but once the next phoneme /b/ becomes available (generating the sequence /noʊb/), *notable* would be discarded as a possibility. This model suggests that with the occurrence of each phoneme in a word, there is a cohort of lexical items compatible with the phoneme sequence up to the most current phoneme. Since activation of allowable candidates might be associated with neural processes, "cohort size" was modeled as the number of lexical items under consideration at each phoneme occurrence. The number of items removed from the cohort at each phoneme, "cohort reduction", was used as an estimate of how informative a given phoneme is. While these two variables treat all lexical

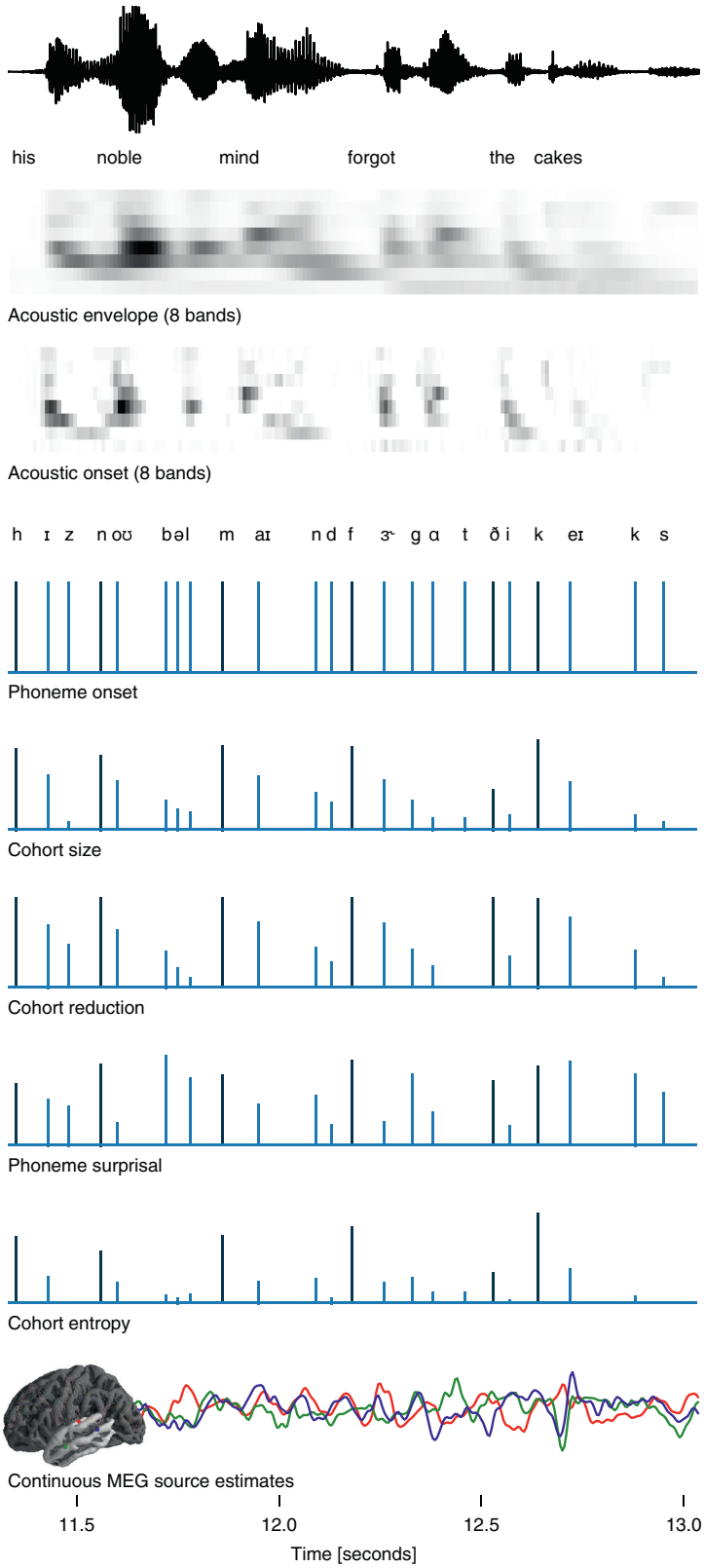

Figure 1. Analysis Framework, Illustrated with an Excerpt from One of the Stimuli

The acoustic waveform (top row) is shown for reference only. Subsequent rows show the predictor variables used to model responses to a single speaker. Acoustic predictors were based on an auditory spectrogram aggregated into 8 frequency bands. For the phoneme-based predictor variables, the initial phoneme of each word is drawn in black, whereas all subsequent phonemes are drawn in blue. The last row contains estimated brain responses from three virtual current dipoles, representative of the modeled signal. The anatomical plot of the cortex is shaded to indicate the temporal lobe, the anatomical region of interest (only the left hemisphere is shown, but both hemispheres were analyzed). See Table S1 for correlations between different predictor variables and Figure S1 for corresponding scatter-plots (of the phoneme-based predictor variables).

items equally, evidence suggests that lexical processing is sensitive to frequency of usage [11, 22]. If this is reflected neurally, a better estimate of phoneme informativeness would be "phoneme surprisal", which reflects how surprising a phoneme is, given the cohort that is currently active, and assuming that the probability of hearing each word is proportional to its frequency in a large speech corpus [23]. Finally, cohort theory suggests that lexical items compete for activation from incomplete input. The degree of competition can be quantified as "cohort entropy", which is defined as the Shannon entropy [24] over all lexical items compatible with the input at the given point in the word. Both phoneme surprisal and cohort entropy have been shown to be associated with reaction times [25, 26] and MEG responses [27–30] to isolated spoken words.

"Word onsets", i.e., word-initial phonemes, were modeled separately from subsequent phonemes (black and blue in Figure 1) to account for the possibility that word onsets might involve different or additional processes, e.g., activation of an initial cohort as opposed to modification of an existing cohort [30, 31].

Responses to Single Speaker Reflect Lexical Processing

MEG recordings from participants listening to a single talker were used to determine which variables significantly predict brain responses. Taken together, the 8 lexical processing variables significantly improved model predictions (t_{max} = 5.93, p < 0.001; significance was assessed by comparing the predictive power of the full model to the ensemble average of 3 models representing the null hypothesis, in which the predictors under investigation were shuffled by randomly re-assigning their values to different phonemes). Because these variables are not independent (see Table S1 and Figure S1), the initial set of variables was reduced to a set in which each variable explained a distinct proportion of the variance in the data. To this end, the significance of each lexical variable was evaluated, and the model was reduced by sequentially excluding non-significant predictors until only significant variables remained (*cf.* [25]).

Two of the eight lexical variables survived this minimization procedure: phoneme surprisal (t_{max} = 4.47, p < 0.001) and cohort entropy (t_{max} = 5.68, p < 0.001). Both variables were already significant in the initial model including all variables (t_{max} = 3.87, p = 0.006; t_{max} = 4.61, p < 0.001), and none of the other 8 were (Table S2). This suggests that surprisal and entropy both account for unique features of brain responses even when controlling for all other variables. More generally, this result suggests that it was possible to distinguish contributions from the different variables even though they were correlated to various degrees. The model resulting from this minimization procedure, henceforth called the "reduced model", is shown in Figure 2. The left column shows anatomical plots, indicating where the predictor significantly improved predictions (for acoustic features and phoneme onset, significance was evaluated against shuffled models in which predictors were time-shifted by 15, 30 or 45 s). The right column shows the filter kernels estimated for the reduced model, the so-called temporal response functions (TRFs). TRFs reflect the estimated response to an elementary temporal feature in the stimulus [13, 32, 33] and are thus a continuous analog of evoked responses to temporally distinct events.

The effect of surprisal was significantly left-lateralized (t_{max} = 4.16, p = 0.001). Lateralization of entropy was not significant in the full dataset (t_{max} = 2.86, p = 0.103), though it became significant when the 3 left-handed subjects were excluded (t_{max} = 4.60, p = 0.005). The anatomical centers of mass of the peak responses to surprisal and entropy did not differ significantly (d = 3 mm, p = 0.701), but the response associated with surprisal peaked significantly earlier than the one associated with entropy (114 ms versus 125 ms, $t_{(25)}$ = 2.45, p = 0.022), suggesting that the two variables may reflect separable stages of speech processing. The temporal order of the effects is consistent with the level of information that the two variables encode in the context of the cohort model and, by implication, the neural mechanisms they might arise from: surprisal is a more local measure of phoneme prediction error, which could be related to updating of a predictive coding mechanisms [27], whereas cohort entropy incorporates information about the cohort of lexical items activated by the current phoneme sequence, possibly reflecting lexical competition [27, 34].

More broadly, such activation of form and lexical item information in the superior and middle temporal lobe is consistent with reports of hemodynamic activity in this region [35–38]—for example, effects of speech intelligibility [39] and generalization across different acoustic realizations of the same sentence [40]. Our results suggest that *acoustic* information is used to update *phonetic* expectations held in the STG by approximately 114 ms and to constrain the activated cohort of *lexical* items by 125 ms. While the earliness of these effects might be surprising, evidence from gating studies suggests that 50–100 ms of input is sufficient to correctly identify the initial phoneme of a word [41], and it is plausible that the cortex uses this information as soon as it becomes available. Furthermore, these latencies were calculated from phoneme onset, without additional consideration of coarticulation cues. Since lexical processing is sensitive to coarticulation cues [5, 42, 43], information about phoneme identity may benefit from priming prior to the nominal phoneme onset. Finally, continuous, meaningful speech provides rich layers of contextual information, and it is plausible that processes underlying lexical perception are facilitated by this information [44].

Neither cohort size nor cohort reduction significantly predicted neural responses. Our results thus support the view that lexical perception is sensitive to language statistics, modeled through frequency of use. None of the predictors associated with word-initial phonemes were significant, suggesting that responses to initial phonemes could not be modeled by lexical distributions. This could indicate that word-initial phonemes are processed differently from subsequent ones [30, 31].

Responses Reflect Lexical Segmentation

The effect of word onset, compared to a null distribution estimated from models in which word onsets were randomly assigned among all phonemes, was highly significant (t_{max} = 4.76, p < 0.001). Despite a numerically larger effect in the left hemisphere, lateralization was not significant (t_{max} = 2.58, p = 0.118). With a latency of 103 ms, the peak was significantly earlier than entropy ($t_{(25)}$ = 2.98, p = 0.006) but not surprisal ($t_{(25)}$ = 0.71, p = 0.487). The spatial center of mass of the response peak was located more superior than both surprisal and entropy (d = 9 mm, p = 0.022; d = 8 mm, p = 0.007). While

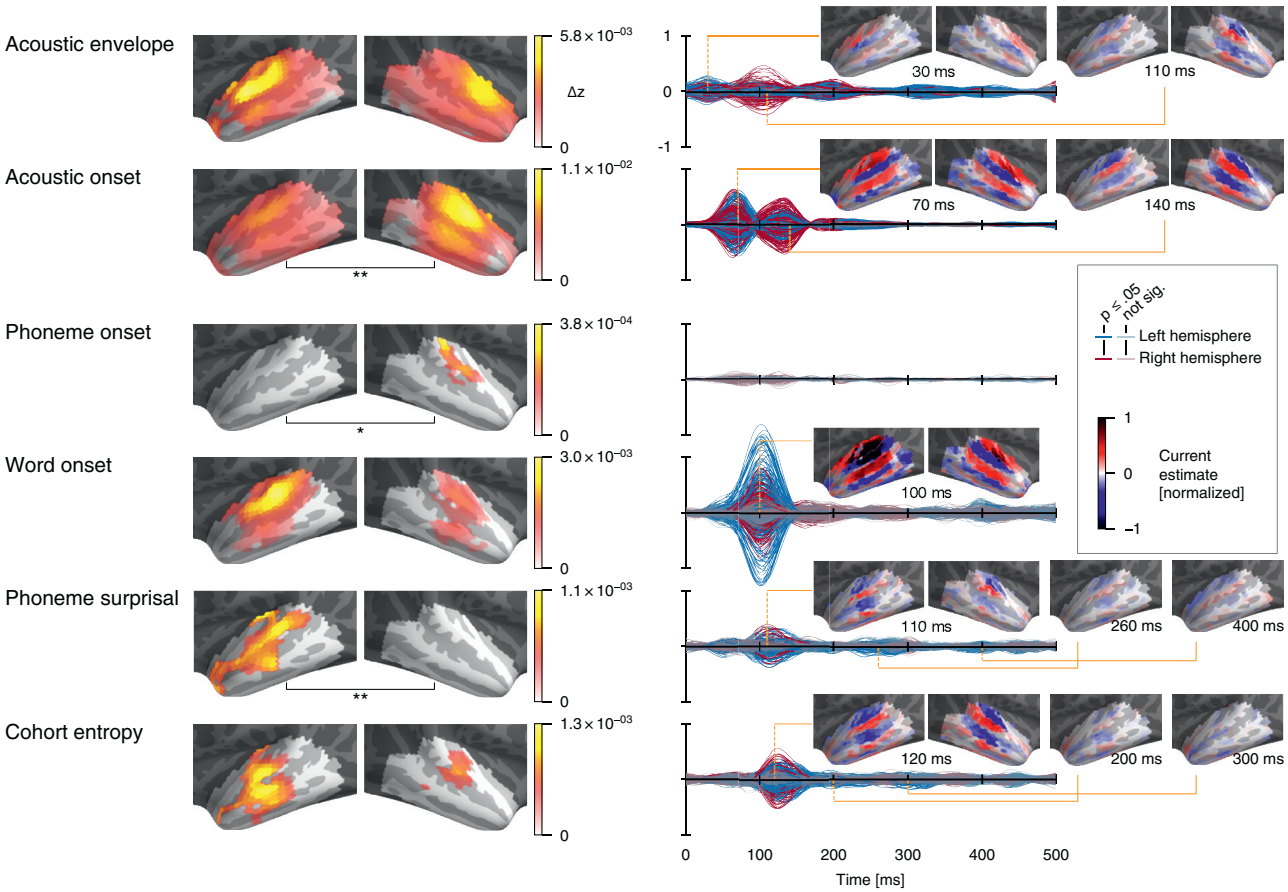

Figure 2. Brain Responses to Single Speaker

Left column: significant predictive power ($p \leq 0.05$, corrected). Colors reflect the difference in *z*-transformed correlation between the full and the appropriately shuffled model. Color-maps are normalized for each predictor to maximize visibility of internal structure, as appropriate for evaluating source localization results: due to spatial dispersion of minimum norm source estimates, effect peaks are relatively accurate estimates, but strong effects can cause spurious spread whose amplitude decreases with distance from the peak. See also Table S2. Right column: Temporal response functions (TRFs) estimated for the reduced model. Each line reflects the TRF at one virtual current dipole, with color coding its location by hemisphere, and saturation coding significance ($p \leq 0.05$, corrected). Anatomical plots display TRFs at certain time points of interest (only significant values are shown), with color coding current direction relative to the cortical surface. Acoustic TRFs were averaged across frequency band for display as visual inspection revealed no major differences apart from amplitude differences between frequency bands. See also Figure S2 and Table S3.

it could be argued that word onsets should be associated with disproportionately large surprisal, it is noteworthy that the word onset TRF peak has the opposite polarity (i.e., current direction) than the surprisal peak, further dissociating the two responses.

A more general implication of these responses to word onsets is that word boundaries should be perceptually salient despite the observation that clear cues for word boundaries are generally missing from speech waveforms [45]. A similar word-onset electroencephalographic response emerged only after listeners learned to segment an artificial language into words, suggesting that it is not a response to local acoustic properties alone [16, 17]. A response tightly locked to word onset suggests that whichever cues listeners use to detect word boundaries [45, 46], boundaries seem to be generally detected as they occur, rather than after incorporating cues occurring subsequent to word onset.

Word-medial phonemes, modeled as an impulse at each phoneme excluding word onsets, were associated with a significant (t_{max} = 3.15, p = 0.005) right-lateralized (t_{max} = 2.75, p = 0.035) response.

Responses to Acoustic Features

Both acoustic predictors were associated with strong bilateral effects (t_{max} = 9.08 and 9.23, both $p < 0.001$). Both were localized close to core auditory cortex, with acoustic onsets somewhat more predictive in the right hemisphere (t_{max} = 4.24, p = 0.007). Significant effects extended over much of the temporal lobe, though the extended area could be due to spatial dispersion of MEG source estimates rather than genuine responses outside of core auditory regions [13]. TRFs to the acoustic envelope exhibited two main peaks at 30 and 106 ms, consistent with earlier results [13, 33, 47]. The latency of two analogous peaks to acoustic onsets at 68 and 131 ms was found to be greater, as

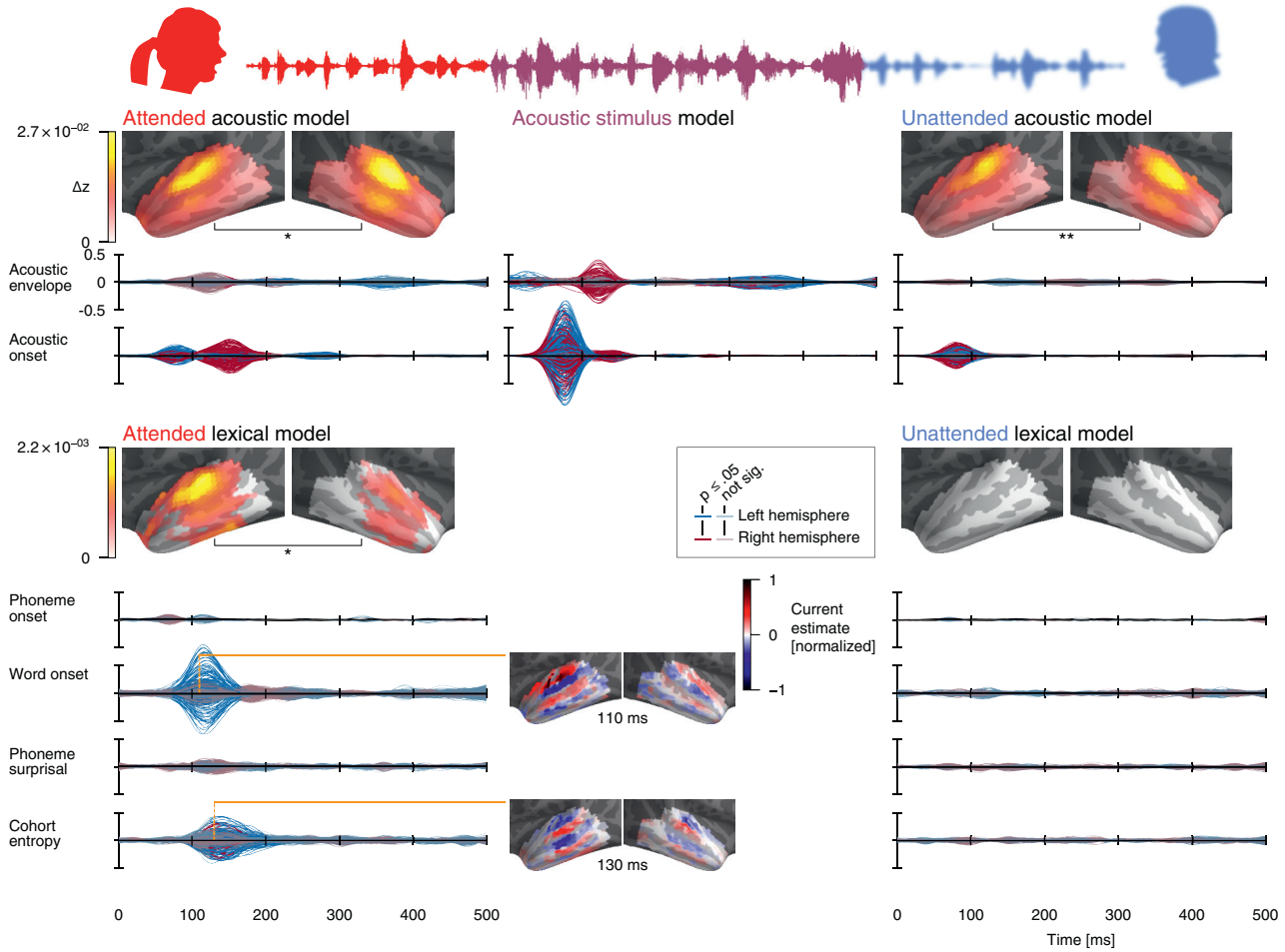

Figure 3. Brain Responses to Two Concurrent Speakers

Details analogous to Figure 2. The three columns display results for the model components for: the attended speech stream (left), the actual acoustic stimulus mixture (middle), and the unattended speech stream (right). The upper part of the figure displays results for acoustic features, the lower part for lexical processing.

expected due to the temporal relationship between the two variables: the time of maximum rising slope precedes the time of maximum amplitude, and is thus earlier compared with specific time points in the neural response. The presence of analogous peaks in the TRFs to both acoustic representations might indicate that they jointly arise from a single, more complex underlying neural response type, reflecting both onset and continuous acoustic properties [48]. On the other hand, spatially, the two response peaks to acoustic onsets were localized posterior to the corresponding acoustic envelope peaks (d = 8 mm, p = 0.002; d = 10 mm, p < 0.001), which might instead indicate that the two responses stem from partially distinct neural populations [49].

Responses to Two Concurrent Speakers Reflect Acoustic, but Not Lexical, Information in Unattended Speech

The variables that significantly predicted responses to a single speaker were used to model acoustic and lexical processing in a version of the cocktail-party paradigm [18, 19]. Participants listened to a single-channel acoustic mixture of a male and a female speaker, attending to one and ignoring the other. This made it possible to test whether the lexical processing observed for a single speaker is restricted to the attended speech stream or whether it occurs also for the unattended stream. Figure 3 shows the predictive power of groups of predictors modeling relevant processing stages and TRFs for the full model fitted to the two-speaker data.

Responses were significantly modulated by acoustic features of both the attended and the unattended speaker (t_{max} = 11.83 and 16.67, both p < 0.001; lateralization t_{max} = 4.17, p = 0.041 and t_{max} = 5.28, p = 0.001). The relative amplitudes of the TRF peaks to acoustic onsets were consistent with previous results [33, 47, 50], with an earlier (~70 ms) peak predominantly reflecting the raw acoustic mixture, and a later (~150 ms) peak predominantly reflecting acoustic energy in the attended speech. Responses to the acoustic envelope almost exclusively reflected processing of the acoustic mixture, suggesting that auditory stream segregation may be predominantly reflected in onset processing.

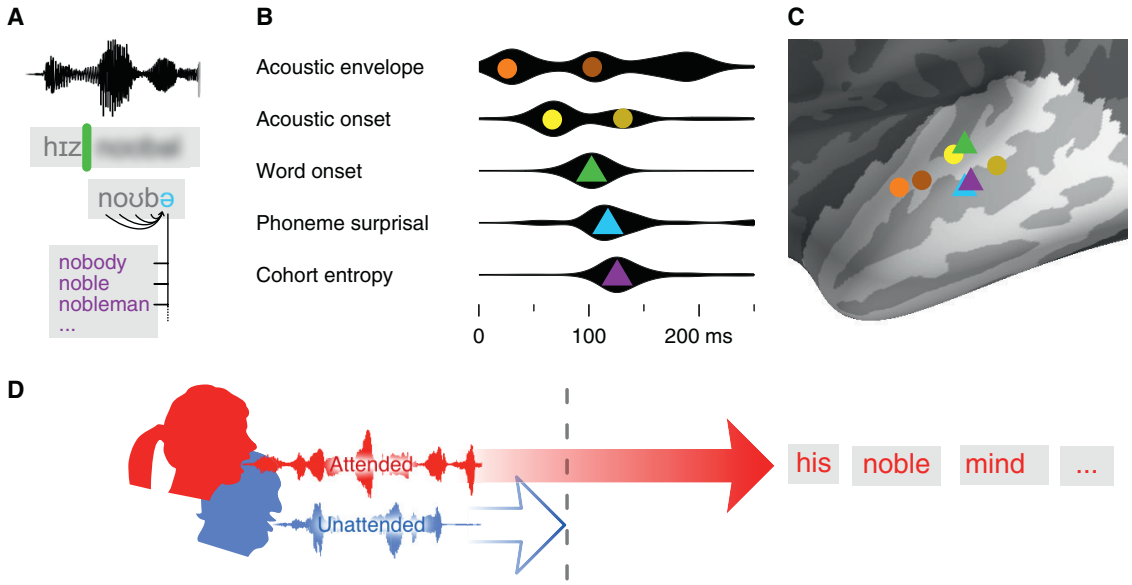

Figure 4. Summary of Results

(A) Illustration of aspects of the cohort model on which significant variables were based: lexical segmentation (word onset), predictive coding based on preceding phoneme sequence (phoneme surprisal) and lexical competition (cohort entropy).

(B) Time course of TRF amplitude for each variable, major peaks marked with symbols corresponding to those used in (C).

(C) Center of mass of average peaks shown in B (see also Table S3 and Figure S2).

(D) Schematic illustration of results of the two-speaker analysis: early acoustic TRF peaks track the processing of the acoustic signal from both speakers, whereas lexical TRFs track processing of only the attended speech.

In contrast to the acoustic models, only the lexical processing model for the attended speech showed a significant effect (t_{max} = 6.08, p < 0.001); lexical properties of the unattended stream did not (t_{max} = 2.90, p = 0.159), and the effect of lexical processing of the attended speech was significantly greater than that for unattended speech (t_{max} = 6.13, p < 0.001). Individually, only word onset and cohort entropy significantly contributed to model predictions (t_{max} = 4.93, p < 0.001 and t_{max} = 3.96, p = 0.002), while surprisal did not (t_{max} = 2.63, p = 0.802). Consistent with this, TRFs to word onset and cohort entropy were significant in the attended speech only. These responses were very similar to the corresponding single speaker responses in shape, although with a significant delay (word onset: 118 versus 103 ms $t_{(25)}$ = 2.52, p = 0.018; entropy: 140 versus 125 ms, $t_{(25)}$ = 2.15, p = 0.042).

While recent research suggests that processing of information contingent on successful word recognition is suppressed for unattended speech [12, 14], these previous findings leave open the possibility that unattended speech is processed up to and including identification of lexical items, but without retrieval of the recognized words' properties. The results presented here indicate that lexical processing of unattended speech is suppressed at the level of detecting word forms. This raises the possibility that lexical processing constitutes a bottleneck in speech perception. Lexical perception is thought to be massively parallel by involving activation of multiple candidate lexical representations through the cohort [10]. The mechanisms implementing this multiple activation might involve mental resources that cannot be shared across parallel instances of the same process, making it impossible for more than one cohort to be represented at the same time.

Conclusion

MEG responses to continuous speech reflect a transformation of the speech signal from acoustic representations, which can be characterized with spectro-temporal receptive fields, to probabilistically driven activation of lexical units. Phonetic cues are rapidly analyzed for their relevance to word perception, updating a lexical processor in the left temporal lobe within ~130 ms (Figures 4A–4C). In the presence of two competing speakers, this transformation is restricted to the attended speech stream (Figure 4D). While the analysis presented here is naturally limited to a specific listening condition and adults with normal hearing, the framework lends itself to studying the influence of different conditions and individual differences on speech processing. Importantly, the methods presented here allow studying lexical processing while listeners are engaged in comprehension of continuous speech without an intrusive extraneous task.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Stimuli
  - MEG data acquisition and preprocessing
  - Source localization

- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Predictor variables
  - Model estimation
  - Model comparisons
  - Test of lateralization
  - Spatio-temporal response functions
  - Response time course
  - Response localization
  - Single speaker analysis
  - Two speaker analysis
  - No detectable effect of repeated presentation
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information includes two figures and three tables and can be found with this article online at https://doi.org/10.1016/j.cub.2018.10.042.

ACKNOWLEDGMENTS

This work was supported by a National Institutes of Health grant R01-DC-014085 (to J.Z.S.) and by a University of Maryland Seed Grant (to L.E.H. and J.Z.S.). We would like to thank Krishna Puvvada for his assistance in designing and preparing the stimuli and Natalia Lapinskaya for her help in collecting data and for excellent technical support.

AUTHOR CONTRIBUTIONS

J.Z.S. and L.E.H. conceived and performed the experiments and secured funding. C.B. and J.Z.S. conceived and performed the analysis and wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: May 18, 2018
Revised: September 6, 2018
Accepted: October 16, 2018
Published: November 29, 2018

REFERENCES

1. McQueen, J.M. (2007). Eight questions about spoken word recognition. In The Oxford Handbook of Psycholinguistics, M.G. Gaskell, ed., pp. 37–53.
2. Kazanina, N., Bowers, J.S., and Idsardi, W. (2017). Phonemes: Lexical access and beyond. Psychon. Bull. Rev. *25*, 560–585.
3. Phillips, C., Pellathy, T., Marantz, A., Yellin, E., Wexler, K., Poeppel, D., McGinnis, M., and Roberts, T. (2000). Auditory cortex accesses phonological categories: an MEG mismatch study. J. Cogn. Neurosci. *12*, 1038–1055.
4. Stevens, K.N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. J. Acoust. Soc. Am. *111*, 1872–1891.
5. Marslen-Wilson, W., and Warren, P. (1994). Levels of perceptual representation and process in lexical access: words, phonemes, and features. Psychol. Rev. *101*, 653–675.
6. Di Liberto, G.M., O'Sullivan, J.A., and Lalor, E.C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. Curr. Biol. *25*, 2457–2465.
7. Chang, E.F., Rieger, J.W., Johnson, K., Berger, M.S., Barbaro, N.M., and Knight, R.T. (2010). Categorical speech representation in human superior temporal gyrus. Nat. Neurosci. *13*, 1428–1432.
8. Mesgarani, N., Cheung, C., Johnson, K., and Chang, E.F. (2014). Phonetic feature encoding in human superior temporal gyrus. Science *343*, 1006–1010.
9. Hickok, G., and Poeppel, D. (2007). The cortical organization of speech processing. Nat. Rev. Neurosci. *8*, 393–402.
10. Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word-recognition. Cognition *25*, 71–102.
11. Norris, D., and McQueen, J.M. (2008). Shortlist B: a Bayesian model of continuous speech recognition. Psychol. Rev. *115*, 357–395.
12. Broderick, M.P., Anderson, A.J., Di Liberto, G.M., Crosse, M.J., and Lalor, E.C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. Curr. Biol. *28*, 803–809.e3.
13. Brodbeck, C., Presacco, A., and Simon, J.Z. (2018). Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. Neuroimage *172*, 162–174.
14. Ding, N., Pan, X., Luo, C., Su, N., Zhang, W., and Zhang, J. (2018). Attention Is Required for Knowledge-Based Sequential Grouping: Insights from the Integration of Syllables into Words. J. Neurosci. *38*, 1178–1188.
15. Ding, N., Melloni, L., Zhang, H., Tian, X., and Poeppel, D. (2015). Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci. *19*, 158–164.
16. Sanders, L.D., Newport, E.L., and Neville, H.J. (2002). Segmenting nonsense: an event-related potential index of perceived onsets in continuous speech. Nat. Neurosci. *5*, 700–703.
17. Sanders, L.D., and Neville, H.J. (2003). An ERP study of continuous speech processing. I. Segmentation, semantics, and syntax in native speakers. Brain Res. Cogn. Brain Res. *15*, 228–240.
18. Cherry, E.C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. J. Acoust. Soc. Am. *25*, 975–979.
19. McDermott, J.H. (2009). The cocktail party problem. Curr. Biol. *19*, R1024–R1027.
20. Kluender, K.R., Coady, J.A., and Kiefte, M. (2003). Sensitivity to change in perception of speech. Speech Commun. *41*, 59–69.
21. Allopenna, P.D., Magnuson, J.S., and Tanenhaus, M.K. (1998). Tracking the Time Course of Spoken Word Recognition Using Eye Movements: Evidence for Continuous Mapping Models. J. Mem. Lang. *38*, 419–439.
22. Dahan, D., Magnuson, J.S., and Tanenhaus, M.K. (2001). Time course of frequency effects in spoken-word recognition: evidence from eye movements. Cognit. Psychol. *42*, 317–367.
23. 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.
24. Shannon, C.E. (1948). A Mathematical Theory of Communication. Bell Syst. Tech. J. *27*, 379–423, 623–656.
25. Balling, L.W., and Baayen, R.H. (2012). Probability and surprisal in auditory comprehension of morphologically complex words. Cognition *125*, 80–106.
26. Wurm, L.H., Ernestus, M.T.C., Schreuder, R., and Baayen, R.H. (2006). Dynamics of the auditory comprehension of prefixed words: Cohort entropies and Conditional Root Uniqueness Points. Ment. Lex. *1*, 125–146.
27. Gagnepain, P., Henson, R.N., and Davis, M.H. (2012). Temporal predictive codes for spoken words in auditory cortex. Curr. Biol. *22*, 615–621.
28. Ettinger, A., Linzen, T., and Marantz, A. (2014). The role of morphology in phoneme prediction: evidence from MEG. Brain Lang. *129*, 14–23.
29. Gwilliams, L., and Marantz, A. (2015). Non-linear processing of a linear speech stream: The influence of morphological structure on the recognition of spoken Arabic words. Brain Lang. *147*, 1–13.
30. 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. *33*, 402–423.

31. Vitevitch, M.S. (2002). Influence of onset density on spoken-word recognition. J. Exp. Psychol. Hum. Percept. Perform. *28*, 270–278.
32. Lalor, E.C., Pearlmutter, B.A., Reilly, R.B., McDarby, G., and Foxe, J.J. (2006). The VESPA: a method for the rapid estimation of a visual evoked potential. Neuroimage *32*, 1549–1561.
33. Ding, N., and Simon, J.Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. USA *109*, 11854–11859.
34. Revill, K.P., Aslin, R.N., Tanenhaus, M.K., and Bavelier, D. (2008). Neural correlates of partial lexical activation. Proc. Natl. Acad. Sci. USA *105*, 13111–13115.
35. Davis, M.H., Coleman, M.R., Absalom, A.R., Rodd, J.M., Johnsrude, I.S., Matta, B.F., Owen, A.M., and Menon, D.K. (2007). Dissociating speech perception and comprehension at reduced levels of awareness. Proc. Natl. Acad. Sci. USA *104*, 16032–16037.
36. Gagnepain, P., Chételat, G., Landeau, B., Dayan, J., Eustache, F., and Lebreton, K. (2008). Spoken word memory traces within the human auditory cortex revealed by repetition priming and functional magnetic resonance imaging. J. Neurosci. *28*, 5281–5289.
37. Lerner, Y., Honey, C.J., Silbert, L.J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J. Neurosci. *31*, 2906–2915.
38. Obleser, J., and Eisner, F. (2009). Pre-lexical abstraction of speech in the auditory cortex. Trends Cogn. Sci. *13*, 14–19.
39. Davis, M.H., and Johnsrude, I.S. (2003). Hierarchical processing in spoken language comprehension. J. Neurosci. *23*, 3423–3431.
40. Dehaene-Lambertz, G., Dehaene, S., Anton, J.-L., Campagne, A., Ciuciu, P., Dehaene, G.P., Denghien, I., Jobert, A., Lebihan, D., Sigman, M., et al. (2006). Functional segregation of cortical language areas by sentence repetition. Hum. Brain Mapp. *27*, 360–371.
41. Tyler, L.K. (1984). The structure of the initial cohort: evidence from gating. Percept. Psychophys. *36*, 417–427.
42. Warren, P., and Marslen-Wilson, W. (1987). Continuous uptake of acoustic cues in spoken word recognition. Percept. Psychophys. *41*, 262–275.
43. Dahan, D., Magnuson, J.S., Tanenhaus, M.K., and Hogan, E.M. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. Lang. Cogn. Process. *16*, 507–534.
44. Altmann, G.T.M., and Kamide, Y. (1999). Incremental interpretation at verbs: restricting the domain of subsequent reference. Cognition *73*, 247–264.
45. Mattys, S.L., and Bortfeld, H. (2016). Speech Segmentation. In Speech Perception and Spoken Word Recognition Current Issues in the Psychology of Language, M.G. Gaskell, and J. Mirkovic, eds. (Abingdon, Oxon: Psychology Press), pp. 55–75.
46. Mattys, S.L., White, L., and Melhorn, J.F. (2005). Integration of multiple speech segmentation cues: a hierarchical framework. J. Exp. Psychol. Gen. *134*, 477–500.
47. Ding, N., and Simon, J.Z. (2013). Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. J. Neurosci. *33*, 5728–5735.
48. David, S.V., Mesgarani, N., Fritz, J.B., and Shamma, S.A. (2009). Rapid synaptic depression explains nonlinear modulation of spectro-temporal tuning in primary auditory cortex by natural stimuli. J. Neurosci. *29*, 3374–3386.
49. Hamilton, L.S., Edwards, E., and Chang, E.F. (2018). A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. Curr. Biol. *28*, 1860–1871.
50. Puvvada, K.C., and Simon, J.Z. (2017). Cortical Representations of Speech in a Multitalker Auditory Scene. J. Neurosci. *37*, 9189–9196.
51. Oldfield, R.C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. Neuropsychologia *9*, 97–113.
52. Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Parkkonen, L., and Hämäläinen, M.S. (2014). MNE software for processing MEG and EEG data. Neuroimage *86*, 446–460.
53. Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., and Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. Front. Neurosci. *7*, 267.
54. Taulu, S., and Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. Phys. Med. Biol. *51*, 1759–1768.
55. Fischl, B. (2012). FreeSurfer. Neuroimage *62*, 774–781.
56. Dale, A.M., and Sereno, M.I. (1993). Improved Localizadon of Cortical Activity by Combining EEG and MEG with MRI Cortical Surface Reconstruction: A Linear Approach. J. Cogn. Neurosci. *5*, 162–176.
57. Hämäläinen, M.S., and Ilmoniemi, R.J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. Med. Biol. Eng. Comput. *32*, 35–42.
58. Lin, F.-H., Witzel, T., Ahlfors, S.P., Stufflebeam, S.M., Belliveau, J.W., and Hämäläinen, M.S. (2006). Assessing and improving the spatial accuracy in MEG source localization by depth-weighted minimum-norm estimates. Neuroimage *31*, 160–171.
59. Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., et al. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage *31*, 968–980.
60. Yang, X., Wang, K., and Shamma, S.A. (1992). Auditory representations of acoustic signals. IEEE Trans. Inf. Theory *38*, 824–839.
61. Boersma, P., and Weenink, D. (2017). Praat: doing phonetics by computer [Computer program] Available at: http://www.praat.org/.
62. David, S.V., Mesgarani, N., and Shamma, S.A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. Network *18*, 191–212.
63. Brodbeck, C. (2018). Eelbrain 0.27 (Zenodo) Available at: https://eelbrain.readthedocs.io.
64. 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.
65. Nichols, T.E., and Holmes, A.P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. *15*, 1–25.
66. Greve, D.N., Van der Haegen, L., Cai, Q., Stufflebeam, S., Sabuncu, M.R., Fischl, B., and Brysbaert, M. (2013). A surface-based analysis of language lateralization and cortical asymmetry. J. Cogn. Neurosci. *25*, 1477–1492.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
MEG data and predictor variables	Digital Repository at the University of Maryland	http://hdl.handle.net/1903/21109
Software and Algorithms		
Presentation Software	Neurobehavioral Systems (https://www.neurobs.com/)	RRID: SCR_002521
Python 2.7	Anaconda	https://www.anaconda.com
MNE-Python	MNE Developers (http://martinos.org/mne/)	RRID: SCR_005972
Eelbrain	Christian Brodbeck (https://pypi.org/project/eelbrain/)	RRID: SCR_014661
Gentle (forced aligner)	Robert M Ochshorn and Max Hawkins	https://lowerquality.com/gentle/
MATLAB	The MathWorks	RRID: SCR_001622
NSL MATLAB Toolbox	Neural Systems Laboratory, University of Maryland	https://isr.umd.edu/Labs/NSL/Software.htm
Other		
'A Child's History of England' by Charles Dickens, chapters 3 and 8	LibriVox	https://librivox.org/a-childs-history-of-england-by-charles-dickens/
157 axial gradiometer whole head MEG at the University of Maryland	KIT, Kanazawa, Japan	N/A

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Christian Brodbeck (brodbeck@umd.edu).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

MEG data were recorded from 28 native speakers of English, recruited by media advertisements from the Baltimore area as control group for another study. Participants with medical, psychiatric or neurological illnesses, head injury, and substance dependence or abuse were excluded. All subjects signed informed consents and were paid for their participation. Data from two participants were excluded, one due to corrupted localizer measurements, and one due to excessive magnetic artifacts associated with dental work. The sample analyzed was composed of 18 male and 8 female participants with mean age 45.2 (range 22 - 61). All subjects provided informed consent in accordance with the University of Maryland Baltimore Institutional Review Board.

Three participants were left handed [51]. Excluding them from analysis changed only one major result: The effect of cohort entropy, which was not significantly lateralized in the full group, became significantly left-lateralized, as mentioned in Results.

METHOD DETAILS

Stimuli

1 min long segments were extracted from audiobook recordings of *A Child's History of England* by *Charles Dickens*, one chapter read by a male and one by a female speaker (https://librivox.org/a-childs-history-of-england-by-charles-dickens/, chapters 3 and 8). Pauses longer than 300 ms were shortened to an interval randomly chosen between 250 and 300 ms, and loudness was matched perceptually. Cocktail party stimuli were generated by additively combining two segments, one from each speaker, with an initial 1 s period containing only the to-be attended speaker.

Four segments were extracted for each speaker: male-1 through 4 and female-1 through 4; mix-1 through 4 were constructed by mixing male-1 and female-1, and so forth. Participants listened four times to mix-1, while attending to one speaker and ignoring the other (which speaker they attended to was counterbalanced across subject), then 4 times to mix-2 while attending to the other speaker. Then, the four segments just heard were all presented once individually. The same procedure was repeated for stimulus segments 3 and 4. After each mix segment, participants answered a question relating to the content of the attended stimulus.

Participants lay supine and were instructed to keep their eyes closed during stimulus presentation (to minimize ocular artifacts and head movement). Stimuli were delivered through foam pad earphones inserted into the ear canal at a comfortably loud listening level.

MEG data acquisition and preprocessing

Continuous MEG data were acquired with the 157 axial gradiometer whole head MEG system (KIT, Kanazawa, Japan) inside a magnetically shielded room (Vacuumschmelze GmbH & Co. KG, Hanau, Germany) at the University of Maryland, College Park. Sensors (15.5 mm diameter) are uniformly distributed inside a liquid-He dewar, spaced ~25 mm apart, and configured as first-order axial gradiometers with 50 mm separation and sensitivity >5 $fT \cdot Hz^{-1/2}$ in the white noise region (> 1 KHz). Data were recorded with an online 200 Hz low-pass filter and a 60 Hz notch filter at a sampling rate of 1 kHz.

Recordings were pre-processed using mne-python [52, 53]. Flat channel responses were automatically detected and excluded. Extraneous artifacts were removed with temporal signal space separation [54]. Data were filtered between 1 and 40 Hz with a zero-phase FIR filter (mne-python 0.15 default settings). Responses time-locked to the onset of the speech stimuli were extracted and downsampled to 100 Hz.

Source localization

Before the MEG recording, each participant's head shape was digitized (Polhemus 3SPACE FASTRAK) and five marker coils were attached to their head. The marker coils were localized with respect to the MEG sensors at the beginning and at the end of the recording session, and these position measurements were used to determine the head position relative to the MEG sensors. The digitized head shape was used to coregister the FreeSurfer [55] "fsaverage" template brain to each subject's head shape using rotation, translation, and uniform scaling.

A source space was defined using four-fold icosahedral subdivision of the white matter surface of the fsaverage brain, with all source dipoles oriented perpendicularly to the cortical surface. Based on this source space, ℓ2 minimum norm current estimates [56, 57] were computed for all data using a depth weighting parameter of 0.8 [58]. Analysis was restricted to the temporal lobe of both hemispheres, based on anatomical labels in the "aparc" parcellation [59]. This resulted in 314 source dipoles in the left hemisphere and 313 source dipoles in the right (see highlighted area in Figure 1).

QUANTIFICATION AND STATISTICAL ANALYSIS

Predictor variables

Predictor variables were generated as uniform time series with a sampling rate of 100 Hz to match the processed MEG data. Figure 1 illustrates the predictor variables, aligned with an excerpt from one of the stimuli.

Responses to the acoustic features of the speech signal were modeled using a model of acoustic transformations in the brainstem, the so called auditory spectrogram [60]. The auditory spectrogram was computed using the NSL Toolbox (https://isr.umd.edu/Labs/NSL/Software.htm) and shifted by −20 ms in order to compensate for the intrinsic delay introduced by this transformation. A predictor reflecting the moment by moment acoustic envelope was generated by summing the auditory spectrogram in 8 logarithmically spaced frequency bands (8 bands were chosen as a compromise between reducing the computational demand for model fitting, while still being able to recognize phonetically relevant acoustic features in the spectrogram). Because brain responses are known to be sensitive to contrast and changes, and phonetic information is often specifically located in acoustic onsets [20], it was important to control for responses to onsets in the acoustic signals. For this reason, an acoustic onset predictor was constructed from the half-wave rectified derivative of the acoustic envelope predictor.

All phoneme-based predictors were modeled as impulses at phoneme onset (see Figure 1). Phoneme onsets in the stimuli were automatically determined by the Gentle forced aligner (https://lowerquality.com/gentle/) and then adjusted by hand where appropriate using Praat [61]. A phonetic lexicon with lexical statistics was generated by combining pronunciations from the CMU Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict) and word frequency statistics from the SUBTLEX subtitle database [23]. Stress information was stripped from all phonemes. Missing pronunciations were manually added, and words occurring in the stimuli but missing from SUBTLEX were assigned a frequency count of 1.

The cohort refers to the set of words compatible with the acoustic input at any point during a word [10]. For each phoneme, the cohort was determined by selecting from the phonetic lexicon those entries that started with the phoneme sequence from the beginning of the word to the current phoneme. The cohort size variable was the log of the number of words in the cohort at each phoneme. The cohort reduction variable was the log of the number of words at the current phoneme minus the number of words at the previous phoneme or, for the initial phoneme, minus the number of words in the whole lexicon. While these two variables are not as widely used as surprisal and entropy (see below), they are potentially more fundamental variables that should be controlled for before drawing conclusions about surprisal and entropy.

While cohort size variables depend only on word counts, the frequency with which individual words occur in the language is known to affect lexical processing [11, 22]. This is taken into account by the measures of phoneme surprisal and entropy. Phoneme surprisal is defined as the inverse of the conditional probability of each phoneme, given the preceding phonemes in the current word:

$$surprisal_i = -\log_2\left(\frac{freq(cohort_i)}{freq(cohort_{i-1})}\right)$$

where $cohort_i$ is the cohort at phoneme with position i, and $freq(c)$ is the summed frequency of all words in cohort c. Cohort entropy is defined as the entropy [24] of the cohort at each phoneme. Entropy at phoneme i is given by:

$$H_i = -\sum_{word}^{cohort_i} p_{word} \log_2 p_{word}$$

where p_{word} is the probability of the given word form, here modeled as the relative frequency count in the SUBTLEX corpus.

To account for the possibility that the initial phoneme of each word is processed differently from the subsequent phonemes (see e.g., [10]), we modeled the initial phoneme of each word separately from the subsequent phonemes for each variable (indicated by a different color of the word-initial phonemes in Figure 1).

Model estimation

For each subject, the localized current at each potential neural source dipole was modeled as a sum of linear convolutions of the stimulus variables with a filter of 500 ms duration. Separate and independent models were estimated for each of the 627 source dipoles. Optimal filters were estimated for all predictor variables concurrently with a coordinate descent algorithm [13, 62] minimizing the $\ell 1$ error between predicted and actual current time course. Filters were generated from a basis of 50 ms Hamming windows, centered at each time point in the kernel. This smoothness constraint on the filters was imposed to improve the reliability of predictions, compensating for the temporal sparseness of the impulse representation of phonemes. Algorithms used for model estimation and statistical analysis are publicly available in the Eelbrain open source Python package [63] (https://github.com/christianbrodbeck/eelbrain).

Model comparisons

Model fit was evaluated using the z-transformed Pearson correlation between estimated and measured responses (the Fisher z-transformation corrects for distortions introduced by the fixed end-points at 1 and -1 of correlation coefficients). Model fit z-maps were smoothed with a Gaussian kernel (STD = 5 mm) to account for granularity caused by local variation in source dipole orientation. To compare the fit of two models, their respective z-maps were compared with related measures t tests. First, a t map was computed for the difference at each source dipole. The resulting map was then processed with threshold-free cluster enhancement (TFCE) [64], and a distribution of the largest expected TFCE value per t map under the null-hypothesis was computed with 10,000 permutations, randomly switching condition labels within subjects [64, 65]. A p value for each dipole was computed by locating the original TFCE-enhanced t value on the permutation distribution. Along with the permutation-based p value, we report t_{max}, the largest t value from the given comparison's t map (using absolute t values for two-tailed tests).

To test for significant contributions of a given predictor, the predictive power of the full model was compared to the average of 3 models consistent with the null hypothesis, which were identical except for the predictor under investigation, each shuffled in a way appropriate for the hypothesis being tested. A predictor was considered significant if it significantly improved model fit across participants over the respective shuffled models. Three shufflings were used to decrease the influence of arbitrary features of a single randomization. This procedure allowed testing for incremental model improvement due to a specific predictor, without introducing bias by changing the degrees of freedom. Under the null hypothesis that there is no significant association between the given predictor and the responses, a shuffled version of the predictor should be equally effective as the properly aligned version. A difference in model fit between the full and the shuffled models thus indicates a significant relationship between predictor and responses.

Test of lateralization

Tests of hemispheric asymmetry were performed by comparing model-fit improvement between the two hemispheres. First, a difference map was computed by subtracting, from the z-values of the full model, the average from the 3 shuffled models. The resulting difference maps from both hemispheres were mapped to the left hemisphere of the "fsaverage_sym" brain [13, 66], masked by the region of significant model improvement in at least one hemisphere, and compared with a two-tailed t test while controlling for multiple comparisons with TFCE as described above.

A potential concern with source localized MEG data could be differential sensitivity in the two hemispheres. However, a laterality test of the reduced model as a whole for the single speaker data indicated that there was no overall difference in predictive power in the left and right hemisphere (t_{max} = 2.94, p = 0.142). Together with the result that significant lateralization was observed in both directions for different variables, this suggests that bias toward one hemisphere was minimal.

Spatio-temporal response functions

Temporal response functions, i.e., the kernels of the optimal filters, were analyzed similarly, but including the additional dimension of time. A spatio-temporal t-map was computed for a one-sample t test against 0. This map was again processed with TFCE and a two-tailed distribution for the maximum TFCE value was computed based on 10,000 permutations. For graphical display only, time series were upsampled to 1000 Hz to minimize visual discretization artifacts.

Response time course

The time points of response peaks were estimated from the group average TRFs. The spatio-temporal response functions were masked by significance, and their absolute values were summed across sources. The resulting time course was resampled to 1000 Hz, and response peaks were identified as local maxima (time courses are displayed in Figure 4B).

In order to compare response peak latencies across the phoneme-based variables, a spatio-temporal mask was generated as the union of the masks for the individual predictors (phoneme onset, word onset, phoneme surprisal, and cohort entropy). As above, the response functions were masked, absolute values were summed across source dipoles and the resulting time course was resampled to 1000 Hz. For pairwise comparisons, the response peak with the largest amplitude was identified for each subject for each variable in the window from 60 – 160 ms. This window was centered around the main peaks of the four variables, which all occurred between 103 and 125 ms, and largely encompassed the temporal extent of significant responses. Peaks from different variables were compared using two-tailed t tests.

Response localization

The main peaks in the response functions were compared as to their spatial localization. Since the responses to the main variables of interest were dominant in the left hemisphere, this analysis was restricted to the left hemisphere. For each response peak in the average response, a spatial map was generated based on the sum of the absolute TRFs in a time window of 60 ms centered on that peak (since response peaks were identified in upsampled time courses and usually lay between two actual TRF samples, this amounted to including 3 samples on either side of the peak). Each map was thresholded at half its maximum value to reduce the influence of spatial leakage of source estimates, and the center of mass was computed as the weighted average of the corresponding FreeSurfer fsaverage right/anterior/superior coordinates.

A permutation test was used to test the null hypothesis that the location of two peaks was indistinguishable. For each subject, the difference vector between the two locations was computed. The length of the average vector served as a statistic of intertest, and its distribution under the null hypothesis was determined in 10,000 permutations in which each vector was rotated by a randomly determined angle. Results of pairwise comparisons are listed in Table S3.

Due to the finite extent of the source space volume, averages of individual subjects' centers of mass were biased toward the center of the source space, making them unsuitable for visualization (Figures 4 and S2). Instead, centers of mass were computed for average TRF peak maps (see Figure S2). For this purpose, individual subject peak maps were normalized and averaged (before thresholding). Then, the center of mass was computed as described above, but using the coordinates of the inflated brain used for visualization. For visualization in Figure S2, the peak maps were smoothed with a Gaussian kernel (STD = 5 mm).

Single speaker analysis

Responses to a single speaker were used to determine variables that reflect lexical processing of phonetic information. To test for an effect of lexical variables without inflated type I error due to multiple comparisons, an initial test was performed against shuffled models in which all 8 lexical variables were shuffled together. Subsequently, the set of lexical variables was reduced to a set in which each variable explained a distinct proportion of the variance. To this end, the model was reduced one predictor at a time by removing the predictor with the largest p value, until only significant predictors were left (see e.g., reference [25] for a similar approach). Once only significant lexical predictors remained (henceforth called the reduced model), the other variables in the model were also evaluated for significance (shown in Figure 2).

The way in which variables were shuffled depended on the nature of the variable and the corresponding null hypothesis: Lexical variables were shuffled by randomly reordering the values (e.g., phoneme surprisal) while leaving the phoneme time locations constant. Acoustic predictors were shuffled by shifting the acoustic stimulus in time by 15, 30 or 45 s (including wrapping around from the end to the beginning). To test whether word onsets were represented neurally, word onsets were randomly assigned to different phoneme locations, while keeping all phoneme locations constant. To test whether phoneme onsets were represented, the time-series modeling non-word initial phonemes (i.e., all phonemes except those at word onset) was time-shifted in the same manner as the acoustic predictors. In each case, all remaining predictors were left unchanged in the control model.

Two speaker analysis

For modeling responses to stimuli with a mixture of two speakers, separate predictors were included for the attended and the unattended speech stream, both based on the reduced single speaker model. In addition, acoustic predictors were generated for the acoustic mixture of the two speakers, i.e., for the raw acoustic stimulus that was actually presented to the participants.

Models were assessed by grouping the predictors modeling each process of interest. Since the acoustic mixture is closely approximated by a linear combination of the attended and the unattended signals (in all bands $r > 0.95$ for the acoustic envelope and $r > 0.88$ for acoustic onsets), the predictive power of the mix could not be assessed independently. Instead, the predictive power of the attended stimulus was assessed by shuffling both the attended and the mix acoustic predictors, and the unattended stimulus was assessed by shuffling both the unattended and the mix acoustic predictors. Nevertheless, acoustic TRFs could be analyzed for all three streams, since the coordinate descent algorithm determines which predictor can reduce the error most efficiently, regardless of whether the same model fit could be achieved by a linear combination of other, less efficient predictors. Lexical processing was as-

sessed separately for attended and unattended streams, by shuffling all the values among phoneme locations, but leaving phoneme locations themselves unchanged. Thus, the model comparison controlled for responses associated with all phonemes independent of lexical processing.

TRF latencies were determined as for the single speaker responses. In order to compare latencies from the two speaker responses to the latencies from the single speaker responses, the same mask and procedure as for the peak analysis of single speaker responses was used, but the time window for analysis was extended by 20 ms to account for the increase in latency clearly present in the averaged responses.

No detectable effect of repeated presentation

A potential concern in the present experiment was that stimuli were repeated multiple times. While each stimulus was a full minute long, making it unlikely that participants were able to form a detailed memory over the course of a few presentations, repeated presentation might nevertheless have led to subtle changes in processing. The effect of previous exposure on lexical responses was assessed separately for clean speech and two-talker mixed speech.

For clean speech, separate models were fit to stimuli based on whether the speech had been previously attended or ignored, respectively, when presented as part of two-talker mixed speech. Listeners attended to the same speaker during each presentation of a mixed stimulus, e.g., always attending to the female-1 stimulus when it was presented mixed with the male-1 stimulus. Consequently, subsequently listening to female-1 alone should essentially amount to a fifth repetition, while listening to male-1 should amount to a new stimulus. TRFs to previously attended and ignored stimuli were compared for all lexical predictors using paired t tests and TFCE, analogous to tests described above. The statistical analysis was restricted to response latencies between 0 and 250 ms, which encompassed the main responses found in the main analysis. None of the TRFs differed significantly (word onset: t_{max} = 3.82, p = 0.176; phoneme surprisal: t_{max} = 4.59, p = 0.208; cohort entropy: t_{max} = 3.03, p = 0.119). For increased power, the analysis was repeated with TRFs masked by the spatio-temporal region in which each predictor was significantly different from 0 in the reduced model. This did not change the conclusions (word onset: t_{max} = 3.82, p = 0.221; phoneme surprisal: t_{max} = 2.54, p = 0.181; cohort entropy: t_{max} = 2.55, p = 0.255). Time courses of responses appeared very similar, and this was confirmed by an analysis of peak latencies, using the same methods as described above for the comparison of peak latencies between predictors (word onset: $t_{(25)}$ = 0.36, p = 0.725; phoneme surprisal: $t_{(25)}$ = 0.65, p = 0.521; cohort entropy: $t_{(25)}$ = 0.78, p = 0.441).

For the two-speaker mix, separate models were fit to the first, second, third, and fourth presentations of the two speaker stimuli. Spatio-temporal one-way ANOVA with four levels representing order of presentation did not reveal any differences in TRFs to lexical processing of the foreground speech between 0 and 250 ms (word onset: f_{max} = 5.01, p = 0.316; phoneme surprisal: f_{max} = 6.65, p = 0.110; cohort entropy: f_{max} = 7.46, p = 0.656; masked by significant response in the main analysis, word onset: f_{max} = 4.36, p = 0.472; cohort entropy: f_{max} = 4.65, p = 0.316). An analysis of peak latency also did not indicate a difference due to how many times a stimulus had been heard (word onset: $f_{(3,75)}$ = 0.06, p = 0.982; cohort entropy: $f_{(3,75)}$ = 0.60, p = 0.616).

DATA AND SOFTWARE AVAILABILITY

Algorithms used for model estimation and statistical analysis are available in the Eelbrain open source Python package [63] (https://github.com/christianbrodbeck/eelbrain). Data are available from the Digital Repository at the University of Maryland (http://hdl.handle.net/1903/21109).

Current Biology, Volume *28*

Supplemental Information

Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech

Christian Brodbeck, L. Elliot Hong, and Jonathan Z. Simon

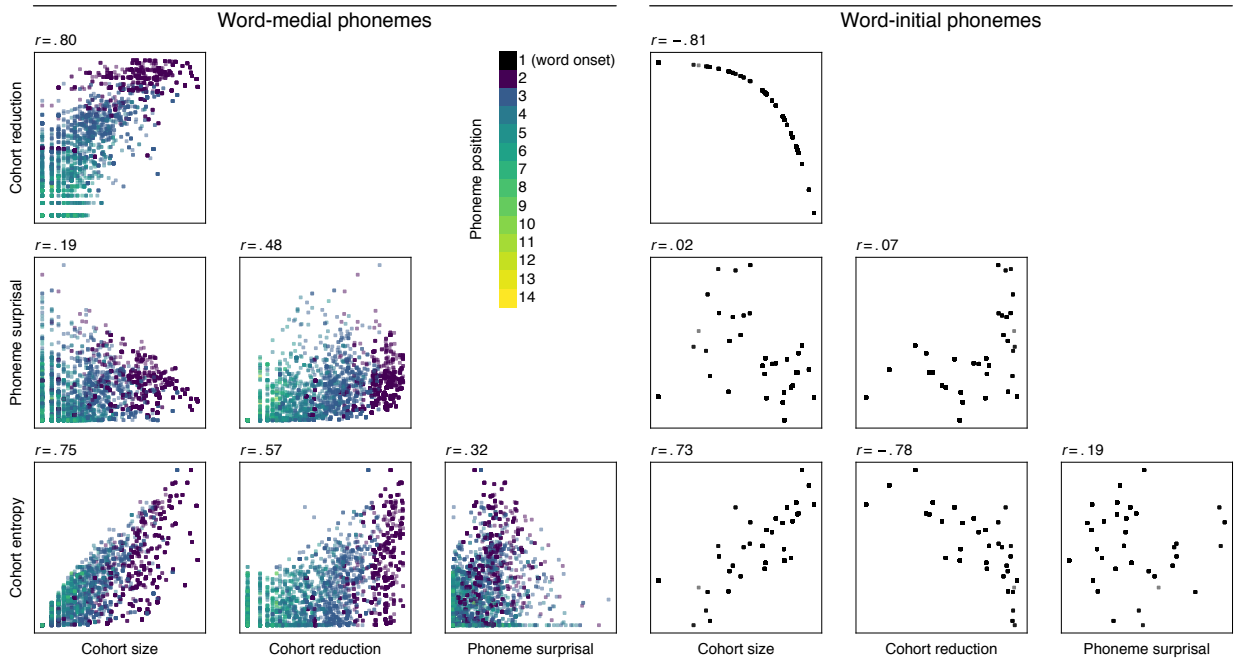

Figure S1. Relationship between predictor variables, across all stimuli. Related to Figure 1 and Table S1. Each data point reflects one phoneme. Corresponding correlation values are also listed in Table S1.

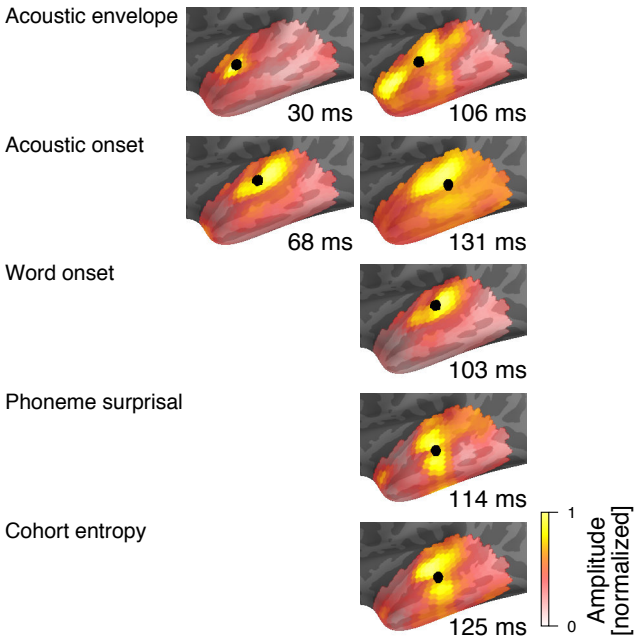

Figure S2. TRF peak maps. Related to Figures 2 and 4 and Table S3. Average of subject maps for all major TRF peaks, averaged in 60 ms windows around peaks. Black circles indicate the center of mass of each map, calculated as described in the Methods section and displayed on Figure 4. See Table S3 for pairwise tests of peak locations.

	Acoustic envelope	Acoustic onset	Cohort size	Cohort reduction	Phoneme surprisal
Acoustic onset	.44				
Word-medial					
Cohort size	.01 - .12	.03 - .19			
Cohort reduction	.01 - .12	.03 - .20	.80		
Phoneme surprisal	.01 - .08	.02 - .13	.19	.48	
Cohort entropy	.00 - .11	.01 - .19	.75	.57	.32
Word-initial					
Cohort size	-.07 - .00	-.03 - .05			
Cohort reduction	-.07 - .00	-.03 - .04	-.81		
Phoneme surprisal	-.07 - .00	-.03 - .04	.02	.07	
Cohort entropy	-.07 - .01	-.03 - .03	.73	-.78	.19

Table S1. Predictor correlation. Related to Figure 1. Pairwise correlation for predictor variables across all stimuli. For correlations between the two acoustic predictors, the correlation reflects all samples across time and center frequency. For correlations between acoustic and phoneme-based predictors, correlations were computed separately for each frequency band across all time samples, and the range (min – max) of correlations with the different frequency bands is given. For correlations between phoneme-based variables, correlations between phoneme values, i.e., the values of the non-zero impulses, were computed. Corresponding scatter-plots for phoneme-based variables are displayed in Figure S1.

		Word-medial				**Word-initial**			
		Cohort size	Cohort reduction	Phoneme surprisal	Cohort entropy	Cohort size	Cohort reduction	Phoneme surprisal	Cohort entropy
1	t_{max}	1.47	2.68	3.87**	4.61***	2.21	2.81	2.88	2.54
	p	.998	.870	.006	< .001	.970	.307	.558	.832
2	t_{max}		2.95	3.71**	5.10***	2.03	3.37	2.51	2.85
	p		.517	.006	< .001	.996	.119	.903	.748
3	t_{max}		3.24	3.95**	4.85***		2.76	2.41	2.50
	p		.263	.004	< .001		.489	.938	.867
4	t_{max}		3.70	3.94**	5.09***		2.78		2.37
	p		.099	.002	< .001		.417		.967
5	t_{max}		3.57	3.74**	5.49***		3.40		
	p		.105	.004	< .001		.086		
6	t_{max}			3.98**	6.04***		3.00		
	p			.002	< .001		.397		
7	t_{max}			4.47***	5.68***				
	p			< .001	< .001				

Table S2. Model reduction. Related to Figure 2. Each row constitutes one step in the model reduction. The row provides t_{max} and *p*-values for each predictor variable (significance marked * ≤ .05; ** ≤ .01; *** ≤ .001). The variable with the lowest non-significant effect size (greatest *p*-value) was excluded for the next row, until only significant predictor variables remained.

		Acoustic envelope		Acoustic onset		Word onset	Phoneme surprisal
		30 ms	106 ms	68 ms	131 ms	103 ms	114 ms
Acoustic envelope	30 ms						
	106 ms	2					
Acoustic onset	68 ms	8**	6				
	131 ms	11***	9***	7**			
Word onset	103 ms	10***	9**	3	5		
Phoneme surprisal	114 ms	5	4	7	7	8*	
Cohort entropy	125 ms	8**	6	7**	4	7**	3

Table S3. Pairwise tests of peak locations. Related to Figures 2 and 4. For each subject, the center of mass of the TRF at a given peak was extracted, and the resulting center coordinates were compared with pairwise permutation-based tests (see Methods section). Each cell displays the distance in mm and corresponding significance test (* ≤ .05; ** ≤ .01; *** ≤ .001). See Figure S2 for average peak distribution maps. Note that center of mass estimates are biased towards the center of the source space volume, and distances are thus smaller than distances between the peaks of average maps.