Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers

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

Short title: Neural speech restoration at the cocktail party

Abstract

Humans are remarkably skilled at listening to one speaker out of an acoustic mixture of several speech sources. Two speakers are easily segregated, 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 magnetoencephalography (MEG) 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 non-overlapping spectro-temporal features are seen for both talkers. When competing talkers' spectro-temporal features mask each other, the individual representations persist, but they occur with a ~20 ms delay. This suggests that the auditory cortex recovers acoustic features that are masked in the mixture, even if they occurred in the ignored speech. The existence of such noise-robust cortical representations, of features present in attended as well as ignored speech, suggests an active cortical stream segregation process, which could explain a range of behavioral effects of ignored background speech.

Keywords

Auditory streaming, stream segregation, continuous speech, temporal response function

Introduction

When listening to an acoustic scene, the signal that arrives at the ears is an additive mixture of the different sound sources. Listeners trying to selectively attend to one of the sources face the task of determining which spectro-temporal features belong to that source [1]. When multiple speech sources are involved, as in the classic cocktail party problem [2], 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 [3,4]. Binaural cues can facilitate segregation of different sound sources based on their location [5], 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 [6]. Here we are specifically interested in the fundamental ability to segregate and attend to one out of multiple speakers even without such external cues.

The neural mechanisms involved in this ability are not well understood, but previous research suggests at least two separable cortical processing stages. In magnetoencephalographic (MEG) responses to multiple talkers [7], the early (~50 ms) cortical 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 [8]. In general, cortical regions further away from core auditory cortex tend to mainly reflect information about the attended speaker [9]. Together, these results suggest a cortical mechanism which, based on a detailed representation of the acoustic input, detects and groups features belonging to the attended source.

A long-standing question is whether early cortical processing of the acoustic mixture is restricted to passive spectro-temporal filtering, or whether it involves active grouping of acoustic features leading to the formation of auditory object representations. The filter theory of attention suggests early representations of physical stimulus characteristics independent of attention, with attention selecting a subset of these for further processing and semantic identification [10,11]. Consistent with this, electroencephalography (EEG) and MEG results suggest that time-locked processing of higher order linguistic features, such as words and meaning, is restricted to the attended speech source [12,13]. However, it is not known whether, in the course of recovering 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 [8]. 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 more active mechanisms to dynamically recover potential speech features, regardless of what stream they belong to.

Selective attention could then rely on these local auditory (proto-) objects to recover the attended speech [14]. This hypothesis critically 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.

Cortical responses to speech reflect a strong representation of the envelope (or spectrogram) of the speech signal [15,16]. Prior work has also shown that acoustic onsets are prominently represented in auditory cortex, both in naturalistic speech [17,18] and in non-speech stimuli [19,20]. Previous studies using paradigms similar to the one used here often predicted brain responses from only envelopes or only onsets [16,21–24], but more recent studies show that both representations explain non-redundant portion of the responses [12,18]. Behaviorally, acoustic onsets are also specifically important for speech intelligibility [25,26]. Here we consider both envelope and onset features, but focus on onset features in particular because of their relevance for stream segregation, as follows [1]. If acoustic elements in different frequency regions are co-modulated over time, they likely stem from the same physical source [27]. A simultaneous onset in distinct frequency bands thus provides sensory evidence that these cross-frequency features originate from the same acoustic source, and should be processed as an auditory object. 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 [28,29]. For example, the onset of a vowel is characterized by a shared onset at the fundamental frequency of the voice and its harmonics. Correspondingly, 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 [30]. This link to object perception thus makes acoustic onsets particularly relevant cues, which might be represented distinctly from envelope cues and used to detect the beginning of local auditory objects, and thus to aid segregation of the acoustic input into different, potentially overlapping auditory objects.

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 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: a single talker in quiet (“clean speech”), and a two-talker mixture, in which a female and a male speaker were mixed at equal perceptual loudness. MEG responses were analyzed as additive, linear response to multiple concurrent stimulus features (see Figure 1). First, cross-validated model comparisons were used to determine which representations significantly improved prediction of the responses. Then, the resultant spectro-temporal response functions (STRFs) were analyzed to gain insight into the nature of the representations.

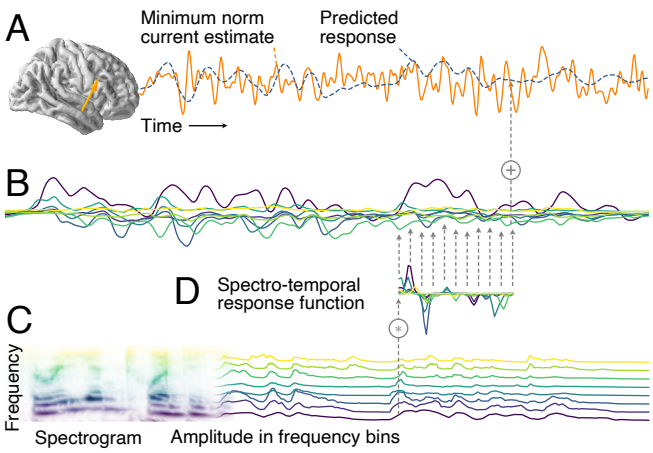

Figure 1. **Additive linear response model based on spectro-temporal response functions (STRFs).** A) MEG responses recorded during stimulus presentation were source localized with distributed minimum norm current estimates. A single virtual source dipole is shown for illustration, with its physiologically 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 a collection of 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 [19]. The two predictors were each binned into 8 frequency bands, such that the MEG responses were predicted from a model of the acoustic stimulus encompassing 16 time series in total. Each of the two predictors was assessed based on how well (left-out) MEG responses were predicted by the full model, compared to a null model in which the relevant predictor was omitted. Both predictors significantly improve predictions (onsets: $t_{max} = 12.00$, $p \leq .001$; envelopes: $t_{max} = 9.39$, $p \leq .001$), with an anatomical distribution consistent with sources in HG and STG bilaterally (Figure 2-B). Since this localization agrees with findings from intracranial recordings [8,17], results were henceforth analyzed in an auditory region of interest (ROI) restricted to these two anatomical landmarks (Figure 2-C). When averaging the model fits in this ROI, almost all subjects showed responses associated with both predictors (Figure 2-D).

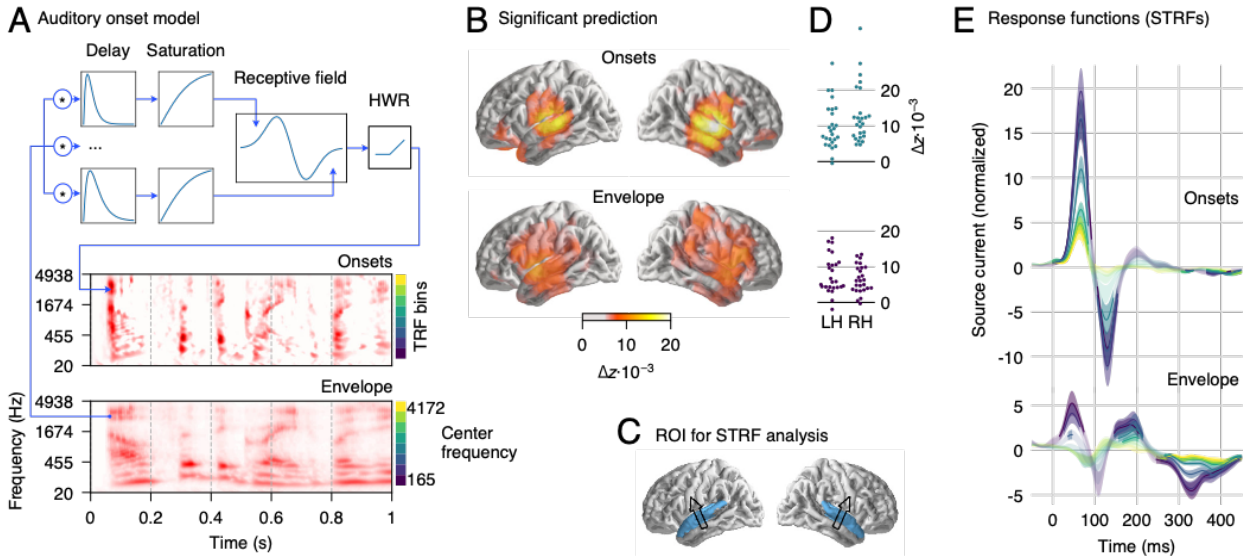

Figure 2. **MEG responses to clean speech**. A) Schematic illustration of the neurally inspired acoustic edge detector model, which was used to generate onset representations. The signal at each frequency band was passed through multiple parallel pathways, with increasing delays, so that an "edge detector" receptive field could detect the occurrence of fast increases. Final half-wave rectification (HWR) removed the negative sections. An excerpt from a gammatone spectrogram ("envelope") and the corresponding onset representation is shown for illustration purposes. B) Regions of significant explanatory power of onset and envelope representations, consistent with a main source in bilateral auditory cortex ($p \leq .05$, 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) Individual subject model fit improvements, averaged over the ROI in the left and right hemisphere (LH/RH). E) Spectro-temporal response functions (STRFs) corresponding to onset and envelope representations in the ROI; the onset STRF exhibits a clear pair of positive and negative peaks, while peaks in the envelope STRF are less well-defined. 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) [31]. Regions in which STRFs differ significantly from 0 are marked with more saturated (less faded) colors ($p \leq .05$, corrected for time/frequency).

Auditory cortical STRFs were summarized for each subject 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 (Figure 2-C, arrows). This current vector is consistent with activity in core auditory cortex and the superior temporal plane. Since MEG sensors are less sensitive to any radial currents, expected from lateral STG areas implicated by intracranial recordings [8], we focus here on the temporal information in STRFs rather than the spatial distribution of sources. STRFs can be interpreted as reflecting different processing stages associated with different latencies, possibly involving multiple regions in the superior temporal lobe. STRFs were initially separately

analyzed by hemisphere, but since none of the reported results interact 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-E). An early, positive peak (average latency 65 ms) is followed by a later, negative peak (126 ms). This structure closely resembles previously described auditory response functions to envelope representations when estimated without consideration of onsets [16]. 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-E), which is consistent with previous tonotopic mapping of speech areas and may follow from the spectral distribution of information in the speech signal [32,33]. 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 S1).

Auditory cortex represents ignored speech

MEG responses to a two-speaker mixture were then used to test for a neural representation of ignored speech. Participants listened to a perceptually-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 stimuli and subjects. Responses were predicted using both onset and envelope representations for: the acoustic mixture, the attended speech source, and the ignored source (Figure 3-A). The underlying rationale is that, since the brain does not have direct access to the individual speech sources, if there is neural activity corresponding to either source separately (above and beyond the mixture), this indicates that cortical responses have reconstructed features of that source. Both predictors representing the ignored speech significantly improve predictions of the responses in the ROI (both $p < .001$, onsets: $t_{max} = 6.70$, envelopes: $t_{max} = 6.28$; group level statistics were evaluated with spatial permutation tests; subject-specific model fits, averaged in the ROI are shown in Figure 3-B). This indicates that acoustic features of the ignored speech are represented neurally even after controlling for features of the mixture and the attended source. The remaining four predictors also significantly increased model fits (all $p < .001$; mixture onsets: $t_{max} = 8.61$, envelopes: $t_{max} = 5.70$; attended onsets: $t_{max} = 6.32$, envelopes: $t_{max} = 7.37$).

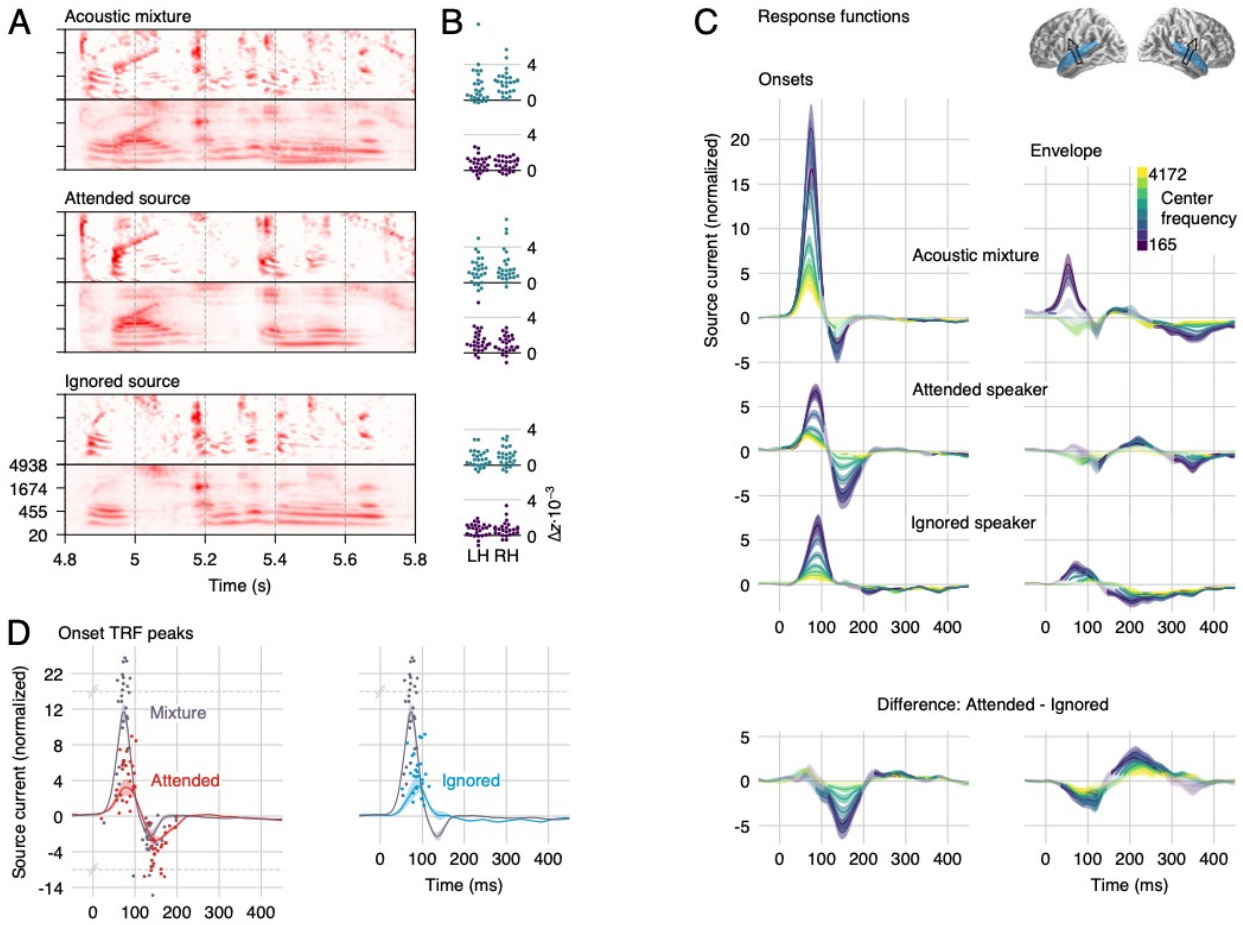

Figure 3. **Responses to the two-speaker mixture, using the stream-based model**. A) The envelope and onset representations of the acoustic mixture and the two speech sources were used to predict MEG responses. B) Individual subject model fit improvement due to each predictor, averaged in the auditory cortex ROI. Each predictor explains neural data not accounted for by the others. C) Auditory cortex STRFs to onsets are characterized by the same positive/negative peak structure as STRFs to a single speaker. The early, positive peak is dominated by the mixture, but also contains speaker-specific information. The second, negative peak is dominated by representations of the attended speaker, followed, to a lesser extent, by those of the mixture. As with responses to a single talker, the envelope STRFs have lower amplitudes, but they do show a strong and well-defined effect of attention. Explicit differences between the attended and ignored representations are shown in the bottom row. Details as in Figure 2. D) The major onset STRF peaks representing individual speech sources are delayed compared to corresponding peaks representing the mixture. To determine latencies, mixture-based and individual-speaker-based STRFs were averaged across frequency (lines with shading for mean ±1 SE). Dots represent the largest positive and negative peak for each subject between 20 and 200 ms. Note that the y-axis is scaled by an extra factor of 4 beyond the indicated break points at y = 14 and -6.

Onset STRFs exhibit the same characteristic positive-negative pattern as for responses to a single talker, but with reliable distinctions between the mixture and the individual speech streams (Figure 3-C, D). 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: 89 ms, t_{25} = 6.92, p < .001; amplitude mixture > attended: t_{25} = 8.41, p < .001; mixture > ignored: t_{25} = 7.66, p < .001). This positive peak is followed by a negative peak only in responses to the mixture (136 ms) and the attended source (150 ms; difference t_{25} = 3.20, p = .004). The amplitude of these negative peaks is statistically indistinguishable (t_{25} = 1.56, p = .132).

The mixture predictor is not completely orthogonal to the source predictors. This 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 Figures S1 and S2).

Envelope processing is strongly modulated by selective attention

While the envelope STRFs seem to be generally less structured than those of the onsets, a comparison of the STRFs to the attended and the ignored source revealed a strong and well-defined effect of attention (Figure 3-C, right column). The attended-ignored difference wave exhibits a negative peak at ~100 ms, consistent with previous work [16], 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 very earliest responses. 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 [28], which would allow early attentional processing of the spectrogram, or envelope features.

Auditory cortex “un-masks” masked onsets

The above 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 is particularly interesting because, while envelopes combine mostly in an additive manner, acoustic onsets may experience *negative* interference. This can be seen in the spectrograms in Figure 3-A: the envelope mixture representation largely looks like a sum of the two stream envelope representations. In contrast, the onset mixture representation has several features that have a *lower* amplitude than the corresponding feature in the relevant source. Neural representations of the source onsets thus suggest that the auditory cortex reconstructs features that are masked in the mixture. This possibility might be related to instances of neural filling-in, where cortical representations show evidence of filling in missing information to repair degraded input signals [34–36]. The latency difference between mixture and source onsets might then reflect additional processing for the recovery of underlying features that are not directly accessible in the sensory input. However, these neural representations of stream-specific onsets could also be the result of an acoustic representation that the mixture is simply transparent to (unlike our gammatone representation). An example 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.

While the two possibilities could both explain the results described so far, they make different predictions regarding responses to masked onsets. A passive mechanism, based on features to

which the mixture is transparent, should apply to overt and masked onsets equally (because they appear masked only in the gammatone representations), and should thus be exhaustively explained by the model used above. On the other hand, an active mechanism that specifically processes masked onsets might generate an additional response modulation associated with masked onsets. To test for such a modulation, we subdivided the stream-based onset predictors to allow for separate responses to overt and masked onsets. The new predictors were implemented as element-wise operations on the onset spectrograms (Figure 4-A). Specifically, for each speech source, the new "masked onsets" predictor models the degree to which an onset in the source is attenuated (masked) in the mixture, i.e. the amount by which a segregated speech source exceeds the (physically presented) resultant acoustic mixture, or zero when it does not: (*max(source – mixture, 0)*). The new "overt onsets" predictor models all other time-frequency points, where an onset in the source is also seen as a comparable onset in the mixture (element-wise *min(mixture, source)*). Note that with this definition the sum of overt and masked onsets exactly equals the original speech source onset representation, i.e., the new predictors model the same onsets, but allow response to differ depending on whether an onset is masked or not. Replacing the two speech-source based onset predictors with the four overt/masked onset predictors significantly improves the model fit (t_{max} = 6.81, p < .001), suggesting that cortical responses indeed distinguish between overt and masked onsets. Each of the four new predictors individually improved MEG response predictions, although masked onsets did so more robustly (attended: t_{max} = 8.42, p < .001; ignored: t_{max} = 5.23, p < .001) than overt onsets (attended: t_{max} = 3.34, p = .027; ignored: t_{max} = 3.82, p = .016; Figure 4-B); this difference could be due to overt source onsets being more similar to the mixture onsets predictor. Critically, the significant effect for masked onsets in the ignored source confirms that the auditory cortex recovers masked onsets even when they occur in the ignored source.

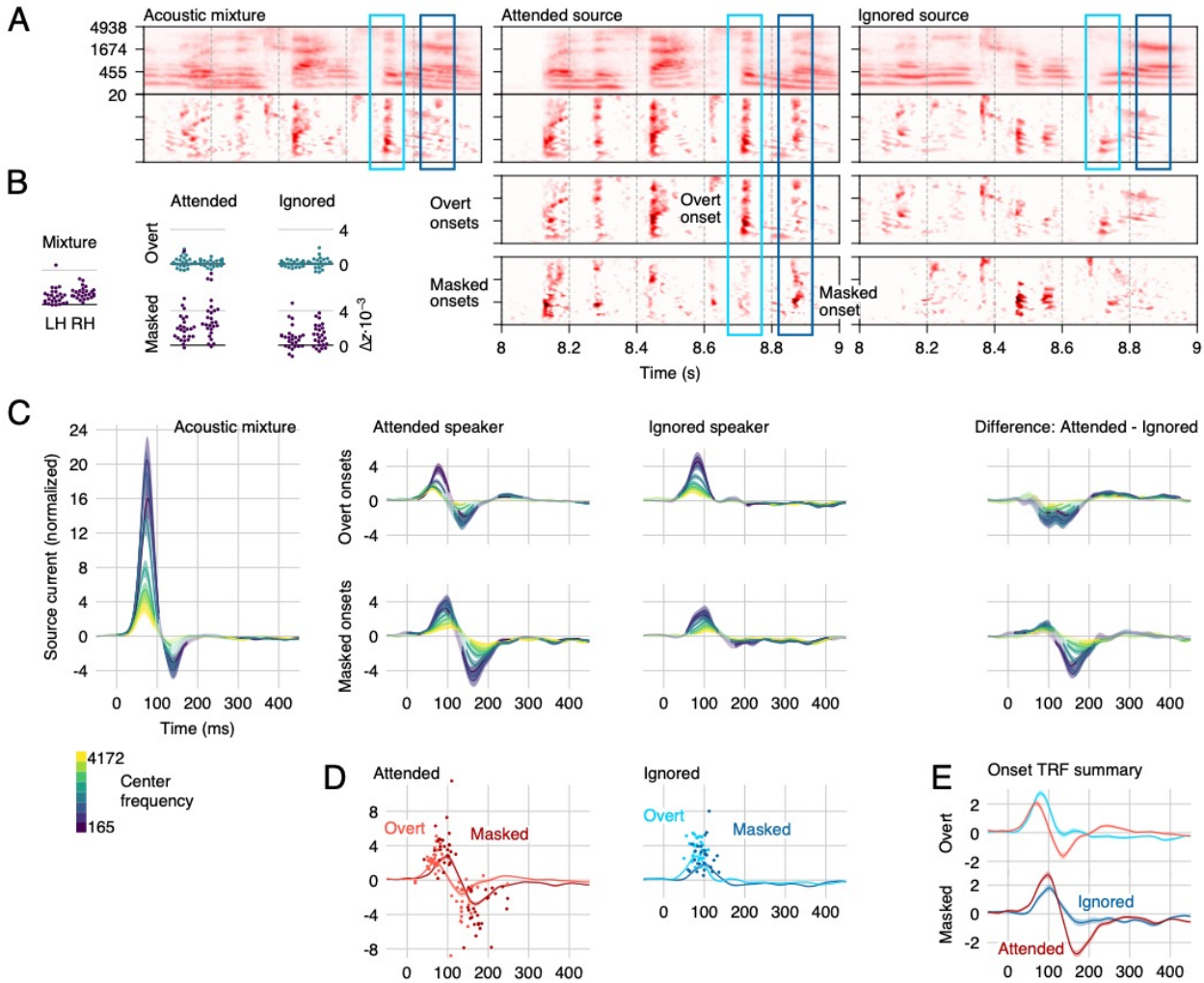

Figure 4. **Responses to overt and masked onsets**. A) Spectrograms (note that in this figure, the onset representations are placed below the envelope representations, to aid visual comparison of the different onset representations) were transformed using element-wise operations to distinguish between overt onsets, i.e., onsets in a source that are apparent in the mixture, and masked onsets, i.e., onsets in a source that are masked in the presence of the other source. Two examples are marked by rectangles: The light blue 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 dark blue rectangle marks a masked (attended) onset, i.e., an onset in the attended source which is not apparent in the mixture. B) All predictors significantly improve the cross-validated model fit (note that improvements were statistically tested with a test sensitive to spatial variation, whereas these plots show single-subject ROI average fits). C) The corresponding overt/masked STRFs exhibit the previously described positive-negative two peaked structure. The first, positive peak is dominated by a representation of the mixture, but also contains segregated features of the two talkers. 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. D) 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. E) Direct comparison of the frequency-averaged onset TRFs highlights the amplitude differences between the peaks. For overt onsets, the negative deflection due to selective attention starts decreasing the response magnitude even near the maximum of the first, positive peak. For masked onsets, the early peak reflecting attended onsets is increased despite the subsequent enhanced negative peak. Results for envelope predictors are omitted from this figure because they are practically indistinguishable from those in Figure 3.

Masked onsets are processed with a delay and an early effect of attention

Model comparison thus indicates that the neural representation of masked onsets is significantly different from that of overt onsets. The analysis of STRFs suggests that this is for at least two reasons (Figure 4-C-E): response latency differences, and a difference in the effect of selective attention.

First, responses to masked onsets are systematically delayed compared to overt onsets (Figure 4-D). This is true for the early, positive peak (mixture: 72 ms; attended/overt: 72 ms, masked: 91 ms, $t_{25} = 2.85$, $p = .009$; ignored/overt: 83 ms, masked: 97 ms, $t_{25} = 6.11$, $p < .001$), as well as the later, negative peak (mixture: 138 ms; attended/overt: 133 ms, masked: 182 ms, $t_{25} = 4.45$, $p < .001$). Thus, at each major peak, representations of masked onsets lag behind representations of overt onsets by at least 15 ms.

Second, for overt onsets, the early representations at the positive peak appear to be independent of the target of selective attention (Figure 4-C, top right). In contrast, for masked onsets, even these early representations are enhanced by attention (Figure 4-C, bottom right). This difference is confirmed in a stream (attended, ignored) by masking (overt, masked) ANOVA on peak amplitudes with a significant interaction ($F_{(1,25)} = 24.45$, $p < .001$). For overt onsets, Figure 4-E might suggest that the early peak is actually enhanced for the ignored speaker; however, this difference can be explained by the early onset of the second, negative response to attended onsets, which overlaps with the earlier peak. This observation makes the early effect of attention for masked onsets all the more impressive, because the early peak is larger *despite* the onset of the subsequent, negative peak (note the steeper slope between positive and negative peak for attended masked onsets). Also note that we consider the effect of attention early relative to the peak structure of the TRFs; in terms of absolute latency, the onset of the effect of attention is actually more similar between masked and overt onsets (cf. Figure 4-E).

Delayed response to masked onsets

Previous research has found that the latency of responses to speech increases with increasing levels of stationary noise [37,38], suggesting a cost to recovering acoustic source information from noise. Our results indicate that this is not a uniform delay for a given perceptual stream, but that the delay varies for each acoustic element based on whether this element is overt or locally masked by the acoustic background. That the delay applies to the individual onsets, rather than the stream as a whole, implies that at this level, the stream is not processed as a homogeneous entity, but that the acoustic elements related to the onsets constitute separate units of analysis.

The addition of background noise is associated with increased response latencies to auditory stimuli at least as early as the brainstem response wave-V [39]. This observed shift in latency is in the sub-millisecond range and may have a mechanistic explanation in terms of different populations of auditory nerve fibers: Background noise saturates high spontaneous rate fibers, and the response is now dominated by somewhat slower, low spontaneous rate fibers [40]. In cortical responses to simple stimuli, like tones, much larger delays are observed in the presence of static noise, in the order of tens of milliseconds [41]. Latency shifts due to absolute signal intensity [42] might be additive with shifts due to noise [43]. Such a non-linear increase in response latency with intensity might be a confounding factor in our analysis, which is based on linear methods: Compared to overt onsets, masked onsets in the ignored talker should generally correspond to weaker onsets in the mixture. Splitting overt and masked onsets might thus improve the model fit because it allows modeling different response latencies for different intensity levels of onsets in the mixture, rather than reflecting a true response to the background speech. In order to control for this possibility, we compared the model fit of the background-aware model with a model allowing for 3 different intensity levels of onsets in the mixture (and without an explicit representation of onsets in the ignored speaker). The background speaker-aware model out-performed the level-aware model (t_{max} = 9.21, p < .001), suggesting that the present results are not explained by this level-based non-linear response to the mixture. Furthermore, a background-unaware non-linearity would not explain the difference in the effect of attention between overt and masked onsets. Together, this suggests that the observed delay is related to recovering acoustic source information, rather than a level-based non-linearity in the response to the mixture. Furthermore, a study of single units in primary auditory cortex found that neurons with delayed, noise-robust responses exhibited response properties suggestive of network effects [44]. This is consistent with the effect of attention found here on the early peak of masked, but not of overt onsets, i.e., the delayed responses to masked onsets also exhibit more evidence of goal-driven processing than the corresponding responses to overt onsets.

In sum, masking causes latency increases at different stages of the auditory system. These latency shifts increase at successive stages of the ascending auditory pathway, as does the preponderance of noise-robust response properties [45]. It is likely that different levels of the auditory system employ different strategies to recover auditory signals of interest, and do so at different time scales. Together with these considerations, our results suggest that the auditory cortex actively recovers masked speech features, and not only of the attended, but also of the ignored speech source.

Early effect of selective attention

Besides the shift in latency, response functions to overt and masked onsets differed in a more fundamental way: While the early, positive response peak to overt onsets did not differentiate between attended and ignored onsets, the early peak to masked onsets contained significantly larger representations of attended onsets (cf. Figure 4-C). Thus, not only do early auditory cortical responses represent 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. This could be achieved, for example, through modulation of excitability based on spectro-temporal prediction of the attended speech signal [46]. 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 any early effect of selective attention [7,16], whereas the latter showed a small but consistent enhancement of feature representations associated with the attended acoustic source signal [8]. Furthermore, an early effect of attention would also be expected based on animal models that show task-dependent modulations of A1 responses [47,48]. Our results show that this discrepancy may depend, in part, on 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 [49]. More generally, it is consistent with results that suggest an influence of attention early on in auditory stream formation [50].

Stages of speech segregation through selective attention

Regardless of whether listening to a single talker or two concurrent talkers, response functions to acoustic onsets are characterized by a prominent positive-negative two peak structure. Brain responses to two concurrent talkers allow separating these response functions into components related to different representations, and thus reveal different processing stages. Figure 5 presents a plausible model incorporating these new findings. Early on, the response is dominated by a representation of the acoustic mixture, with a preliminary segregation, possibly implemented through spectral filters [8]. This is followed by restoration of speech features that are masked in the mixture, regardless of speaker, but with a small effect of selective attention suggesting a more active mechanism. Finally, later responses are dominated by selective attention, suggesting a clearly separated representation of the attended speaker as would be expected in successful streaming.

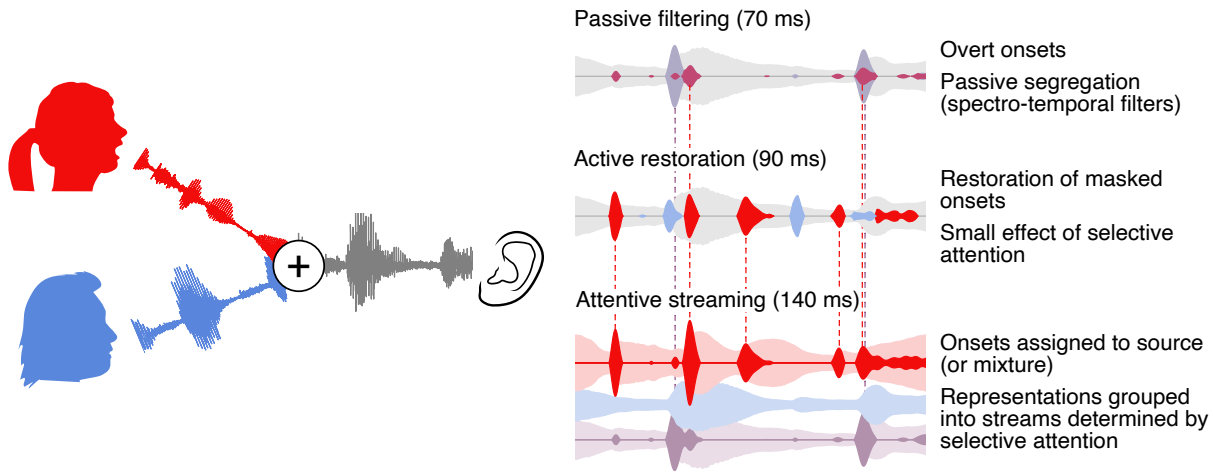

Figure 5. **Model of onset-based stream segregation**. A model of cortical processing stages compatible with the results reported here. *Left*: The auditory scene, with additive mixture of the waveforms from the attended and the ignored speakers (red and blue, respectively). *Right*: Illustration of cortical representations at different processing stages. *Passive filtering*: At an early stage, onsets are extracted from the acoustic mixture and representations are partially segregated, possibly based on frequency. This stage corresponds to the early positive peak in onset TRFs. *Active Restoration*: A subsequent stage also includes representations of onsets in the underlying speech sources that are masked in the mixture, corresponding to the first peak in TRFs to masked onsets. At this stage, a small effect of attention suggests a preliminary selection of onsets with a larger likelihood of belonging to the attended speaker. *Streaming*: Finally, at a third stage, the response to onsets from the ignored speaker is suppressed, suggesting that now the two sources are clearly segregated (see also [8]). This stage corresponds to the second, negative peak, which is present in TRFs to mixture and attended onsets but not to ignored onsets.

An open question concerns how the overt and masked feature representations are temporally integrated. Representations of masked onsets were consistently delayed compared to those of overt onsets by approximately 20 ms (see Figure 4-D). This latency difference 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 [51].

Broadly, the new results are consistent with previous findings that early cortical responses are dominated by the acoustic mixture, rather than receiving pre-segregated representations of the individual streams [7,8]. However, the new results do show evidence of an earlier, partial segregation, in the form of representations of acoustic onsets which are segregated from the mixture, though not grouped into separate streams. Since these early representations do not

strictly distinguish between the attended and the ignored speaker they likely play the role of an intermediate step in extracting the information needed to selectively attend to one of the two speakers. 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 [14,52].

Implications for 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 [22]. 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 [53]. 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 process words in more than one speech source at a time, even when the sources are spatially separated [54]. Furthermore, demonstrations of lexical processing of ignored speech are rare and usually associated with specific perceptual conditions such as dichotic presentation [55]. Consistent with this, recent EEG/MEG evidence suggests that unattended speech is not processed in a time-locked fashion at the lexical [12] or semantic [13] 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 [56], might sometimes capture attention, which could happen based on acoustic rather than lexical analysis [57]. Finally, at very low SNRs the behavioral pattern can invert, and a background talker can be associated with better performance than stationary noise maskers [53]. In such conditions, there might be a benefit of being able to segregate the ignored speech source and use this information strategically [21].

An open question is how the auditory system deals with the presence of multiple background speakers. When there are multiple background speakers, does the auditory system attempt to unmask the different speakers all separately, or are they represented as a unified background [7]? An attempt to isolate speech features even from multiple background talkers might contribute to the overly detrimental effect of babble noise with a small number of talkers [58].

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 and 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 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.

Materials and Methods

Participants

The data analyzed here have been previously used in an unrelated analysis [12] and can be retrieved from the Digital Repository at the University of Maryland (http://hdl.handle.net/1903/21109). 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 participants provided written 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, respectively). 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 (such that either speaker was deemed equally easy to attend to). 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, approximately 70 dB SPL.

Participants listened four times to mix-1, while attending to one speaker and ignoring the other (which speaker they attended to was counterbalanced across participants), 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 better than 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 [59]. Flat channels were automatically detected and excluded. Extraneous artifacts were removed with temporal signal space separation [60]. 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 [61] was then used to remove ocular and cardiac artifacts. Responses time-locked to the onset of the speech stimuli were extracted and resampled 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 participants' 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 [62] ''fsaverage'' template brain was coregistered to each participant'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 [63,64] 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 [65].

Predictor variables

Predictor variables were based on gammatone spectrograms sampled at 256 frequencies, ranging from 20 to 5000 Hz in ERB space [66], resampled to 1 kHz and scaled with exponent 0.6 [67].

Acoustic onset representations were computed by applying an auditory edge detection model [19] 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.

Onset representations of the two speakers were split into masked and overt onsets using element-wise operations on the onset spectrograms. Masked onsets were defined by the extent to which onsets were larger in the source than in the mixture:

$$o_{masked} = \max\left(o_{source} - o_{mixture}, 0\right)$$

Overt onsets were onsets that were not masked, i.e., speech source onsets that were also visible in the mixture:

$$o_{overt} = o_{source} - o_{masked} \equiv min(\, o_{source}, o_{mixture})$$

Using this procedure, approximately 67% of total onset magnitudes ($\ell 1$ norm) was assigned to overt onsets, and 33% to masked onsets.

For model estimation, envelope and onset 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. As part of the reverse correlation procedure, each predictor time series (i.e., each frequency bin) was scaled by its $\ell 1$ norm over time.

For testing an intensity-based nonlinear response (see *Delayed response to masked onsets*), the onset predictor was split into three separate predictors, one for each of three intensity levels. For each of the 8 frequency bins, individual onsets were identified as contiguous non-zero elements; each onset was assigned an intensity based on the sum of its elements, and the onsets were then assigned to one of three predictors based on intensity tertiles (calculated separately for each band). This resulted in three 8-band onset spectrograms modeling low, medium and high intensity onsets.

Reverse correlation

Spectro-temporal response functions (STRFs) were computed independently for each virtual current source [see 68]. 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 an iterative coordinate descent algorithm [69] to minimize the $\ell 1$ error.

For model evaluation, left-out data were predicted using 4-fold cross-validation. Folds were created by assigning successive trials to the different folds in order (1, 2, 3, 4, 1, 2, …). In an outer loop, the responses in each fold were predicted with STRFs estimated from the remaining 3 folds. These predictions, combined, served to calculate the correlation between measured and predicted responses used for model tests. In an inner loop, each of the 3 estimation folds was in turn used as validation set for STRFs trained on the two remaining folds. STRFs were iteratively improved based on the maximum error reduction in the training set (the steepest coordinate descent) and validated in the validation set. Whenever a predictor time series (i.e., one spectrogram bin) would have caused an increasing in the error in the validation set, the kernel for this predictor was frozen, continuing until all predictors were frozen (see [70] for further

details). The 3 STRFs from the inner loop were averaged to predict responses in the left-out testing data.

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 $T = [0, \dots, 500)$ ms. Model quality was quantified as the Fisher *z*-scored Pearson correlation between actual and predicted response. Maps of *z*-values were smoothed with a Gaussian kernel (SD = 5 mm). In order to test the predictive power of each predictor, a corresponding null model was generated by removing that predictor. For each predictor, the model quality of the full model was compared with the model quality of the corresponding null model using a mass-univariate related measures *t*-test with threshold-free cluster enhancement [71] and a null distribution based on 10,000 permutations [72 for further details]. In addition to *p*-values we report t_{max} for each comparison, the largest *t*-value in the significant area, for effect size comparison. For single talker speech (Figure 2) this test included the whole cortex (as labeled by FreeSurfer). For subsequent tests of the two talker condition the same tests were used, but the test area was restricted to the auditory ROI comprising the STG and transverse temporal gyrus in each hemisphere.

Initially, responses to speech in noise (Figure 3) was predicted from:

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

Where *mix* = mixture, *att* = attended, *ign* = ignored. Masked onsets (Figure 4) were analyzed with:

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

In order to test for a level-dependent nonlinear response to onsets in the mixture, this model was compared with the following:

$$MEG \sim o_{mix-low} + o_{mix-mid} + o_{mix-high} + o_{att,overt} + o_{att,masked} + e_{mix} + e_{att} + e_{ign}$$

Where *mix-low, -mid* and *-high* = mixture low, mid and high intensity. This model has the same number of predictors, but assumes no awareness of onsets in the ignored speaker.

STRF analysis

To evaluate STRFs, the corresponding model was refit with $T = [-100, \dots, 500)$ ms to include an estimate of baseline activity (due to occasional edge artifacts, STRFs are displayed between -50 to 450 ms). Using the same 4-fold split of the data as for model fits, 4 STRF estimates were averaged, each using one fold of the data for validation and the remaining 3 for training. Since predictors and responses were $\ell 1$ normalized for the reverse correlation, and STRFs were analyzed in this normalized space, STRFs provide an SNR-like measure of response strength at different latencies for each subject.

Auditory STRFs were computed for each subject and hemisphere as a weighted sum of STRFs in the auditory ROI encompassing the STG and transverse temporal (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 these weights were normalized within the ROI.

In order to make STRFs more comparable across subjects, they were smoothed on the frequency axis with a Hamming window of width 7 bins. STRFs were statistically analyzed in the time range $[0, \dots, 450)$ 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 [73].

For visualization and peak analysis, STRFs were upsampled to 500 Hz. Peak latencies were computed by first averaging auditory STRFs along the frequency axis, and then finding the largest or smallest value in each subject's TRF in a window of $[20, 200)$ ms for single speaker and stream-based analysis (Figures 2 and 3) or $[20, 250)$ ms for the masked onset analysis (Figure 4). Reported peak latencies are always average latencies across subject.

References

1. Bregman AS. Auditory scene analysis: the perceptual organization of sound. Cambridge, Mass: MIT Press; 1990.

2. Cherry EC. Some Experiments on the Recognition of Speech, with One and with Two Ears. J Acoust Soc Am. 1953;25: 975–979. doi:10.1121/1.1907229

3. McDermott JH. The cocktail party problem. Curr Biol. 2009;19: R1024–R1027. doi:10.1016/j.cub.2009.09.005

4. Middlebrooks JC, Simon JZ, Popper AN, Fay RR, editors. The Auditory system at the cocktail party. Cham: Springer International Publishing; 2017.

5. Brungart DS, Simpson BD. The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal. J Acoust Soc Am. 2002;112: 664–676. doi:10.1121/1.1490592

6. Kidd G, Mason CR, Swaminathan J, Roverud E, Clayton KK, Best V. Determining the energetic and informational components of speech-on-speech masking. J Acoust Soc Am. 2016;140: 132–144. doi:10.1121/1.4954748

7. Puvvada KC, Simon JZ. Cortical Representations of Speech in a Multitalker Auditory Scene. J Neurosci. 2017;37: 9189–9196. doi:10.1523/JNEUROSCI.0938-17.2017

8. O'Sullivan J, Herrero J, Smith E, Schevon C, McKhann GM, Sheth SA, et al. Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. Neuron. 2019; S0896627319307809. doi:10.1016/j.neuron.2019.09.007

9. Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, et al. Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a “Cocktail Party.” Neuron. 2013;77: 980–991. doi:10.1016/j.neuron.2012.12.037

10. Broadbent DE. Perception and communication. London: Pergamon Press; 1958.

11. Lachter J, Forster KI, Ruthruff E. Forty-five years after Broadbent (1958): Still no identification without attention. Psychol Rev. 2004;111: 880–913. doi:10.1037/0033-295X.111.4.880

12. Brodbeck C, Hong LE, Simon JZ. Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. Curr Biol. 2018;28: 3976-3983.e5. doi:10.1016/j.cub.2018.10.042

13. Broderick MP, Anderson AJ, Liberto GMD, Crosse MJ, Lalor EC. Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. Curr Biol. 2018;28: 803-809.e3. doi:10.1016/j.cub.2018.01.080

14. Shinn-Cunningham BG. Object-based auditory and visual attention. Trends Cogn Sci. 2008;12: 182–186. doi:10.1016/j.tics.2008.02.003

15. Lalor EC, Foxe JJ. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. Eur J Neurosci. 2010;31: 189–193. doi:10.1111/j.1460-9568.2009.07055.x

16. Ding N, Simon JZ. Emergence of neural encoding of auditory objects while listening to competing speakers. Proc Natl Acad Sci U S A. 2012;109: 11854–9. doi:10.1073/pnas.1205381109

17. Hamilton LS, Edwards E, Chang EF. A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. Curr Biol. 2018;28: 1860-1871.e4. doi:10.1016/j.cub.2018.04.033

18. Daube C, Ince RAA, Gross J. Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. Curr Biol. 2019;29: 1924-1937.e9. doi:10.1016/j.cub.2019.04.067

19. Fishbach A, Nelken I, Yeshurun Y. Auditory Edge Detection: A Neural Model for Physiological and Psychoacoustical Responses to Amplitude Transients. J Neurophysiol. 2001;85: 2303–2323. doi:10.1152/jn.2001.85.6.2303

20. Zhou Y, Wang X. Cortical Processing of Dynamic Sound Envelope Transitions. J Neurosci. 2010;30: 16741–16754. doi:10.1523/JNEUROSCI.2016-10.2010

21. Fiedler L, Wöstmann M, Herbst SK, Obleser J. Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions. NeuroImage. 2019;186: 33–42. doi:10.1016/j.neuroimage.2018.10.057

22. Hausfeld L, Riecke L, Valente G, Formisano E. Cortical tracking of multiple streams outside the focus of attention in naturalistic auditory scenes. NeuroImage. 2018;181: 617–626. doi:10.1016/j.neuroimage.2018.07.052

23. Petersen EB, Wöstmann M, Obleser J, Lunner T. Neural tracking of attended versus ignored speech is differentially affected by hearing loss. J Neurophysiol. 2017;117: 18–27. doi:10.1152/jn.00527.2016

24. Fiedler L, Wöstmann M, Graversen C, Brandmeyer A, Lunner T, Obleser J. Single-channel in-ear-EEG detects the focus of auditory attention to concurrent tone streams and mixed speech. J Neural Eng. 2017;14: 036020. doi:10.1088/1741-2552/aa66dd

25. Stilp CE, Kluender KR. Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. Proc Natl Acad Sci U S A. 2010;107: 12387–12392. doi:10.1073/pnas.0913625107

26. Koning R, Wouters J. The potential of onset enhancement for increased speech intelligibility in auditory prostheses. J Acoust Soc Am. 2012;132: 2569–2581. doi:10.1121/1.4748965

27. Elhilali M, Ma L, Micheyl C, Oxenham AJ, Shamma SA. Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. Neuron. 2009;61: 317–329. doi:10.1016/j.neuron.2008.12.005

28. Bregman AS, Ahad P, Kim J, Melnerich L. Resetting the pitch-analysis system: 1. Effects of rise times of tones in noise backgrounds or of harmonics in a complex tone. Percept Psychophys. 1994;56: 155–162. doi:10.3758/BF03213894

29. Bregman AS, Ahad PA, Kim J. 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. 1994;96: 2694–2703. doi:10.1121/1.411277

30. Hukin RW, Darwin CJ. Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification. Percept Psychophys. 1995;57: 191–196. doi:10.3758/BF03206505

31. Loftus GR, Masson MEJ. Using confidence intervals in within-subject designs. Psychon Bull Rev. 1994;1: 476–490. doi:10.3758/BF03210951

32. Moerel M, De Martino F, Formisano E. Processing of Natural Sounds in Human Auditory Cortex: Tonotopy, Spectral Tuning, and Relation to Voice Sensitivity. J Neurosci. 2012;32: 14205–14216. doi:10.1523/JNEUROSCI.1388-12.2012

33. Hullett PW, Hamilton LS, Mesgarani N, Schreiner CE, Chang EF. Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. J Neurosci. 2016;36: 2014–2026. doi:10.1523/JNEUROSCI.1779-15.2016

34. Cervantes Constantino F, Simon JZ. Dynamic cortical representations of perceptual filling-in for missing acoustic rhythm. Sci Rep. 2017;7: 17536. doi:10.1038/s41598-017-17063-0

35. Leonard MK, Baud MO, Sjerps MJ, Chang EF. Perceptual restoration of masked speech in human cortex. Nat Commun. 2016;7: 13619. doi:10.1038/ncomms13619

36. Cervantes Constantino F, Simon JZ. Restoration and Efficiency of the Neural Processing of Continuous Speech Are Promoted by Prior Knowledge. Front Syst Neurosci. 2018;12: 56. doi:10.3389/fnsys.2018.00056

37. Ding N, Simon JZ. Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical Representation of Speech. J Neurosci. 2013;33: 5728–5735. doi:10.1523/JNEUROSCI.5297-12.2013

38. Billings CJ, McMillan GP, Penman TM, Gille SM. Predicting Perception in Noise Using Cortical Auditory Evoked Potentials. J Assoc Res Otolaryngol. 2013;14: 891–903. doi:10.1007/s10162-013-0415-y

39. Burkard RF, Sims D. A Comparison of the Effects of Broadband Masking Noise on the Auditory Brainstem Response in Young and Older Adults. Am J Audiol. 2002;11: 13–22. doi:10.1044/1059-0889(2002/004)

40. Mehraei G, Hickox AE, Bharadwaj HM, Goldberg H, Verhulst S, Liberman MC, et al. Auditory Brainstem Response Latency in Noise as a Marker of Cochlear Synaptopathy. J Neurosci. 2016;36: 3755–3764. doi:10.1523/JNEUROSCI.4460-15.2016

41. Billings CJ, Tremblay KL, Stecker GC, Tolin WM. Human evoked cortical activity to signal-to-noise ratio and absolute signal level. Hear Res. 2009;254: 15–24. doi:10.1016/j.heares.2009.04.002

42. Drennan DP, Lalor EC. Cortical Tracking of Complex Sound Envelopes: Modeling the Changes in Response with Intensity. eneuro. 2019;6: ENEURO.0082-19.2019. doi:10.1523/ENEURO.0082-19.2019

43. Teschner MJ, Seybold BA, Malone BJ, Hüning J, Schreiner CE. Effects of Signal-to-Noise Ratio on Auditory Cortical Frequency Processing. J Neurosci. 2016;36: 2743–2756. doi:10.1523/JNEUROSCI.2079-15.2016

44. Malone BJ, Heiser MA, Beitel RE, Schreiner CE. Background noise exerts diverse effects on the cortical encoding of foreground sounds. J Neurophysiol. 2017;118: 1034–1054. doi:10.1152/jn.00152.2017

45. Rabinowitz NC, Willmore BDB, King AJ, Schnupp JWH. Constructing Noise-Invariant Representations of Sound in the Auditory Pathway. Zador AM, editor. PLoS Biol. 2013;11: e1001710. doi:10.1371/journal.pbio.1001710

46. Lakatos P, Musacchia G, O'Connel MN, Falchier AY, Javitt DC, Schroeder CE. The Spectrotemporal Filter Mechanism of Auditory Selective Attention. Neuron. 2013;77: 750–761. doi:10.1016/j.neuron.2012.11.034

47. Fritz J, Shamma S, Elhilali M, Klein D. Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat Neurosci. 2003;6: 1216–1223. doi:10.1038/nn1141

48. Atiani S, Elhilali M, David SV, Fritz JB, Shamma SA. Task Difficulty and Performance Induce Diverse Adaptive Patterns in Gain and Shape of Primary Auditory Cortical Receptive Fields. Neuron. 2009;61: 467–480. doi:10.1016/j.neuron.2008.12.027

49. Shinn-Cunningham BG, Lee AKC, Oxenham AJ. A sound element gets lost in perceptual competition. Proc Natl Acad Sci. 2007;104: 12223–12227. doi:10.1073/pnas.0704641104

50. Carlyon RP. How the brain separates sounds. Trends Cogn Sci. 2004;8: 465–471. doi:10.1016/j.tics.2004.08.008

51. van Wassenhove V, Grant KW, Poeppel D. Temporal window of integration in auditory-visual speech perception. Neuropsychologia. 2007;45: 598–607. doi:10.1016/j.neuropsychologia.2006.01.001

52. Elhilali M, Xiang J, Shamma SA, Simon JZ. Interaction between Attention and Bottom-Up Saliency Mediates the Representation of Foreground and Background in an Auditory Scene. Griffiths TD, editor. PLoS Biol. 2009;7: e1000129. doi:10.1371/journal.pbio.1000129

53. Brungart DS. Informational and energetic masking effects in the perception of two simultaneous talkers. J Acoust Soc Am. 2001;109: 1101–1109. doi:10.1121/1.1345696

54. Kidd G, Arbogast TL, Mason CR, Gallun FJ. The advantage of knowing where to listen. J Acoust Soc Am. 2005;118: 12.

55. Rivenez M, Darwin CJ, Guillaume A. Processing unattended speech. J Acoust Soc Am. 2006;119: 4027–4040. doi:10.1121/1.2190162

56. Wood N, Cowan N. 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. 1995;21: 255–260. doi:10.1037/0278-7393.21.1.255

57. Woods KJP, McDermott JH. Schema learning for the cocktail party problem. Proc Natl Acad Sci. 2018;115: E3313–E3322. doi:10.1073/pnas.1801614115

58. Simpson SA, Cooke M. Consonant identification in N-talker babble is a nonmonotonic function of N. J Acoust Soc Am. 2005;118: 2775–2778. doi:10.1121/1.2062650

59. Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, et al. MNE software for processing MEG and EEG data. NeuroImage. 2014;86: 446–460. doi:10.1016/j.neuroimage.2013.10.027

60. Taulu S, Simola J. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. Phys Med Biol. 2006;51: 1759. doi:10.1088/0031-9155/51/7/008

61. Bell AJ, Sejnowski TJ. An Information-Maximization Approach to Blind Separation and Blind Deconvolution. Neural Comput. 1995;7: 1129–1159. doi:10.1162/neco.1995.7.6.1129

62. Fischl B. FreeSurfer. NeuroImage. 2012;62: 774–781. doi:10.1016/j.neuroimage.2012.01.021

63. Hämäläinen MS, Ilmoniemi RJ. Interpreting magnetic fields of the brain: minimum norm estimates. Med Biol Eng Comput. 1994;32: 35–42. doi:10.1007/BF02512476

64. Dale AM, Sereno MI. Improved Localizadon of Cortical Activity by Combining EEG and MEG with MRI Cortical Surface Reconstruction: A Linear Approach. J Cogn Neurosci. 1993;5: 162–176. doi:10.1162/jocn.1993.5.2.162

65. Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage. 2006;31: 968–980. doi:10.1016/j.neuroimage.2006.01.021

66. Heeris J. Gammatone Filterbank Toolkit. 2018. Available: https://github.com/detly/gammatone

67. Biesmans W, Das N, Francart T, Bertrand A. Auditory-Inspired Speech Envelope Extraction Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario. IEEE Trans Neural Syst Rehabil Eng. 2017;25: 402–412. doi:10.1109/TNSRE.2016.2571900

68. Brodbeck C, Presacco A, Simon JZ. Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. NeuroImage. 2018;172: 162–174. doi:10.1016/j.neuroimage.2018.01.042

69. David SV, Mesgarani N, Shamma SA. Estimating sparse spectro-temporal receptive fields with natural stimuli. Netw Comput Neural Syst. 2007;18: 191–212. doi:10.1080/09548980701609235

70. Brodbeck C, Das P, Brooks TL, Reddigari S. Eelbrain 0.31. Zenodo; 2019. doi:10.5281/ZENODO.3564850

71. Smith SM, Nichols TE. Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage. 2009;44: 83–98. doi:10.1016/j.neuroimage.2008.03.061

72. Brodbeck C, Brooks TL, Das P, Reddigari S. Eelbrain 0.30. Zenodo; 2019. doi:10.5281/zenodo.2653785

73. Maris E, Oostenveld R. Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods. 2007;164: 177–190. doi:10.1016/j.jneumeth.2007.03.024