Auditory Cortex Tracks Masked Acoustic Onsets in Background Speech: Evidence for Early 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 multiple speech sources, even in the absence of binaural cues. Previous research on the neural representations underlying this ability suggests that the auditory cortex primarily represents only the unsegregated acoustic mixture in its early responses, and then selectively processes features of the attended speech at longer latencies (from ~85 ms). The mechanism by which the attended source signal is segregated from the mixture, however, and to what degree an ignored source may also be segregated and separately processed, is not understood. We show here, in human magnetoencephalographic responses to a two-talker mixture, an early neural representation of acoustic onsets in the ignored speech source, over and above onsets of the mixture and the attended source. This suggests that the auditory cortex initially reconstructs acoustic onsets belonging to any speech source, critically, even when those onsets are acoustically masked by another source. Overt onsets in the unseparated acoustic mixture were processed with a lower latency (~70 ms) than masked onsets in either source (~90 ms), suggesting a neural processing cost to the recovery of the masked onsets. Because acoustic onsets precede sustained source-specific information in the acoustic spectrogram, these representations of onsets are cues available for subsequent processing, including full stream segregation. Furthermore, these findings suggest that even bottom-up saliency of objects in the auditory background may rely on active cortical processing, explaining several behavioral effects of background speech.

Significance Statement

The ability to comprehend speech in the presence of multiple talkers is required frequently in daily life, and yet it is compromised in a variety of populations, for example in healthy aging. Here we address a longstanding question concerning the neural mechanisms supporting this ability: to what extent does the auditory cortex process and represent an interfering speech signal despite the fact that it is not being attended? We find that auditory cortex not only represents acoustic onsets in an ignored speech source, it does so even when those onsets are masked by the attended talker. This suggests that auditory cortex reconstructs and processes acoustic features of ignored speech, even in its effort to selectively process the attended speech.

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.

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 this is a nontrivial problem because the spectrograms of the different sources often have strong overlap (see Figure 3-A). Nevertheless, human listeners are remarkably skilled at focusing on one out of multiple talkers. Binaural cues can support segregation of different sound sources based on their location (1), 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 (2).

The mechanisms involved in this ability are not well understood. Previous research suggests that the auditory cortex dominantly represents features of the acoustic mixture in Heschl's gyrus (HG) starting before 50 ms, and more selectively processes features belonging to the attended signal in the superior temporal gyrus (STG) starting around 85 ms latency (3–5). Furthermore, time-locked processing of higher order linguistic features seems to be restricted to the attended speech source (6, 7). It is not known whether, in the course of recovering features of the attended source, the auditory cortex also segregates features of the ignored source from the mixture. A conservative hypothesis is that primary auditory cortex represents acoustic features of the mixture invariantly, and attentional mechanisms select only those representations that are relevant for the attended stream. Alternatively, the auditory cortex could employ some means to recover and represent potential speech features, even if obscured in the mixture, regardless of what stream they belong to, and attentional mechanism could then selectively process those features associated with the attended speech. An extreme possibility, discussed in the scene analysis literature, is that different sound sources could be fully segregated and individually represented, with attention merely selecting one of multiple readily available auditory stream representations (8).

Here we aim to distinguish between these hypotheses by analyzing auditory cortical representations of two concurrent speech sources. An important cue for segregating an acoustic source from a mixture is temporal coherence of different acoustic features (9). We focus in particular on acoustic onset features, i.e., acoustic edges corresponding to a frequency-specific increase in acoustic energy. A simultaneous onset of acoustic elements in distinct frequency bands is a strong cue that these different elements originate from the same speech source. Accordingly, shared acoustic onsets promote perceptual grouping of acoustic components into a single auditory object, such as a complex tone and, vice versa, separate onsets lead to perceptual segregation (10, 11). 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 is often perceived as a separate tone rather than as a component of the vowel (12). Acoustic onsets are very prominently represented in auditory cortex, both in naturalistic speech (13, 14) and in non-speech stimuli (15), and are important for speech intelligibility (16).

We used human magnetoencephalographic (MEG) responses to a continuous two-talker mixture to determine whether the auditory cortex reliably tracks acoustic onset or envelope features of

the ignored speech. 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, 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.

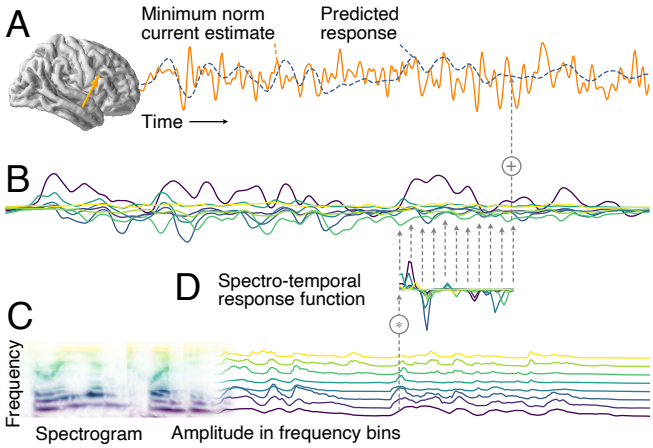

Figure 1. **Additive linear response model based on spectro-temporal response functions (STRFs)**. A) MEG responses recorded during stimulus presentation were source localized with distributed minimum norm current estimates. A single virtual source dipole is shown for illustration, with its physiologically 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 speech presented in quiet were predicted from the gammatone spectrogram of the stimulus, as well as a spectrogram of acoustic onsets (Figure 2-A). Acoustic onsets were derived from a neural model of auditory edge detection (17). 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 improved 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 (13), results were henceforth analyzed in a region of interest (ROI) restricted to these two anatomical landmarks (Figure 2-C).

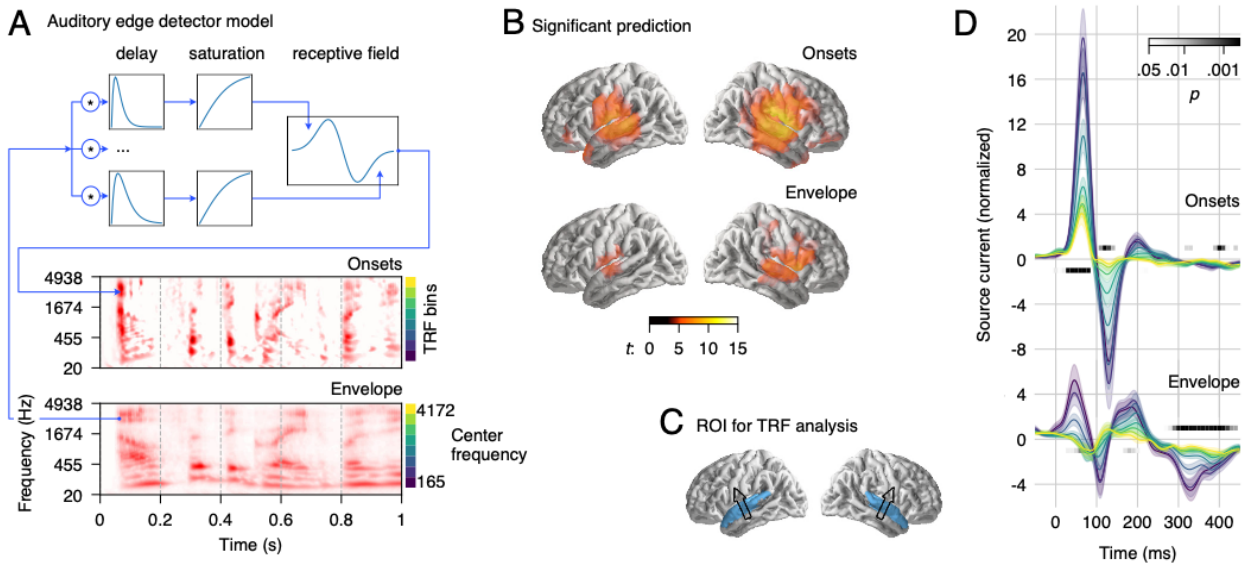

Figure 2. **Acoustic onset responses to clean speech**. A) Schematic illustration of the acoustic edge detector model, along with an excerpt from a gammatone spectrogram ("envelope") and the corresponding onset representation. B) Regions of significant explanatory power of onset- and envelope representations, 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 current direction of 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. 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 (18). Regions in which STRFs differ significantly from 0 (in any band) are marked with horizontal gray bars.

Auditory cortical STRFs were generated separately for each participant and hemisphere using a spatial filter based on principal component analyses of overall STRF power in the ROI. The average direction of that spatial filter replicated the direction of the well-known auditory MEG response with mainly vertical orientation (Figure 2-C). STRFs were initially 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 exhibited a well-defined two-peaked shape, consistent across frequency bands (Figure 2-D). They closely resembled previously described auditory response functions to envelope representations, when these were used without consideration of onsets (3). In comparison, envelope STRFs in the present results were diminished and exhibited 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 had disproportionately higher amplitude at lower frequencies (Figure 2-D). This is consistent with tonotopic mapping of speech areas and may follow from the spectral distribution of information in the speech signal (19, 20). An explanation based on signal properties is also supported by our simulations, in which equal TRFs for each band were simulated, and yet higher frequency bands resulted in lower amplitude responses (see 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 trials 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, including the two predictors representing the ignored speech significantly improved 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 in addition to features of the mixture and the attended source. Separate tests suggested that this result can be ascribed specifically to onset representations (t_{max} = 4.89, p < .001), whereas envelope representations of the ignored source did not significantly improve the model fit (t_{max} = -2.59, p = 1).

Taken individually, onsets in each of the three streams significantly improved predictions (t_{max} ≥ 4.89, p < .001), but none of the envelope representations did (all t_{max} ≤ -0.40, p = 1). This lack of predictive power for the envelope predictors, when tested individually, is likely due to 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 do indicate a reliable representation of ignored speech over and above representations of the acoustic mixture and the attended source.

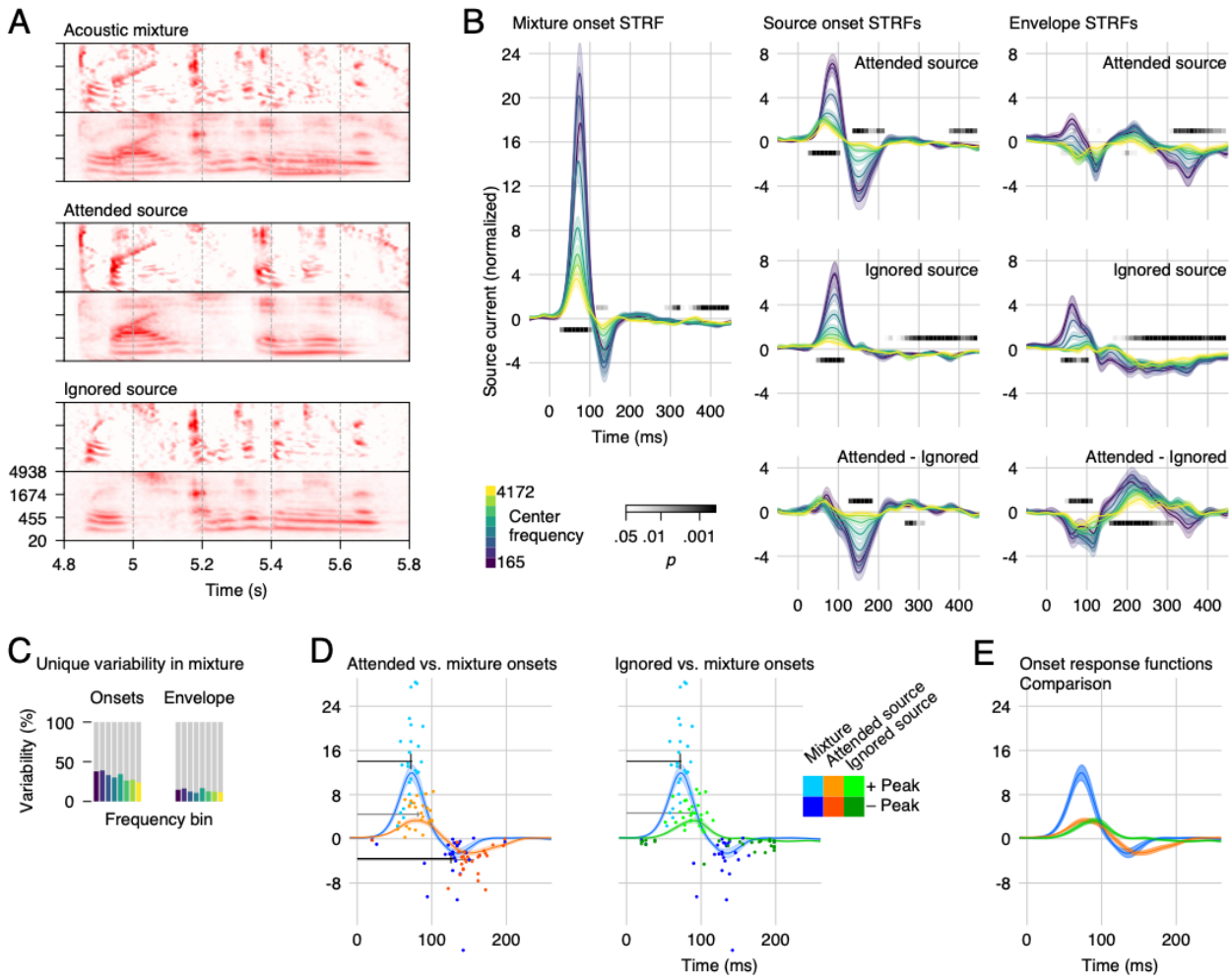

Figure 3. **Response functions to the two-speaker mixture, using the stream-based model**. A) The envelope and onsets of the acoustic mixture and the two speech sources were used to predict MEG responses. B) Auditory cortex STRFs to onsets in the mixture exhibit a large positive peak (72 ms) followed 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. 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 could not be explained from the individual speech sources (with a -500 – 500 ms temporal integration window). D) The major peaks to onsets in the speech sources are delayed compared to corresponding peaks to the mixture. To determine latencies, mixture-based and individual-speaker-based STRFs were averaged across frequency (lines with shading for 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. D) Direct comparison of onset response functions averaged across frequency, ±1 SE.

Onset STRFs exhibited 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 occurred earlier and had 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 was followed by a negative peak only in responses to the mixture (126 ms) and the attended source (150 ms; difference $t_{25} = 4.36$, $p < .001$). In contrast to the corresponding positive peak, the amplitude of these negative peaks was statistically indistinguishable ($t_{25} = 0.36$, $p = .722$). STRFs to the ignored source did not exhibit a detectable corresponding negative peak, as seen in Figure 3-C where participants' peaks cluster around the time window edges instead of at a characteristic latency.

The fact that the mixture predictor is not 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).

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. A comparison of 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 (3), 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, suggesting that when onset responses are accounted for separately from envelope responses, even early envelope processing is influenced by attention.

Auditory cortex recovers masked onsets

The results using these stream-based predictors suggest that the auditory cortex represents acoustic onsets in both speech sources separately, in addition to onsets in the acoustic mixture. This suggests a marked degree of abstraction from the acoustic input, involving early reconstruction of features of the inferred, underlying speech sources. This is further supported by the latency analysis, which suggests that representations of reconstructed source onsets are processed separately from onsets heard in the mixture. This latency difference might also be indicative of some additional processing cost, as reflected in the delay of the representation of reconstructed onsets. Such an added processing cost, however, might be larger for masked onsets, i.e. onsets in one of the sources that are obscured in the mixture, compared to onsets which are overt in the mixture. The model used in the last section is not well suited to capture such an effect, since it does not differentiate between masked and overt source onsets.

To test for a distinct response associated with the recovery of masked onsets in speech sources, we generated a new predictor to reflect masked onsets only, regardless of which source they originated from. This predictor was implemented as an element-wise comparison-based combination of onset spectrogram representations. Specifically, at each frequency- and time point, the predictor uses the (larger) source onset value but only by the amount it is over and

above the corresponding onset in the mixture, i.e., *max(0, max(attended, ignored) – mixture)*. This additional predictor improved predictions of brain responses in the ROI bilaterally (t_{max} = 8.12, $p < .001$), suggesting that responses in the auditory cortex indeed differentiate between overt and masked onsets.

Masked onsets are processed with a delay

Model comparison thus indicates that the neural representation of masked onsets differs from that of overt onsets. This implies that the influence of attention should also be assessed separately for overt and masked onsets. The previously used predictors do not allow this in a straight-forward manner, 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 frequency- and time-points in which an onset in the source is also accompanied by an onset in the mixture (element-wise *min(mixture, source)*), and the "masked onset" 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 ($t_{max} \leq 4.87$, $p < .001$). 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.

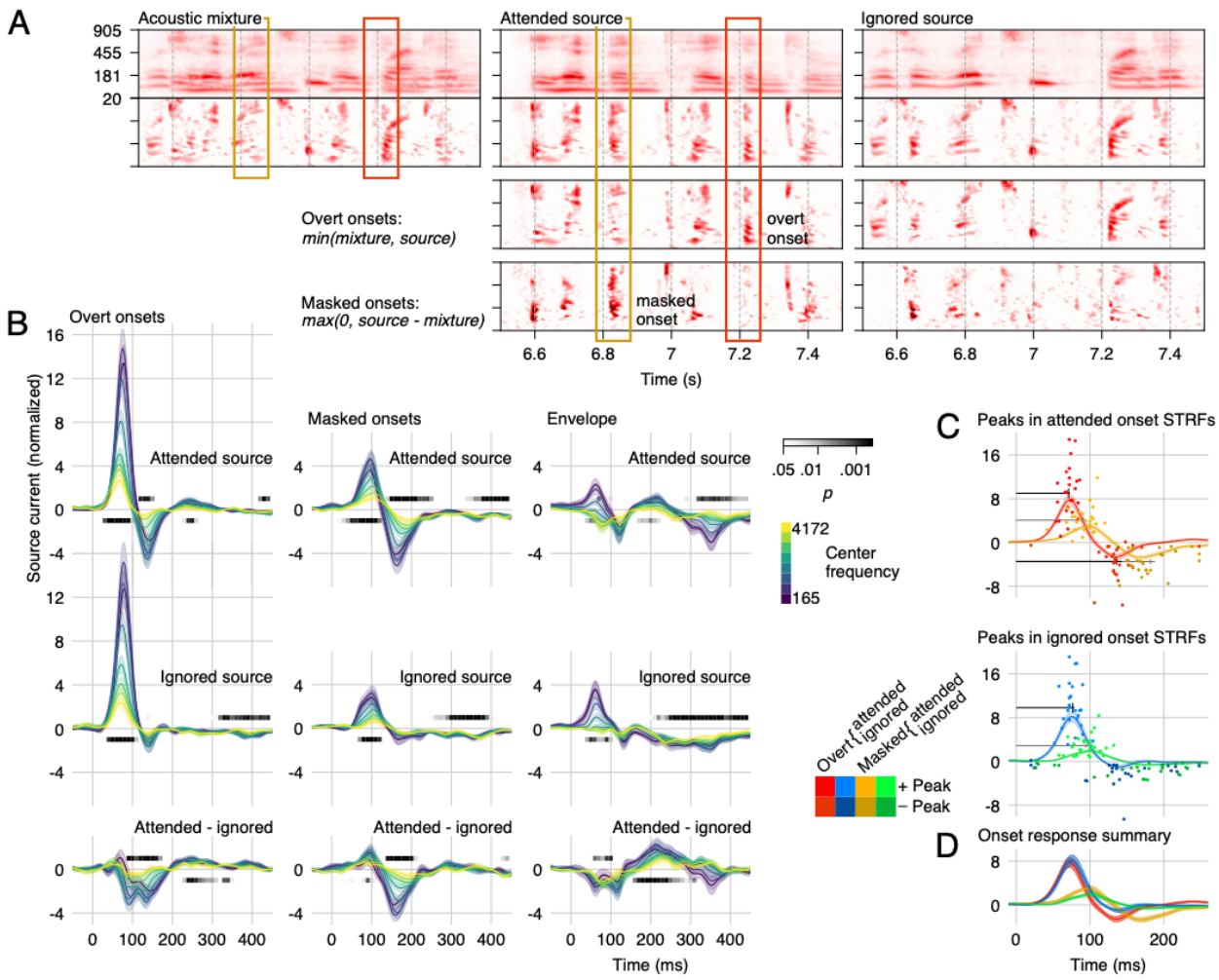

Figure 4. **Response functions to overt and masked onsets**. A) Spectrograms were transformed using element-wise operations to distinguish between overt onset, i.e., onsets in a source that are apparent in the mixture, and masked onsets, i.e., onsets in a source that are masked by the other source. Two examples are marked by rectangles: The yellow rectangle marks a region with a masked onset, i.e., an onset in the attended source which is not apparent in the mixture. The red square marks an overt onset, with an onset in the attended source that also corresponds to an onset in the mixture. B) STRFs exhibited the previously described positive-negative two peaked structure. For overt onsets, only the second, negative peak was modulated by attention. For obscured onsets, even the first peak exhibited a small degree of attentional modulation. C) Responses to masked onsets were 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 onset STRFs, averaged across frequency, ±1 SE.

The STRFs to each stream's overt onsets exhibited an early positive peak at ~74 ms that did not differentiate between onsets originating from the attended and unattended source, followed by a negative peak at ~140 ms with increased 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 (4). 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 reliably than onsets in the ignored source.

STRFs to masked onsets exhibited a similar positive-negative pattern as STRFs to overt onsets, but now with a consistent temporal delay of approximately 20 ms (Figure 4-C). The delay was 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. Even the earlier, positive peak was significantly larger for attended compared to ignored onsets. Thus, 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. While this might indicate that the two sources are segregated at this level, it does not necessarily mean that both sources are represented as individuated streams. Another explanation could be that masked onsets are evaluated early on, based on some available features, as to their likelihood of belonging to the attended source. Onsets that are more likely to belong to the attended source might then be represented more strongly, without yet being ascribed to one or the other source exclusively. Overall, the difference between the attended and ignored source suggests that 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 the auditory background is not as fully elaborated as the attended foreground (21).

Increasing abstraction over time

Responses to overt and masked onsets exhibited a comparable positive-negative two peak structure. While the first, positive peak was much larger for overt compared to masked onsets, the second, negative peak was of comparable magnitude (see Figure 4-D). This trend was confirmed in 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 was 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 (22), 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 could 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 there do exist phonetic contrasts where a distortion of a few tens of milliseconds would be relevant). Indeed, in audio-visual speech perception, temporal misalignment between auditory and visual input can actually be tolerated up to more than 100 ms (23).

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. 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 (24). At moderate signal to noise ratios (SNRs), a second talker causes much more interference with speech perception than a comparable non-speech masker and, interestingly, this interference manifests not just in the inability to hear attended words, but in intrusions of words from the ignored talker (25). 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 (26). Demonstrations of semantic processing of ignored speech are rare and usually associated with specific perceptual conditions such as dichotic presentation (27). Consistent with this, recent EEG/MEG evidence suggests that unattended speech is not processed in a time-locked fashion at the lexical (6) or semantic (7) level. The results presented here, showing systematic recovery of acoustic features from the ignored speech source, suggest a potential explanation for the increased interference from speech as opposed to other maskers. Representing onsets in two 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 (28), might sometimes capture attention, which could happen based on acoustic rather than lexical analysis (29). Finally, at very low SNRs this behavioral pattern can invert, and a background talker can be associated with better performance than stationary noise maskers (25). In such

conditions, there might be a benefit of being able to segregate the ignored speech source and use this information strategically (30).

Conclusions

How do listeners succeed in selectively listening to one of two concurrent talkers? Our results suggest that representations of acoustic onsets play 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. This recovery of masked onsets seems to be a cognitively costly process, reflected in a temporal delay of about 20 ms compared to overt onsets. Given the importance of temporal coherence for identifying auditory objects (31), it is likely that the onset representations play a key role in linking concurrent onsets at different frequency regions, and thus in segregating elements from the two auditory sources. While acoustic onsets are themselves relevant features for some phonetic contrasts, they also often precede informative regions in the spectrogram, such as the spectral detail of voiced segments. The onsets might thus also serve as cues to spectral regions in which relevant information is more likely to occur subsequently (10). Onsets might thus be used to decide which spectro-temporal features to group into an auditory object, and to further analyze as a perceptual entity. In our analysis, responses to these spectro-temporal features subsequent to onsets was modeled in the envelope predictors. If onsets are used to group features and allocate attention to information in the envelope, then this might explain why responses to the envelope predictors were affected by attention so early on.

Materials and Methods

Participants

The data analyzed here have been previously used in an unrelated analysis (6). 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 chapters (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 (32). Flat channels were automatically detected and excluded. Extraneous artifacts were removed with temporal signal space separation (33). 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 (34) 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 (35) ''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 (36, 37) 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 (38).

Predictor variables

Predictor variables were based on gammatone spectrograms sampled at 256 frequencies, ranging from 20 to 5000 Hz in ERB space (39), resampled to 1 kHz and scaled with exponent 0.6 (40). 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 (17) 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 $T = [-500, \dots, 500)$ (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 41). 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 (42) 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 43 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 $T = [0, \dots, 500)$ 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 (44) and a null distribution based on 10,000 permutations (43 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 $T = [-100, \dots, 500)$ 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 $[0, \dots, 450)$ 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 (45).

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 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.

References

1. D. S. Brungart, B. D. Simpson, 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 (2002).

2. G. Kidd, *et al.*, Determining the energetic and informational components of speech-on-speech masking. *J. Acoust. Soc. Am.* **140**, 132–144 (2016).

3. N. Ding, J. Z. Simon, Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 11854–9 (2012).

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

5. J. O'Sullivan, *et al.*, Hierarchical Encoding of Attended Auditory Objects in Multi-talker Speech Perception. *Neuron*, S0896627319307809 (2019).

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

7. M. P. Broderick, A. J. Anderson, G. M. D. Liberto, M. J. Crosse, E. C. Lalor, Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Curr. Biol.* **28**, 803-809.e3 (2018).

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

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

10. A. S. Bregman, P. Ahad, J. Kim, L. Melnerich, 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 (1994).

11. A. S. Bregman, P. A. Ahad, J. Kim, 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 (1994).

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

13. L. S. Hamilton, E. Edwards, E. F. Chang, A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Curr. Biol.* **28**, 1860-1871.e4 (2018).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

28. N. Wood, N. Cowan, 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 (1995).

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

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

31. S. Teki, M. Chait, S. Kumar, S. Shamma, T. D. Griffiths, Segregation of complex acoustic scenes based on temporal coherence. *eLife* **2** (2013).

32. A. Gramfort, *et al.*, MNE software for processing MEG and EEG data. *NeuroImage* **86**, 446–460 (2014).

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

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

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

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

37. A. M. Dale, M. I. Sereno, Improved Localizadon of Cortical Activity by Combining EEG and MEG with MRI Cortical Surface Reconstruction: A Linear Approach. *J. Cogn. Neurosci.* **5**, 162–176 (1993).

38. R. S. Desikan, *et al.*, An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968–980 (2006).

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

40. W. Biesmans, N. Das, T. Francart, A. Bertrand, 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 (2017).

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

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

43. C. Brodbeck, T. L. Brooks, P. Das, S. Reddigari, *Eelbrain 0.30* (Zenodo, 2019) https:/doi.org/10.5281/zenodo.2653785.

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

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