Dynamic Processing of Background Speech at the Cocktail Party: Evidence for Early Active Cortical Stream Segregation

Christian Brodbeck[1,*], Alex Jiao[2], L. Elliot Hong[3] & Jonathan Z. Simon[1,2,4]

1) Institute for Systems Research, University of Maryland, College Park, Maryland 20742, U.S.A
2) Department of Electrical and Computer Engineering, University of Maryland, College Park, Maryland 20742, U.S.A
3) Department of Psychiatry, Maryland Psychiatric Research Center, University of Maryland School of Medicine, Baltimore, Maryland 21201, U.S.A
4) Department of Biology, University of Maryland, College Park, Maryland 20742, U.S.A

* christianbrodbeck@me.com

Abstract

Humans are remarkably skilled at listening to one speaker out of an acoustic mixture of speech sources, even without binaural cues, but the neural mechanisms underlying this ability are not well understood. One possibility is that early cortical processing performs a spectro-temporal decomposition of the acoustic mixture, allowing the attended speech to be reconstructed via optimally weighted recombinations that discount spectro-temporal regions where sources heavily overlap. Using human magnetoencephalographic responses to a two-talker mixture, we show evidence for an alternative possibility, in which early, active segregation occurs even for strongly spectro-temporally overlapping regions. Early (~70 ms) responses to individual non-overlapping spectro-temporal features are seen for both talkers. When competing talkers' spectro-temporal features mask each other, the individual representations persist but now with a ~20 ms delay, suggesting an active segregation process with a temporal processing cost. Such active segregation could explain behavioral effects of ignored background speech.

Keywords

Auditory streaming, auditory cortex, continuous speech, temporal response function

Introduction

When listening to an acoustic scene, the acoustic signal that arrives at the ears is an additive mixture of the different sound sources. Listeners trying to selectively attend to one of the sound sources face the task of deciding which spectro-temporal features belong to that source. When multiple speech sources are involved, as in the classic cocktail party problem (Cherry, 1953), this is a nontrivial problem because the spectrograms of the different sources often have strong overlap. Nevertheless, human listeners are remarkably skilled at focusing on one out of multiple talkers (McDermott, 2009; Middlebrooks et al., 2017). Binaural cues can support segregation of different sound sources based on their location (Brungart and Simpson, 2002), but are not necessary for this ability, since listeners are able to selectively attend even when two speech signals are mixed into a monophonic signal and presented with headphones (Kidd et al., 2016).

The mechanisms involved in this ability are not well understood, but previous research suggests at least two separable cortical processing stages. In magnetoencephalographic (MEG) recordings of subjects listening to multiple talkers (Puvvada and Simon, 2017), the early (~50 ms) cortical response component is better described as a response to the acoustic mixture than as the sum of the responses to the individual (segregated) source signals, consistent with an early unsegregated representation of the mixture. In contrast, the later (> 85 ms) response component is dominated by the attended (segregated) source signal. Recent direct cortical recordings largely confirm this picture, suggesting that early responses in Heschl's Gyrus (HG) reflect a spectro-temporal decomposition of the acoustic mixture that is largely unaffected by selective attention, whereas later responses in the superior temporal gyrus (STG) dynamically change to represent the attended speaker (O'Sullivan et al., 2019). In general, cortical regions further away from core auditory cortex tend to mainly reflect information about the attended speaker (Zion Golumbic et al., 2013). These results suggest a cortical mechanism which detects and groups features belonging to the attended source, extracted from a detailed representation of the acoustic input. A long-standing question concerning this mechanism is whether early representations are restricted to passive spectro-temporal filtering or whether they also involve active extraction of acoustic features, and to what degree such features may also be actively segregated and represented as auditory objects.

While time-locked processing of words and higher order linguistic features seems to be restricted to the attended speech source (Brodbeck et al., 2018a; Broderick et al., 2018), it is not known whether, in the course of recovering features of the attended source, the auditory cortex also segregates acoustic features of the ignored source from the mixture. Individual intracranially recorded HG responses to a two-speaker mixture can be selective for either one of the speakers, but this selectivity can be explained merely by spectral response characteristics favoring the spectrum of a given speaker over the other (O'Sullivan et al., 2019). A conservative hypothesis is thus that early auditory cortical responses represent acoustic features of the mixture based on stable (possibly pre-defined) spectro-temporal receptive fields, allowing the attended speech to be segregated through an optimally weighted combination of these responses. Alternatively, the auditory cortex could employ an active process to dynamically recover

and represent potential speech features regardless of what stream they belong to, and in doing so provide selective attention with more elaborate representations (Carlyon, 2004). Critically, this hypothesis predicts the existence of representations of acoustic features from an ignored speech source, even when those features are not apparent in the acoustic mixture, i.e., when those features are masked by acoustic energy from another source. Here we report evidence for such representations in human MEG responses.

Interference between individual speech sources, for a small number of talkers, is relatively hard to discern in acoustic *envelope* features (e.g., the speech spectrogram); quantitatively, the envelope of the acoustic mixture is strongly correlated with the sum of the envelopes of the individual speech sources (Puvvada and Simon, 2017). Instead, we here include analysis with respect to acoustic *onset* features, i.e. spectro-temporally local increases in acoustic energy. Acoustic onsets are prominently represented in auditory cortex, both in naturalistic speech (Daube et al., 2019; Hamilton et al., 2018) and in non-speech stimuli (Fishbach et al., 2001; Zhou and Wang, 2010), and are also important for speech intelligibility (Stilp and Kluender, 2010). With acoustic onset features, interference is more pronounced because ongoing acoustic energy in one source can hide an increase in another source (see below for empirical verification of this claim). Behaviorally, acoustic onset features are highly relevant because they are important cues for auditory object perception. Simultaneous onsets in distinct frequency bands create a strong cue that these cross-frequency features originate from the same acoustic source. Accordingly, shared acoustic onsets promote perceptual grouping of acoustic elements into a single auditory object, such as a complex tone and, vice versa, separate onsets lead to perceptual segregation (Bregman et al., 1994a, 1994b). For example, the onset of a vowel is characterized by a shared onset at the fundamental frequency of the voice and its harmonics. If the onset of a formant is artificially offset by as little as 80 ms, it can be perceived as a separate tone rather than as a component of the vowel (Hukin and Darwin, 1995). Because onsets are such strong cues for feature grouping, they are also important cues for segregating distinct acoustic sources from a mixture (Elhilali et al., 2009).

We analyzed human MEG responses to a continuous two-talker mixture to determine to what extent the auditory cortex reliably tracks acoustic onset or envelope features of the ignored speech (above and beyond reliable tracking of the attended speech and the mixture). Participants listened to 1-minute long continuous audiobook segments, spoken by a male or a female speaker. Segments were presented in two conditions: as speech in quiet (“clean speech”), and as a two-talker mixture, in which a female and a male speaker were mixed at equal loudness. MEG responses were analyzed as additive, linear response to multiple concurrent stimulus features (see Figure 1). First, model comparison was used to determine which representations significantly improved prediction of the responses. Then, spectro-temporal response functions (STRFs) were analyzed to gain insight into the nature of the representations.

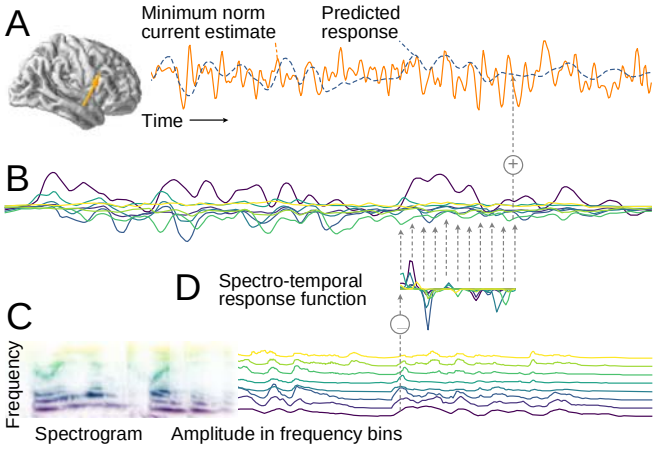

Figure 1. **Additive linear response model based on spectro-temporal response functions (STRFs)**. A) MEG responses recorded during stimulus presentation were source localized with distributed minimum norm current estimates. A single virtual source dipole is shown for illustration, with its physiologically measured response and the response prediction of a model. Model quality was assessed by the correlation between the measured and the predicted response. B) The model's predicted response is the sum of tonotopically separate response contributions generated by convolving the stimulus envelope at each frequency (C) with the estimated temporal response function (TRF) of the corresponding frequency (D). TRFs quantify the influence of a predictor variable on the response at different time lags. The stimulus envelopes at different frequencies can be considered multiple parallel predictor variables, as shown here by the gammatone spectrogram (8 spectral bins); the corresponding TRFs as a group constitute the spectro-temporal response function (STRF). Physiologically, the component responses (B) can be thought of as corresponding to responses in neural subpopulations with different frequency tuning, with MEG recording the sum of those currents.

Results and Discussion

Auditory cortex represents acoustic onsets

MEG responses to clean speech were predicted from the gammatone spectrogram of the stimulus, and, simultaneously, from the spectrogram of acoustic onsets (Figure 2-A). Acoustic onsets were derived from a neural model of auditory edge detection (Fishbach et al., 2001). Both predictors were binned into 8 frequency bands, for a total of 16 predictor time series. Each of the two predictors was assessed based on how well the correct model predicted MEG responses, compared to null models in which the relevant predictor was temporally misaligned with the responses. Both predictors significantly improve predictions ($p \leq 0.001$), with an anatomical distribution consistent with sources in HG and STG bilaterally (Figure 2-B). Since this localization agrees with findings from intracranial recordings (Hamilton et al., 2018; O'Sullivan et al., 2019), results were henceforth analyzed in an auditory region of interest (ROI) restricted to these two anatomical landmarks (Figure 2-C).

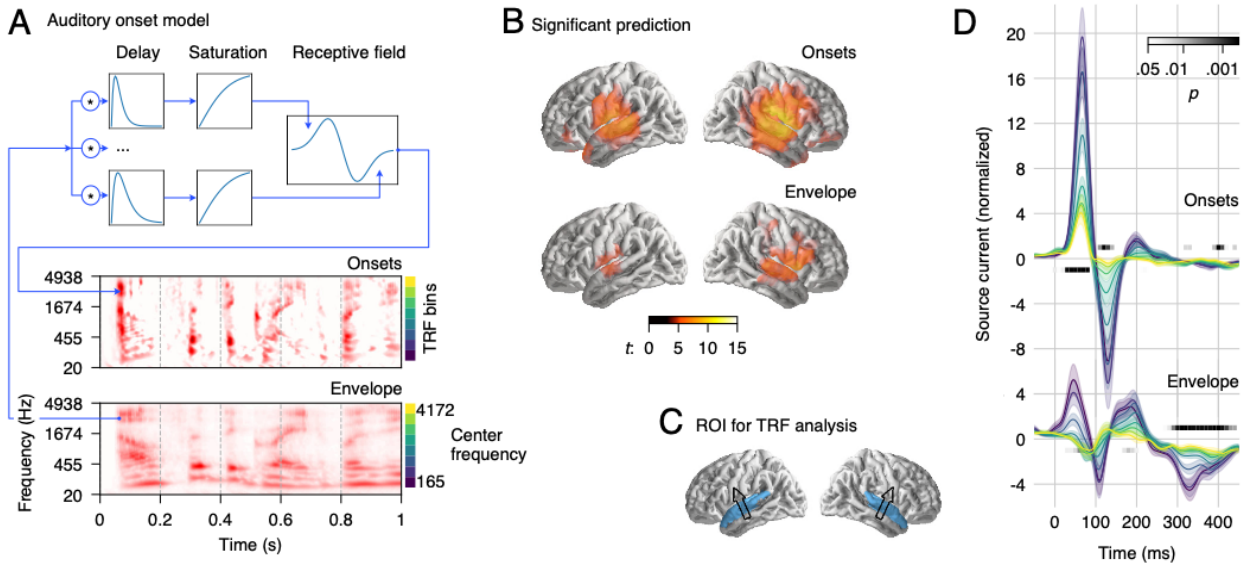

Figure 2. **Acoustic onset responses to clean speech**. A) Schematic illustration of the neurally inspired acoustic edge detector model, which was used to generate onset representations, along with an excerpt from a gammatone spectrogram ("envelope") and the corresponding onset representation. B) Regions of significant explanatory power of onset and envelope representations, consistent with a main source in auditory cortex bilaterally ($p \leq .05$, corrected for whole brain analysis). C) Region of interest (ROI) used for the analysis of response functions, including superior temporal gyrus and Heschl's gyrus. An arrow indicates the average dominant current direction in the ROI (upward current), determined through the first principal component of response power. D) Spectro-temporal response functions corresponding to onset and envelope representations in the ROI; the initial peaks of the onset representation are more pronounced and well-defined than those of the envelope representation. Different color curves reflect the frequency bins as indicated next to the onset and envelope spectrograms in panel A. Shaded areas indicate the within-subject standard error (SE) (Loftus and Masson, 1994). Regions in which STRFs differ significantly from 0 (in any band) are marked with horizontal gray bars.

Auditory cortical STRFs were computed for each participant and hemisphere using a subject-specific spatial filter based on principal component analyses of overall STRF power in the ROI. The average direction of that spatial filter replicates the direction of the well-known auditory MEG response with dominantly vertical orientation (Figure 2-C). STRFs were initially separately analyzed by hemisphere, but since none of the reported results interacted significantly with hemisphere the results shown are collapsed across hemisphere to simplify presentation.

STRFs to acoustic onsets exhibit a well-defined two-peaked shape, consistent across frequency bands (Figure 2-D). They closely resemble previously described auditory response functions to envelope representations when estimated without consideration of onsets (Ding and Simon, 2012). In comparison, envelope STRFs in the present results are diminished and exhibit a less well-defined structure. This is consistent with acoustic onsets explaining a large portion of the signal usually attributed to the envelope; indeed, when the model was refitted with only the envelope predictor,

excluding the onset predictor, the envelope STRFs exhibited that canonical pattern and with larger amplitudes (not shown).

STRFs have disproportionately higher amplitudes at lower frequencies (Figure 2-D), which is consistent with previous tonotopic mapping of speech areas and may follow from the spectral distribution of information in the speech signal (Hullett et al., 2016; Moerel et al., 2012). This explanation is also supported by simulations, where responses were generated using equal TRFs for each band, and yet estimated STRFs exhibited higher amplitudes in lower frequency bands (Figure SI-1).

Auditory cortex represents onsets of ignored speech

MEG responses to a two-speaker mixture were then used to test for a neural representation of ignored speech. Participants listened to an equal loudness mixture of a male and a female talker and were instructed to attend to one talker and ignore the other. The speaker to be attended was counterbalanced across stimulus and subjects. Responses were predicted using the onset and envelope representations for the acoustic mixture, the attended speech source and the ignored source (Figure 3-A). Taken together, the two predictors representing the ignored speech significantly improve predictions of the responses in the ROI (t_{max} = 8.32, p < .001). This indicates that acoustic features of the ignored speech are represented neurally even after controlling for features of the mixture and the attended source. Separate tests suggest that this result can be ascribed specifically to onset representations (t_{max} = 4.89, p < .001), whereas the envelope representations of the ignored source does not significantly improve the model fit (t_{max} = -2.59, p = 1).

Taken individually, each of the three onset representations significantly improves model predictions (t_{max} ≥ 4.89, p < .001), but none of the envelope representations does (all t_{max} ≤ -0.40, p = 1). This lack of predictive power for the envelope predictors, when tested individually, is likely due to their high collinearity. Intuitively, the envelope of the mixture can be approximated relatively well by the sum of the envelopes of the individual streams (cf. Figure 3-A). More formally, the proportion of the variability in the mixture representations that cannot be predicted from the two sources is small for the envelopes, but substantially larger for the onsets (Figure 3-C). Accordingly, when the mixture envelope predictor was removed from the model, the two source envelope predictors became significant individually (attended: t_{max} = 4.72, p = .002; ignored: t_{max} = 2.93, p = .042). Thus, as far as the envelope representations are concerned, the nature of the stimulus representations prevents a conclusive distinction between representations of the acoustic mixture and the ignored source. In contrast, onset representations indicate a reliable representation of ignored speech over and above representations of the acoustic mixture and the attended source.

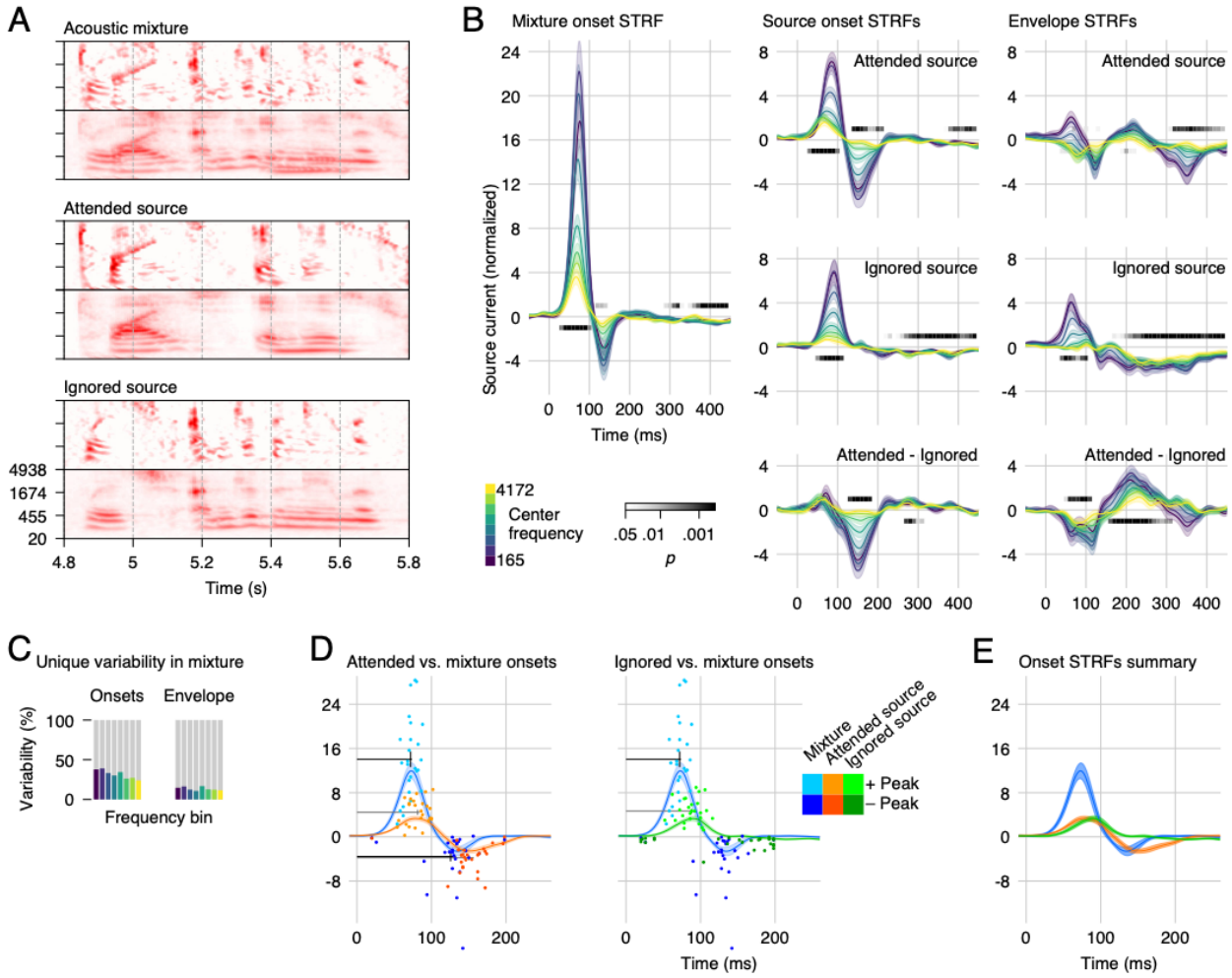

Figure 3. **Response functions to the two-speaker mixture, using the stream-based model**. A) The envelope and onsets of the acoustic mixture and the two speech sources were used to predict MEG responses. B) Auditory cortex STRFs to onsets in the mixture exhibit a large positive peak (72 ms) followed by a smaller negative peak (126 ms). STRFs to attended and ignored onsets both exhibit an early positive peak (81 and 88 ms), followed only in attended onsets by a negative peak (150 ms). This effect of attention on the negative peak is confirmed by the attended – ignored STRF differences. Envelope STRFs have lower amplitudes but show a strong effect of attention. C) Compared to envelope representations, acoustic onset representations are better suited for distinguishing segregated sources from the mixture. Colored portions indicate proportion of the variability of the mixture predictors that cannot be explained from the individual speech sources (with a -500 – 500 ms temporal integration window). D) The major peaks from onsets in the speech sources are delayed compared to corresponding peaks from the mixture; the early, positive peak is also diminished. To determine latencies, mixture-based and individual-speaker-based STRFs were averaged across frequency (lines with shading for mean ±1 SE). Colored dots represent the largest positive and negative peak for each participant between 20 and 200 ms; the peaks corresponding to individual speakers are delayed with respect the corresponding peaks for the mixture. Horizontal bars indicate average amplitude and latency ±1 SE. E)

Direct comparison of the same onset response functions averaged across frequency, ±1 SE. Peaks from onsets in individual speech sources are delayed compared to corresponding peaks from the mixture; the early, positive peak is also diminished.

Onset STRFs exhibit the same characteristic positive-negative pattern as for speech in quiet, but with reliable distinctions between the mixture and the individual speech streams (Figure 3-B, left and middle columns, Figure 3-D & E). The early, positive peak occurs earlier and has a larger amplitude for onsets in the mixture than for onsets in either of the sources (latency mixture: 72 ms; attended: 81 ms, $t_{25} = 4.47$, $p < .001$; ignored: 88 ms, $t_{25} = 6.92$, $p < .001$; amplitude mixture > attended: $t_{25} = 8.60$, $p < .001$; mixture > ignored: $t_{25} = 7.92$, $p < .001$). This positive peak is followed by a negative peak only in responses to the mixture (126 ms) and the attended source (150 ms; difference $t_{25} = 4.36$, $p < .001$). The amplitude of these negative peaks is statistically indistinguishable ($t_{25} = 0.36$, $p = .722$). STRFs to the ignored source do not exhibit a detectable corresponding negative peak, as seen in Figure 3-D where participants' peaks cluster around the time window edges instead of at a characteristic latency.

The fact that the mixture predictor is not completely orthogonal to the source predictors might raise a concern that a true response to the mixture might cause spurious responses to the sources. Simulations using the same predictors as used in the experiment suggest, however, that such contamination is unlikely to have occurred (see Figure SI-1).

Envelope processing is strongly modulated by selective attention

In contrast to onsets, the different envelope predictors did not contain enough independent information to distinguish between a representation of the ignored source and a representation of the mixture. However, a comparison of the STRFs to the attended and the ignored source revealed a strong effect of attention (Figure 3-B, right column). The attended-ignored difference wave exhibits a negative peak at ~100, consistent with previous work (Ding and Simon, 2012), and an additional positive peak at ~200 ms. In contrast to previous work, however, a robust effect of attention on the envelope representation starts almost as early as the earliest responses at all. Thus, accounting for responses to onset features separately reveals that envelope processing is thoroughly influenced by attention. The reason for this might be that onsets often precede informative regions in the spectrogram, such as the spectral detail of voiced segments. The onsets might thus serve as cues to direct attention to specific regions in the spectrogram (Bregman et al., 1994a), which would allow early attentional processing of the spectrogram, or envelope features.

Auditory cortex "un-masks" masked onsets

The analysis using the stream-based predictors suggests that the auditory cortex represents acoustic onsets in both speech sources separately, in addition to onsets in the acoustic mixture. This result could be due to a passive process, whose action does not depend on whether a source feature is masked or not, or an active process that can isolate onsets in an individual speech stream even if masked by the other speech stream. An example of the former could be a secondary stage of onset extraction after spectral sharpening. Such spectral sharpening might resolve some of the overlap among the two speakers, and thus reveal additional source specific onsets.

Alternatively, the auditory cortex might employ a more active process to infer features of the underlying speech sources, akin to neural filling-in (Cervantes Constantino and Simon, 2017; Leonard et al., 2016). The latency difference might then reflect an additional processing cost for the recovery of underlying features that are not directly accessible in the sensory input. Such an added processing cost in particular might explain why background speech is cognitively taxing (see *Processing of "ignored" acoustic sources* below).

While the two possibilities could both explain the results described so far, they make different predictions regarding responses to masked onsets. A passive process, which applies to overt and masked onsets equally, should be exhaustively explained by the model used above. On the other hand, an active process applied specifically to masked onsets might generate an additional response modulation associated with masked onsets. To test for such a modulation, we created a new predictor reflecting masked onsets only, regardless of which source they originated from. This predictor was implemented as an element-wise comparison of the onset spectrogram representations. Specifically, at each time-frequency point, the predictor uses the larger source onset value, but only if, and by the amount, it is larger than the corresponding onset in the mixture, i.e., *max(0, max(attended, ignored) – mixture)*. This additional predictor improves predictions of brain responses in the ROI bilaterally (t_{max} = 8.12, p < .001), supporting the hypothesis of an active process for recovering masked onsets.

Masked onsets are processed with a delay

Model comparison thus indicates that the neural processing of masked onsets is significantly different from that of overt onsets. This indicates that the influence of attention should also be assessed separately for overt and masked onsets. The previously used predictors do not allow this, however, because the speech sources were modeled as unified streams, combining overt and masked onsets. To separate effects of masking and attention, the information from the previously used onset predictors was recombined to generate a new set of predictors (Figure 4-A). Specifically, for each speech source, the new "overt onsets" predictor models time-frequency points in which an onset in the source is also seen as an onset in the mixture (element-wise *min(mixture, source)*), whereas the "masked onsets" predictor models the degree to which an onset in the source is attenuated (masked) in the mixture (*max(0, source – mixture)*). This model thus disentangles the effect of attention (attended vs ignored source) from whether an onset is overt in the mixture or masked. All four predictors significantly improved MEG response predictions (t_{max} ≤ 4.87, p < .001). In particular, this was also true for masked onsets in the ignored source (t_{max} = 4.87, p < .001), confirming that the auditory cortex recovers masked onsets even when they occur in the ignored source.

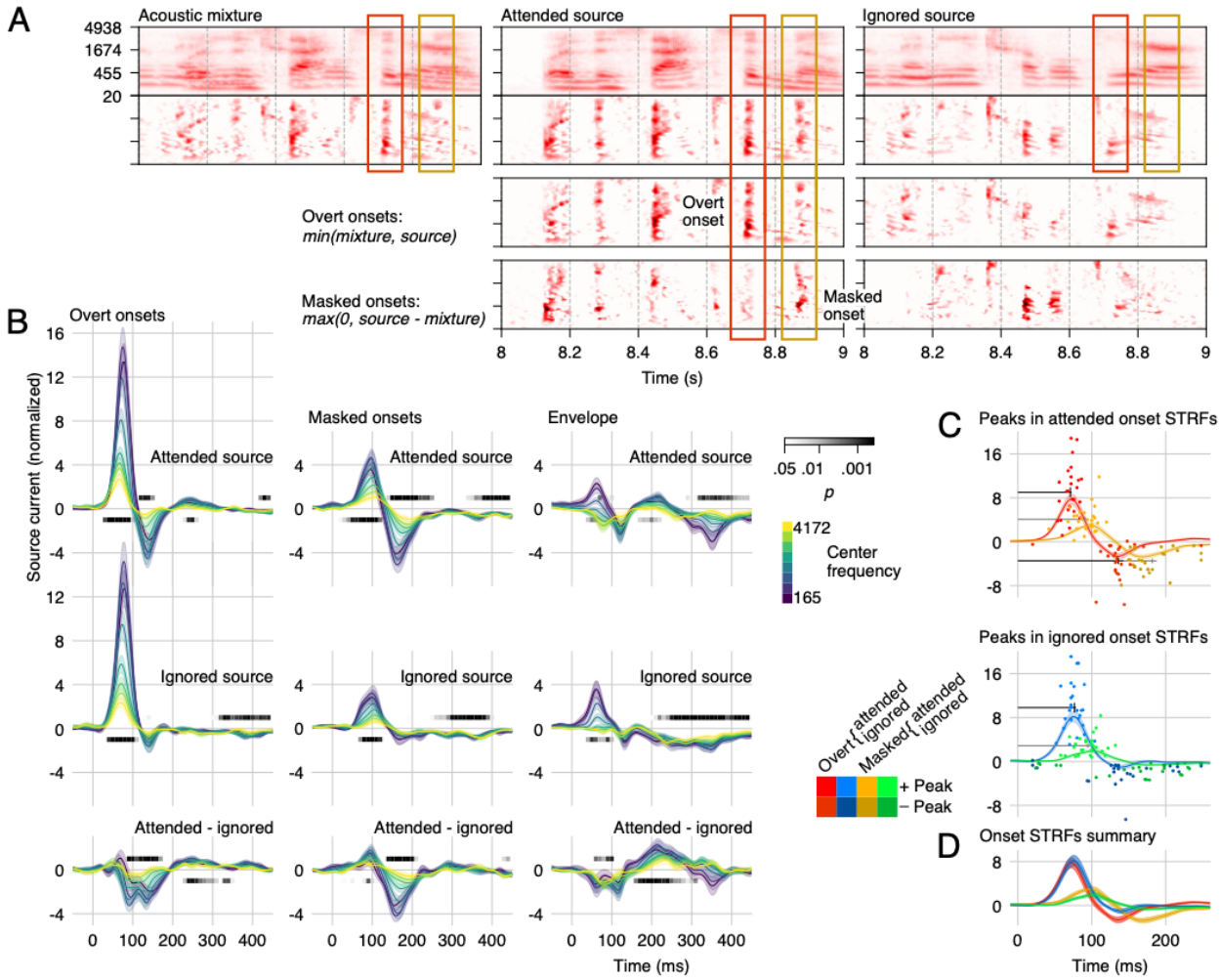

Figure 4. **Response functions to overt and masked onsets**. A) Spectrograms were transformed using element-wise operations to distinguish between overt onset, i.e., onsets in a source that are apparent in the mixture, and masked onsets, i.e., onsets in a source that are masked by the other source. Two examples are marked by rectangles: The red rectangle marks a region with an overt (attended) onset, i.e., an onset in the attended source that also corresponds to an onset in the mixture. The yellow rectangle marks a masked (attended) onset, i.e., an onset in the attended source which is not apparent in the mixture. B) The corresponding overt/masked STRFs exhibit the previously described positive-negative two peaked structure. For overt onsets, only the second, negative peak is modulated by attention. For masked onsets, even the first peak exhibits a small degree of attentional modulation. Envelope STRFs are largely unchanged by the new onset model. C) Responses to masked onsets are consistently delayed compared to responses to overt onsets. Details are analogous to Figure 3-D, except that the time window for finding peaks was extended to 20 – 250 ms to account for the longer latency of masked onset response functions. D) Direct comparison of the frequency-averaged onset STRFs highlights the amplitude differences between the peaks.

The STRFs to each stream's overt onsets exhibit an early positive peak at ~74 ms that does not differentiate between onsets originating from the attended and unattended source, followed by a negative peak at ~140 ms with enhanced amplitude for the attended source (Figure 4-B, left column). This suggests that the cortical processing stage corresponding to the first peak represents onsets in the acoustic mixture without regard to their acoustic source (Puvvada and Simon, 2017). By the time of the second peak, however, the cortical representations distinguish between the two sources, with onsets in the attended source being represented more prominently than onsets in the ignored source.

STRFs to masked onsets exhibit a similar positive-negative pattern as STRFs to overt onsets, but with important differences. Critically there is a consistent temporal delay (Figure 4-C). The delay is significant for both streams' positive peak (attended overt: 71 ms, masked: 91 ms, $t_{25} = 6.77$, $p < .001$; unattended overt: 77 ms, masked: 95 ms, $t_{25} = 7.23$, $p < .001$), as well as for the negative peak to attended onsets (overt: 136 ms, masked: 182 ms; $t_{25} = 4.72$, $p < .001$). For masked onsets in the ignored source, there is no evidence for a consistent negative peak at all, as can be seen in Figure 4-C where data points are spread throughout the time window. While these numbers might suggest an increase in the temporal delay from the first (~20 ms) to the second peak (46 ms), the difference for the second peak might be inflated by an additional positive peak for overt, attended onsets (see Figure 4-B).

Early effect of selective attention

Another difference for the response functions to masked onsets is that even the earlier, positive peak was significantly larger for attended compared to ignored onsets (Figure 4-B). Thus, the auditory cortex not only represents masked onsets, but these representations are substantively affected by whether the onset belongs to the attended or the ignored source. This distinction could have several causes. In the extreme, it could indicate that the two streams are completely segregated and represented as two distinct auditory objects. However, it might also be due to a weighting of features based on their likelihood of belonging to the attended source, for example, through modulation of excitability based on spectro-temporal prediction of the attended speech signal (Lakatos et al., 2013). Thus, onsets that are more likely to belong to the attended source might be represented more strongly, without yet being ascribed to one of the sources exclusively.

One discrepancy in previous studies using extra- and intracranial recordings is that the former were unable to detect an early effect of selective attention (Ding and Simon, 2012; Puvvada and Simon, 2017), whereas the latter showed a small but consistent enhancement of feature representations associated with the attended acoustic source signal (O'Sullivan et al., 2019). Furthermore, an early effect of attention would also be expected based on animal models that show task-dependent modulations of A1 responses (Atiani et al., 2009). Our results show that this discrepancy might be in part due to which acoustic features are analyzed: while overt acoustic onsets were not associated with an effect of selective attention, masked onsets were.

Overall, the early difference between the attended and ignored source suggests that acoustic information from the ignored source is represented to a lesser degree than information from the attended source. This is consistent with evidence from

psychophysics suggesting that auditory representations of background speech are not as fully elaborated as those of the attended foreground (Shinn-Cunningham et al., 2007).

Increasing abstraction over time

Responses to overt and masked onsets exhibit a comparable positive-negative two peak structure. While the first, positive peak is much larger for overt compared to masked onsets, the second, negative peak is of comparable magnitude (see Figure 4-D). This trend is confirmed by a peak (positive, negative) by masking (overt, masked) ANOVA of attended STRF peak amplitudes with a significant interaction ($F_{(1,25)}$=33.45, $p < .001$; in order to compare positive and negative peaks, peak amplitudes of the negative peak were multiplied by -1). One may infer, then, that at the earlier stage, the response is dominated by bottom-up processing of the acoustic stimulus, with a much smaller contribution reflecting the internally generated, recovered source properties. At the later stage, this distinction disappears, and the responses reflecting overt and masked onsets are of comparable magnitude. Similarly, the earliest stage of the mixture onset representations did not distinguish onsets in the attended source from onsets in the ignored source, but subsequent response peaks to overt and masked onsets showed increasing attention-based separation. Broadly, this pattern of results is consistent with a succession of processing stages, with early stages dominated by bottom-up activation from the input signal, gradually leading to later stages with task-driven, internally generated representations.

Attentive processing is not strictly time-locked

While the response magnitude to overt and masked onsets thus seems to be adjusted at subsequent processing stages, the response latency is not. Representations of masked onsets were consistently delayed compared to those of overt onsets by approximately 20 ms (see Figure 4-D). Previous research found that the latency of the representation of speech increased with increasing levels of stationary noise (Ding and Simon, 2013), suggesting a processing cost to recovering acoustic source information from noise. Our results suggest that this is not a uniform delay for a given perceptual stream, but that the delay varies by whether an acoustic element is overt or locally masked by the acoustic background. The delay might thus arise from a variable processing cost that depends on the local acoustic environment.

This latency difference between representations of overt and masked onsets entails that upstream speech processing mechanisms may receive different packages of information about the attended speech source with some temporal desynchronization. While this might imply a need for a higher order corrective mechanism, it is also possible that upstream mechanisms are tolerant to this small temporal distortion. A misalignment of 20 ms is small compared to the normal temporal variability encountered in speech (although phonetic contrasts do exist where a distortion of a few tens of milliseconds would be relevant). Indeed, in audio-visual speech perception, temporal misalignment up to 100 ms between auditory and visual input can be tolerated (van Wassenhove et al., 2007).

Processing of “ignored” acoustic sources

The interference in speech perception from a second talker can be very different from the interference caused by non-speech sounds. For instance, music is cortically segregated from speech even when both signals are unattended, consistent with a more automatic segregation, possibly due to distinctive differences in acoustic signal properties (Hausfeld et al., 2018). In contrast, at moderate signal to noise ratios (SNRs), a second talker causes much more interference with speech perception than a comparable non-speech masker. Interestingly, this interference manifests not just in the inability to hear attended words, but in intrusions of words from the ignored talker (Brungart, 2001). The latter fact in particular has been interpreted as evidence that ignored speech might be segregated and processed to a relatively high level. On the other hand, listeners seem to be unable to access words in more than one speech source at a time, even when the sources are spatially separated (Kidd et al., 2005). Furthermore, demonstrations of semantic processing of ignored speech are rare and usually associated with specific perceptual conditions such as dichotic presentation (Rivenez et al., 2006). Consistent with this, recent EEG/MEG evidence suggests that unattended speech is not processed in a time-locked fashion at the lexical (Brodbeck et al., 2018a) or semantic (Broderick et al., 2018) level. The results described here, showing systematic recovery of acoustic features from the ignored speech source, suggest a potential explanation for the increased interference from speech compared to other maskers. Representing onsets in two speech sources could be expected to increase cognitive load compared to detecting onsets of a single source in stationary noise. These representations of ignored speech might also act as bottom-up cues and cause the tendency for intrusions from the ignored talker. They might even explain why a salient and overlearned word, such as one’s own name (Wood and Cowan, 1995), might sometimes capture attention, which could happen based on acoustic rather than lexical analysis (Woods and McDermott, 2018). Finally, at very low SNRs the behavioral pattern can invert, and a background talker can be associated with better performance than stationary noise maskers (Brungart, 2001). In such conditions, there might be a benefit of being able to segregate the ignored speech source and use this information strategically (Fiedler et al., 2019).

Overall, these results are highly consistent with object-based models of auditory attention, in which perception depends on an interplay between bottom-up analysis and formation of local structure, and top-down selection and global grouping, or streaming (Shinn-Cunningham, 2008).

Conclusions

How do listeners succeed in selectively listening to one of two concurrent talkers? Our results suggest that active recovery of acoustic onsets plays a critical role. Early responses in the auditory cortex represent not only overt acoustic onsets, but also reconstruct acoustic onsets in the speech sources that are masked in the mixture, even if they originate from the ignored speech source. This suggests that early responses, in addition to representing a spectro-temporal decomposition of the mixture, actively reconstruct acoustic features that could originate from either speech source. Consequently, these early responses make comparatively complex acoustic features

from both speech sources available for downstream processes, thus enabling both selective attention, as well as bottom-up effects of salience and interference.

Acknowledgements

This work was supported by a National Institutes of Health grant R01-DC014085 (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 Shihab Shamma for several fruitful discussions, 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. designed experiment and secured funding. C.B. and J.Z.S. analyzed data and wrote the manuscript. A.J., C.B. and J.Z.S. performed simulations.

Methods

Participants

The data analyzed here have been previously used in an unrelated analysis (Brodbeck et al., 2018a). MEG responses were recorded from 28 native speakers of English, recruited by media advertisements from the Baltimore area. Participants with medical, psychiatric or neurological illnesses, head injury, and substance dependence or abuse were excluded. All subjects provided informed consent in accordance with the University of Maryland Baltimore Institutional Review Board 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, resulting in a final sample of 18 male and 8 female participants with mean age 45.2 (range 22 - 61).

Stimuli

Two chapters were selected from an audiobook recording 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). Four 1-minute long segments were extracted from each chapter (referred to as male-1 through 4 and female 1 through 4). Pauses longer than 300 ms were shortened to an interval randomly chosen between 250 and 300 ms, and loudness was matched perceptually. Two-talker 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 (mix-1 through 4 were constructed by mixing male-1 and female-1, through 4).

Procedure

During MEG data acquisition, participants lay supine and were instructed to keep their eyes closed 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.

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. After each segment, participants

answered a question relating to the content of the attended stimulus. Then, the four segments just heard were all presented once each, as single talkers. The same procedure was repeated for stimulus segments 3 and 4.

Data acquisition and preprocessing

Brain responses were recorded with a 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 (Gramfort et al., 2014). Flat channels were automatically detected and excluded. Extraneous artifacts were removed with temporal signal space separation (Taulu and Simola, 2006). Data were filtered between 1 and 40 Hz with a zero-phase FIR filter (mne-python 0.15 default settings). Extended infomax independent component analysis (Bell and Sejnowski, 1995) was then used to remove ocular and cardiac artifacts. Responses time-locked to the onset of the speech stimuli were extracted and downsampled to 100 Hz. For responses to the two-talker mixture, the first second of data, in which only the to-be attended talker was heard, was discarded.

Five marker coils attached to subjects' head served to localize the head position with respect to the MEG sensors. Two measurements, one at the beginning and one at the end of the recording were averaged. The FreeSurfer (Fischl, 2012) ''fsaverage'' template brain was coregistered to each subject's digitized head shape (Polhemus 3SPACE FASTRAK) using rotation, translation, and uniform scaling. A source space was generated using four-fold icosahedral subdivision of the white matter surface, with source dipoles oriented perpendicularly to the cortical surface. Minimum $\ell 2$ norm current estimates (Dale and Sereno, 1993; Hämäläinen and Ilmoniemi, 1994) were computed for all data. Initial analysis was performed on the whole brain as identified by the FreeSurfer "cortex" label. Subsequent analyses were restricted to sources in the STG and Heschl's gyrus as identified in the ''aparc'' parcellation (Desikan et al., 2006).

Predictor variables

Predictor variables were based on gammatone spectrograms sampled at 256 frequencies, ranging from 20 to 5000 Hz in ERB space (Heeris, 2018), resampled to 1 kHz and scaled with exponent 0.6 (Biesmans et al., 2017). At this point, different stimulus representations were computed. Spectrograms were then binned into 8 frequency bands equally spaced in ERB space (omitting frequencies below 100 Hz because the female speaker had little power below that frequency) and resampled to match the MEG data.

Acoustic onset representations were computed by applying an auditory edge detection model (Fishbach et al., 2001) independently to each frequency band of the spectrogram. The model was implemented with a delay layer with 10 delays ranging from τ_2 = 3 to 5 ms, a saturation scaling factor of $C = 30$, and a receptive field based on

the derivative of a Gaussian window with $SD = 2$ ms. Negative values in the resulting onset spectrogram were set to 0.

The linear dependence between different predictor variables (Figure 3-C) was estimated by treating each predictor time series in turn as the dependent measure and predicting it from the other predictors through a kernel with $T = [-500, \ldots, 500)$ (see next section). For example, segments [male-1, female-2, male-3. female-4] were combined, and each of the 8 bands in this predictor were predicted from [[female-1, mix-1], [male-2, mix-2], …] (including all 8 bands). The same parameters were used as for fitting neural models, except that no temporal basis function was used. The measure of interest was the proportion of the ($\ell 1$) variability of the dependent variable that could not be explained from a linear combination of the other variables.

Reverse correlation

Spectro-temporal response functions (STRFs) were computed independently for each virtual current source (see Brodbeck et al., 2018b). The neural response at time t, y_t was predicted from the sum of N predictor variables x_n convolved with a corresponding response function h_n of length T:

$$\hat{y}_t = \sum_{n}^{N} \sum_{\tau}^{T} h_{n,\tau} \cdot x_{i,t-\tau}$$

STRFs were generated from a basis of 50 ms wide Hamming windows and were estimated using am iterative coordinate descent algorithm (David et al., 2007) to minimize the $\ell 1$ error. Early stopping was based on 4-fold split of the data, freezing each h_n when it lead to an increase of error in the testing data (see Brodbeck et al., 2019 for further details).

Model tests

Each spectrogram comprising of 8 time series (frequency bins) was treated as an individual predictor. Speech in quiet was modeled using the (envelope) spectrogram and acoustic onsets:

$$MEG \sim o + e$$

Where o=onsets and e=envelope. Models were estimated with STRFs with $T = [0, \ldots, 500)$ ms. In order to test the predictive power of each predictor, three corresponding null models were generated by temporally misaligning the predictor with the response by cyclically shifting the predictor for each segment by 15, 30 and 45 seconds. Model quality was quantified as the Pearson correlation between actual and predicted response. For each predictor, the model quality of the full model was compared with the average model quality of the three corresponding null models using a mass-univariate related measures t-test with threshold-free cluster enhancement (Smith and Nichols, 2009) and a null distribution based on 10,000 permutations (Brodbeck et al., 2019 for further details).

Initially, responses to speech in noise was predicted from:

$$MEG \sim o_{mix} + o_{att} + o_{ign} + e_{mix} + e_{att} + e_{ign}$$

Where *mix*=mixture, *att*=attended, *ign*=ignored. Based on evaluation of this model, e_{mix} was dropped (Figure 3). Masked onsets (Figure 4) were analyzed with:

$$MEG \sim o_{att,overt} + o_{ign,overt} + o_{att,masked} + o_{ign,masked} + e_{att} + e_{ign}$$

STRF tests

To evaluate STRFs, the corresponding model (only correctly aligned predictors) was refit with $T = [-100, \ldots, 500)$ ms to include an estimate of baseline activity (due to occasional edge artifacts, STRFs are displayed between -50 to 450 ms).

Auditory STRFs were computed for each subject and hemisphere as a weighted sum of STRFs in the region of interest (ROI) encompassing the STG and Heschl's gyrus. Weights were computed separately for each subject and hemisphere. First, each source point was assigned a vector with direction orthogonal to the cortical surface, and length equal to the total TRF power for responses to clean speech (sum of squares over time, frequency and predictor). The ROI direction was then determined as the first principal component of these vectors, with the sign adjusted to be positive on the inferior-superior axis. A weight was then assigned to each source as the dot product of this direction with the source's direction, and weights were normalized across the ROI.

In order to make TRFs more comparable across subjects, they were smoothed on the frequency axis with a Hamming window of width 7. STRFs were statistically analyzed in the time range $[0, \ldots, 450)$ ms using mass-univariate *t*-tests and ANOVAs, with *p*-values calculated from null distributions based on the maximum statistic (*t, F*) in 10,000 permutations (Maris and Oostenveld, 2007).

References

Atiani, S., Elhilali, M., David, S.V., Fritz, J.B., and Shamma, S.A. (2009). Task Difficulty and Performance Induce Diverse Adaptive Patterns in Gain and Shape of Primary Auditory Cortical Receptive Fields. Neuron *61*, 467–480.

Bell, A.J., and Sejnowski, T.J. (1995). An Information-Maximization Approach to Blind Separation and Blind Deconvolution. Neural Comput. *7*, 1129–1159.

Biesmans, W., Das, N., Francart, T., and Bertrand, A. (2017). Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario. IEEE Trans. Neural Syst. Rehabil. Eng. *25*, 402–412.

Bregman, A.S., Ahad, P., Kim, J., and Melnerich, L. (1994a). Resetting the pitch-analysis system: 1. Effects of rise times of tones in noise backgrounds or of harmonics in a complex tone. Percept. Psychophys. *56*, 155–162.

Bregman, A.S., Ahad, P.A., and Kim, J. (1994b). Resetting the pitch□analysis system. 2. Role of sudden onsets and offsets in the perception of individual components in a cluster of overlapping tones. J. Acoust. Soc. Am. *96*, 2694–2703.

Brodbeck, C., Hong, L.E., and Simon, J.Z. (2018a). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. Curr. Biol. *28*, 3976-3983.e5.

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

Brodbeck, C., Brooks, T.L., Das, P., and Reddigari, S. (2019). Eelbrain 0.30 (Zenodo).

Broderick, M.P., Anderson, A.J., Liberto, G.M.D., 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.

Brungart, D.S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. J. Acoust. Soc. Am. *109*, 1101–1109.

Brungart, D.S., and Simpson, B.D. (2002). The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. J. Acoust. Soc. Am. *112*, 664–676.

Carlyon, R.P. (2004). How the brain separates sounds. Trends Cogn. Sci. *8*, 465–471.

Cervantes Constantino, F., and Simon, J.Z. (2017). Dynamic cortical representations of perceptual filling-in for missing acoustic rhythm. Sci. Rep. *7*, 17536.

Cherry, E.C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. J. Acoust. Soc. Am. *25*, 975–979.

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.

Daube, C., Ince, R.A.A., and Gross, J. (2019). Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. Curr. Biol. *29*, 1924-1937.e9.

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.

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.

Ding, N., and Simon, J.Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. Proc. Natl. Acad. Sci. U. S. A. *109*, 11854–11859.

Ding, N., and Simon, J.Z. (2013). Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical Representation of Speech. J. Neurosci. *33*, 5728–5735.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A.J., and Shamma, S.A. (2009). Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. Neuron *61*, 317–329.

Fiedler, L., Wöstmann, M., Herbst, S.K., and Obleser, J. (2019). Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. NeuroImage *186*, 33–42.

Fischl, B. (2012). FreeSurfer. NeuroImage *62*, 774–781.

Fishbach, A., Nelken, I., and Yeshurun, Y. (2001). Auditory Edge Detection: A Neural Model for Physiological and Psychoacoustical Responses to Amplitude Transients. J. Neurophysiol. *85*, 2303–2323.

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.

Hämäläinen, M.S., and Ilmoniemi, R.J. (1994). Interpreting magnetic fields of the brain: minimum norm estimates. Med. Biol. Eng. Comput. *32*, 35–42.

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.e4.

Hausfeld, L., Riecke, L., Valente, G., and Formisano, E. (2018). Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. NeuroImage *181*, 617–626.

Heeris, J. (2018). Gammatone Filterbank Toolkit.

Hukin, R.W., and Darwin, C.J. (1995). Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification. Percept. Psychophys. *57*, 191–196.

Hullett, P.W., Hamilton, L.S., Mesgarani, N., Schreiner, C.E., and Chang, E.F. (2016). Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. J. Neurosci. *36*, 2014–2026.

Kidd, G., Arbogast, T.L., Mason, C.R., and Gallun, F.J. (2005). The advantage of knowing where to listen. J Acoust Soc Am *118*, 12.

Kidd, G., Mason, C.R., Swaminathan, J., Roverud, E., Clayton, K.K., and Best, V. (2016). Determining the energetic and informational components of speech-on-speech masking. J. Acoust. Soc. Am. *140*, 132–144.

Lakatos, P., Musacchia, G., O'Connel, M.N., Falchier, A.Y., Javitt, D.C., and Schroeder, C.E. (2013). The Spectrotemporal Filter Mechanism of Auditory Selective Attention. Neuron *77*, 750–761.

Leonard, M.K., Baud, M.O., Sjerps, M.J., and Chang, E.F. (2016). Perceptual restoration of masked speech in human cortex. Nat. Commun. *7*, 13619.

Loftus, G.R., and Masson, M.E.J. (1994). Using confidence intervals in within-subject designs. Psychon. Bull. Rev. *1*, 476–490.

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods *164*, 177–190.

McDermott, J.H. (2009). The cocktail party problem. Curr. Biol. *19*, R1024–R1027.

Middlebrooks, J.C., Simon, J.Z., Popper, A.N., and Fay, R.R. (2017). The Auditory system at the cocktail party (Cham: Springer International Publishing).

Moerel, M., De Martino, F., and Formisano, E. (2012). Processing of Natural Sounds in Human Auditory Cortex: Tonotopy, Spectral Tuning, and Relation to Voice Sensitivity. J. Neurosci. *32*, 14205–14216.

O'Sullivan, J., Herrero, J., Smith, E., Schevon, C., McKhann, G.M., Sheth, S.A., Mehta, A.D., and Mesgarani, N. (2019). Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. Neuron S0896627319307809.

Puvvada, K.C., and Simon, J.Z. (2017). Cortical Representations of Speech in a Multitalker Auditory Scene. J. Neurosci. *37*, 9189–9196.

Rivenez, M., Darwin, C.J., and Guillaume, A. (2006). Processing unattended speech. J. Acoust. Soc. Am. *119*, 4027–4040.

Shinn-Cunningham, B.G. (2008). Object-based auditory and visual attention. Trends Cogn. Sci. *12*, 182–186.

Shinn-Cunningham, B.G., Lee, A.K.C., and Oxenham, A.J. (2007). A sound element gets lost in perceptual competition. Proc. Natl. Acad. Sci. *104*, 12223–12227.

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.

Stilp, C.E., and Kluender, K.R. (2010). Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. Proc. Natl. Acad. Sci. U. S. A. *107*, 12387–12392.

Taulu, S., and Simola, J. (2006). Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. Phys. Med. Biol. *51*, 1759.

van Wassenhove, V., Grant, K.W., and Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. Neuropsychologia *45*, 598–607.

Wood, N., and Cowan, N. (1995). The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel? J. Exp. Psychol. Learn. Mem. Cogn. *21*, 255–260.

Woods, K.J.P., and McDermott, J.H. (2018). Schema learning for the cocktail party problem. Proc. Natl. Acad. Sci. *115*, E3313–E3322.

Zhou, Y., and Wang, X. (2010). Cortical Processing of Dynamic Sound Envelope Transitions. J. Neurosci. *30*, 16741–16754.

Zion Golumbic, E.M., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D., Simon, J.Z., et al. (2013). Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party." Neuron *77*, 980–991.