

Auditory Cortex Tracks Masked Acoustic Onsets in Background Speech: Evidence for Early Cortical Stream Segregation

Christian Brodbeck*¹, Alex Jiao², L. Elliot Hong³ & 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

17 Abstract

18 Humans are remarkably skilled at listening to one speaker out of an acoustic mixture of multiple
19 speech sources, even in the absence of binaural cues. Previous research on the neural
20 representations underlying this ability suggests that the auditory cortex primarily represents only
21 the unsegregated acoustic mixture in its early responses, and then selectively processes features
22 of the attended speech at longer latencies (from ~85 ms). The mechanism by which the attended
23 source signal is segregated from the mixture, however, and to what degree an ignored source
24 may also be segregated and separately processed, is not understood. We show here, in human
25 magnetoencephalographic responses to a two-talker mixture, an early neural representation of
26 acoustic onsets in the ignored speech source, over and above onsets of the mixture and the
27 attended source. This suggests that the auditory cortex initially reconstructs acoustic onsets
28 belonging to any speech source, critically, even when those onsets are acoustically masked by
29 another source. Overt onsets in the unseparated acoustic mixture were processed with a lower
30 latency (~70 ms) than masked onsets in either source (~90 ms), suggesting a neural processing
31 cost to the recovery of the masked onsets. Because acoustic onsets precede sustained source-
32 specific information in the acoustic spectrogram, these representations of onsets are cues
33 available for subsequent processing, including full stream segregation. Furthermore, these
34 findings suggest that even bottom-up saliency of objects in the auditory background may rely on
35 active cortical processing, explaining several behavioral effects of background speech.

36 Significance Statement

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

45 Author contributions

46 J.Z.S. and L.E.H. designed experiment and secured funding. C.B. and J.Z.S. analyzed data and
47 wrote the manuscript. A.J., C.B. and J.Z.S. performed simulations.

48

49 Introduction

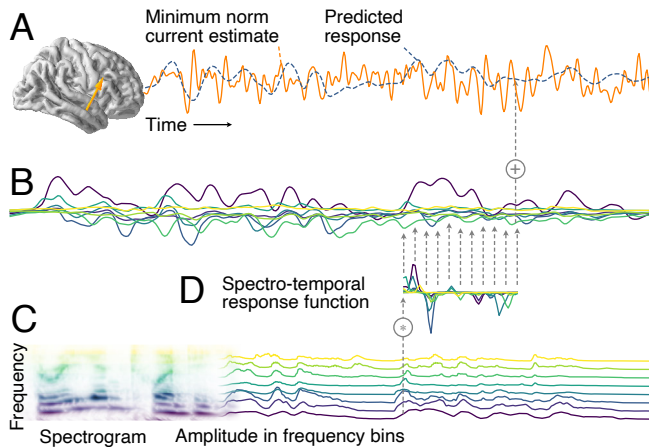
50 When listening to an acoustic scene, the acoustic signal that arrives at the ears is an additive
51 mixture of the different sound sources. Listeners trying to selectively attend to one of the sound
52 sources face the task of deciding which spectro-temporal features belong to that source. When
53 multiple speech sources are involved this is a nontrivial problem because the spectrograms of
54 the different sources often have strong overlap (see Figure 3-A). Nevertheless, human listeners
55 are remarkably skilled at focusing on one out of multiple talkers. Binaural cues can support
56 segregation of different sound sources based on their location (1), but are not necessary for this
57 ability, since listeners are able to selectively attend even when two speech signals are mixed into
58 a monophonic signal and presented with headphones (2).

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

75 Here we aim to distinguish between these hypotheses by analyzing auditory cortical
76 representations of two concurrent speech sources. An important cue for segregating an acoustic
77 source from a mixture is temporal coherence of different acoustic features (9). We focus in
78 particular on acoustic onset features, i.e., acoustic edges corresponding to a frequency-specific
79 increase in acoustic energy. A simultaneous onset of acoustic elements in distinct frequency
80 bands is a strong cue that these different elements originate from the same speech source.
81 Accordingly, shared acoustic onsets promote perceptual grouping of acoustic components into a
82 single auditory object, such as a complex tone and, vice versa, separate onsets lead to
83 perceptual segregation (10, 11). For example, the onset of a vowel is characterized by a shared
84 onset at the fundamental frequency of the voice and its harmonics. If the onset of a formant is
85 artificially offset by as little as 80 ms, it is often perceived as a separate tone rather than as a
86 component of the vowel (12). Acoustic onsets are very prominently represented in auditory
87 cortex, both in naturalistic speech (13, 14) and in non-speech stimuli (15), and are important for
88 speech intelligibility (16).

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

91 the ignored speech. Participants listened to 1-minute long continuous audiobook segments,
92 spoken by a male or a female speaker. Segments were presented in two conditions: as speech in
93 quiet, and as a two-talker mixture, in which a female and a male speaker were mixed at equal
94 loudness. MEG responses were analyzed as additive, linear response to multiple concurrent
95 stimulus features (see Figure 1). First, model comparison was used to determine which
96 representations significantly improved prediction of the responses. Then, spectro-temporal
97 response functions (STRFs) were analyzed to gain insight into the nature of the representations.



98

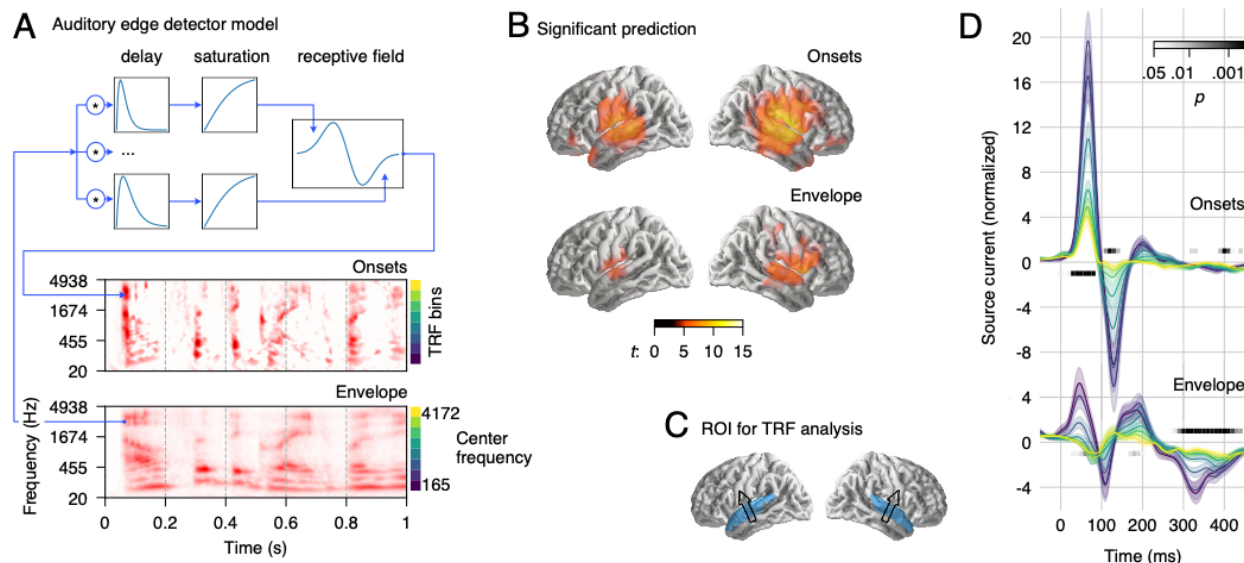
99 **Figure 1. Additive linear response model based on spectro-temporal response functions (STRFs).**
100 A) MEG responses recorded during stimulus presentation were source localized with distributed
101 minimum norm current estimates. A single virtual source dipole is shown for illustration, with its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

113 Results and Discussion

114 Auditory cortex represents acoustic onsets

115 MEG responses to speech presented in quiet were predicted from the gammatone spectrogram
116 of the stimulus, as well as a spectrogram of acoustic onsets (Figure 2-A). Acoustic onsets were
117 derived from a neural model of auditory edge detection (17). Both predictors were binned into 8
118 frequency bands, for a total of 16 predictor time series. Each of the two predictors was assessed
119 based on how well the correct model predicted MEG responses, compared to null models in
120 which the relevant predictor was temporally misaligned with the responses. Both predictors
121 significantly improved predictions ($p \leq 0.001$), with an anatomical distribution consistent with

122 sources in HG and STG bilaterally (Figure 2-B). Since this localization agrees with findings from
123 intracranial recordings (13), results were henceforth analyzed in a region of interest (ROI)
124 restricted to these two anatomical landmarks (Figure 2-C).



125
126 Figure 2. **Acoustic onset responses to clean speech.** A) Schematic illustration of the acoustic edge
127 detector model, along with an excerpt from a gammatone spectrogram (“envelope”) and the
128 corresponding onset representation. B) Regions of significant explanatory power of onset- and
129 envelope representations, consistent with a main source in auditory cortex bilaterally ($p \leq .05$,
130 corrected for whole brain analysis). C) Region of interest (ROI) used for the analysis of response
131 functions, including superior temporal gyrus and Heschl’s gyrus. An arrow indicates the average
132 current direction of the ROI (upward current), determined through the first principal component
133 of response power. D) Spectro-temporal response functions corresponding to onset and
134 envelope representations in the ROI. Different color curves reflect the frequency bins as
135 indicated next to the onset and envelope spectrograms in panel A. Shaded areas indicate the
136 within-subject standard error (18). Regions in which STRFs differ significantly from 0 (in any
137 band) are marked with horizontal gray bars.

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

144 STRFs to acoustic onsets exhibited a well-defined two-peaked shape, consistent across frequency
145 bands (Figure 2-D). They closely resembled previously described auditory response functions to
146 envelope representations, when these were used without consideration of onsets (3). In
147 comparison, envelope STRFs in the present results were diminished and exhibited a less well-
148 defined structure. This is consistent with acoustic onsets explaining a large portion of the signal
149 usually attributed to the envelope; indeed, when the model was refitted with only the envelope

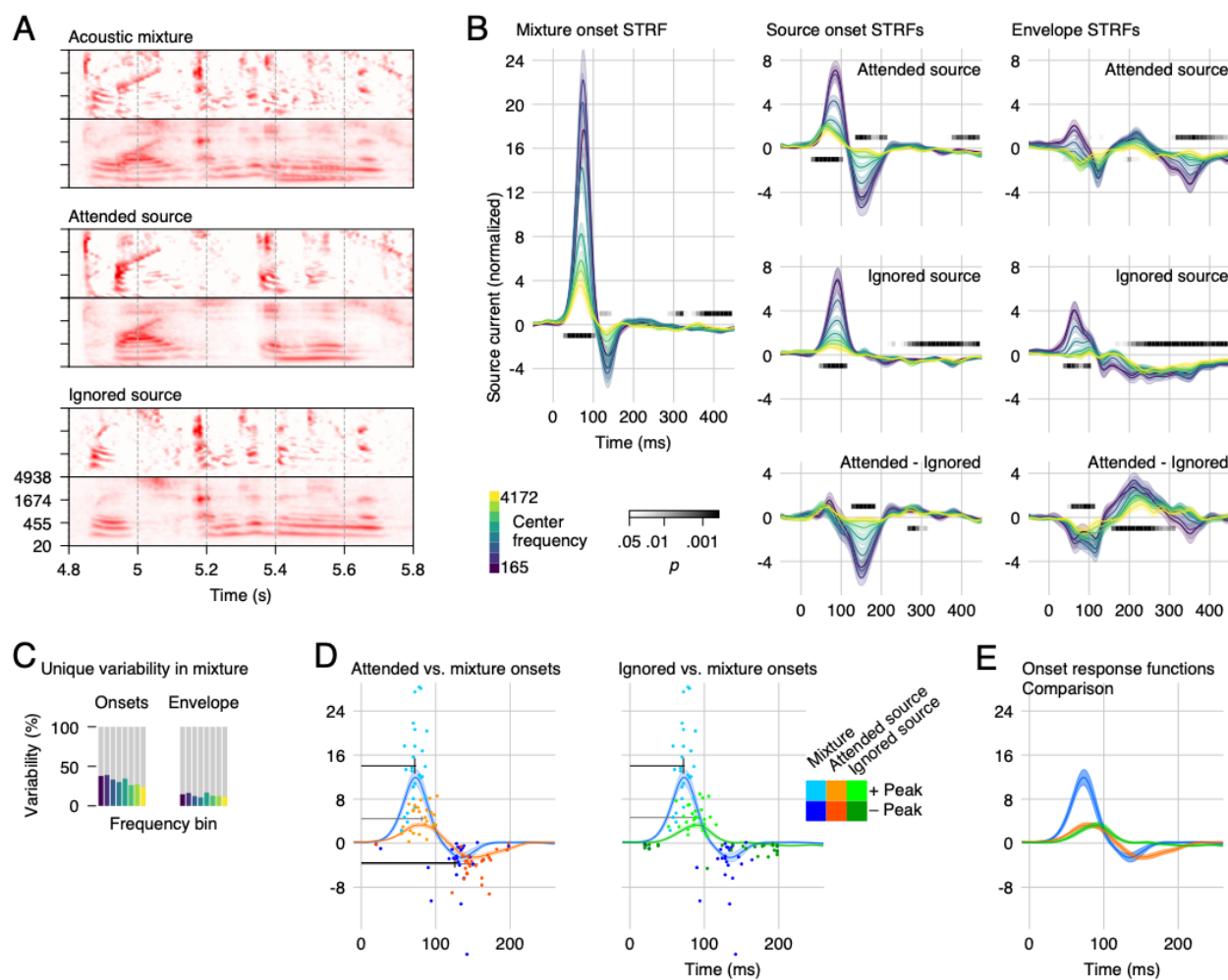
150 predictor, excluding the onset predictor, the envelope STRFs exhibited that canonical pattern
151 and with larger amplitudes (not shown).

152 STRFs had disproportionately higher amplitude at lower frequencies (Figure 2-D). This is
153 consistent with tonotopic mapping of speech areas and may follow from the spectral distribution
154 of information in the speech signal (19, 20). An explanation based on signal properties is also
155 supported by our simulations, in which equal TRFs for each band were simulated, and yet higher
156 frequency bands resulted in lower amplitude responses (see Figure SI-1).

157 Auditory cortex represents onsets of ignored speech

158 MEG responses to a two-speaker mixture were then used to test for a neural representation of
159 ignored speech. Participants listened to an equal loudness mixture of a male and a female talker
160 and were instructed to attend to one talker and ignore the other. The speaker to be attended
161 was counterbalanced across trials and subjects. Responses were predicted using the onset and
162 envelope representations for the acoustic mixture, the attended speech source and the ignored
163 source (Figure 3-A). Taken together, including the two predictors representing the ignored
164 speech significantly improved predictions of the responses in the ROI ($t_{max} = 8.32, p < .001$). This
165 indicates that acoustic features of the ignored speech are represented neurally in addition to
166 features of the mixture and the attended source. Separate tests suggested that this result can be
167 ascribed specifically to onset representations ($t_{max} = 4.89, p < .001$), whereas envelope
168 representations of the ignored source did not significantly improve the model fit ($t_{max} = -2.59, p =$
169 1).

170 Taken individually, onsets in each of the three streams significantly improved predictions ($t_{max} \geq$
171 $4.89, p < .001$), but none of the envelope representations did (all $t_{max} \leq -0.40, p = 1$). This lack of
172 predictive power for the envelope predictors, when tested individually, is likely due to high
173 collinearity. Intuitively, the envelope of the mixture can be approximated relatively well by the
174 sum of the envelopes of the individual streams (cf. Figure 3-A). More formally, the proportion of
175 the variability in the mixture representations that cannot be predicted from the two sources is
176 small for the envelopes, but substantially larger for the onsets (Figure 3-C). Accordingly, when
177 the mixture envelope predictor was removed from the model, the two source envelope
178 predictors became significant individually (attended: $t_{max} = 4.72, p = .002$; ignored: $t_{max} = 2.93, p$
179 $= .042$). Thus, as far as the envelope representations are concerned, the nature of the stimulus
180 representations prevents a conclusive distinction between representations of the acoustic
181 mixture and the ignored source. In contrast, onset representations do indicate a reliable
182 representation of ignored speech over and above representations of the acoustic mixture and
183 the attended source.



184

185 Figure 3. **Response functions to the two-speaker mixture, using the stream-based model.** A) The
 186 envelope and onsets of the acoustic mixture and the two speech sources were used to predict
 187 MEG responses. B) Auditory cortex STRFs to onsets in the mixture exhibit a large positive peak
 188 (72 ms) followed by a smaller negative peak (126 ms). STRFs to attended and ignored onsets
 189 both exhibit an early positive peak (81 and 88 ms), followed only in attended onsets by a
 190 negative peak (150 ms). This effect of attention on the negative peak is confirmed by the
 191 attended – ignored STRF differences. C) Compared to envelope representations, acoustic onset
 192 representations are better suited for distinguishing segregated sources from the mixture.
 193 Colored portions indicate proportion of the variability of the mixture predictors that could not be
 194 explained from the individual speech sources (with a -500 – 500 ms temporal integration
 195 window). D) The major peaks to onsets in the speech sources are delayed compared to
 196 corresponding peaks to the mixture. To determine latencies, mixture-based and individual-
 197 speaker-based STRFs were averaged across frequency (lines with shading for 1 SE). Colored dots
 198 represent the largest positive and negative peak for each participant between 20 and 200 ms;
 199 the peaks corresponding to individual speakers are delayed with respect the corresponding
 200 peaks for the mixture. Horizontal bars indicate average amplitude and latency ± 1 SE. D) Direct
 201 comparison of onset response functions averaged across frequency, ± 1 SE.

202 Onset STRFs exhibited the same characteristic positive-negative pattern as for speech in quiet,
203 but with reliable distinctions between the mixture and the individual speech streams (Figure 3-B,
204 left and middle columns, Figure 3-D & E). The early, positive peak occurred earlier and had a
205 larger amplitude for onsets in the mixture than for onsets in either of the sources (latency
206 mixture: 72 ms; attended: 81 ms, $t_{25} = 4.47$, $p < .001$; ignored: 88 ms, $t_{25} = 6.92$, $p < .001$;
207 amplitude mixture > attended: $t_{25} = 8.60$, $p < .001$; mixture > ignored: $t_{25} = 7.92$, $p < .001$). This
208 positive peak was followed by a negative peak only in responses to the mixture (126 ms) and the
209 attended source (150 ms; difference $t_{25} = 4.36$, $p < .001$). In contrast to the corresponding
210 positive peak, the amplitude of these negative peaks was statistically indistinguishable ($t_{25} =$
211 0.36 , $p = .722$). STRFs to the ignored source did not exhibit a detectable corresponding negative
212 peak, as seen in Figure 3-C where participants' peaks cluster around the time window edges
213 instead of at a characteristic latency.

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

218 In contrast to onsets, the different envelope predictors did not contain enough independent
219 information to distinguish between a representation of the ignored source and a representation
220 of the mixture. A comparison of STRFs to the attended and the ignored source revealed a strong
221 effect of attention (Figure 3-B, right column). The attended-ignored difference wave exhibits a
222 negative peak at ~ 100 , consistent with previous work (3), and an additional positive peak at ~ 200
223 ms. In contrast to previous work, however, a robust effect of attention on the envelope
224 representation starts almost as early as the earliest responses at all, suggesting that when onset
225 responses are accounted for separately from envelope responses, even early envelope
226 processing is influenced by attention.

227 [Auditory cortex recovers masked onsets](#)

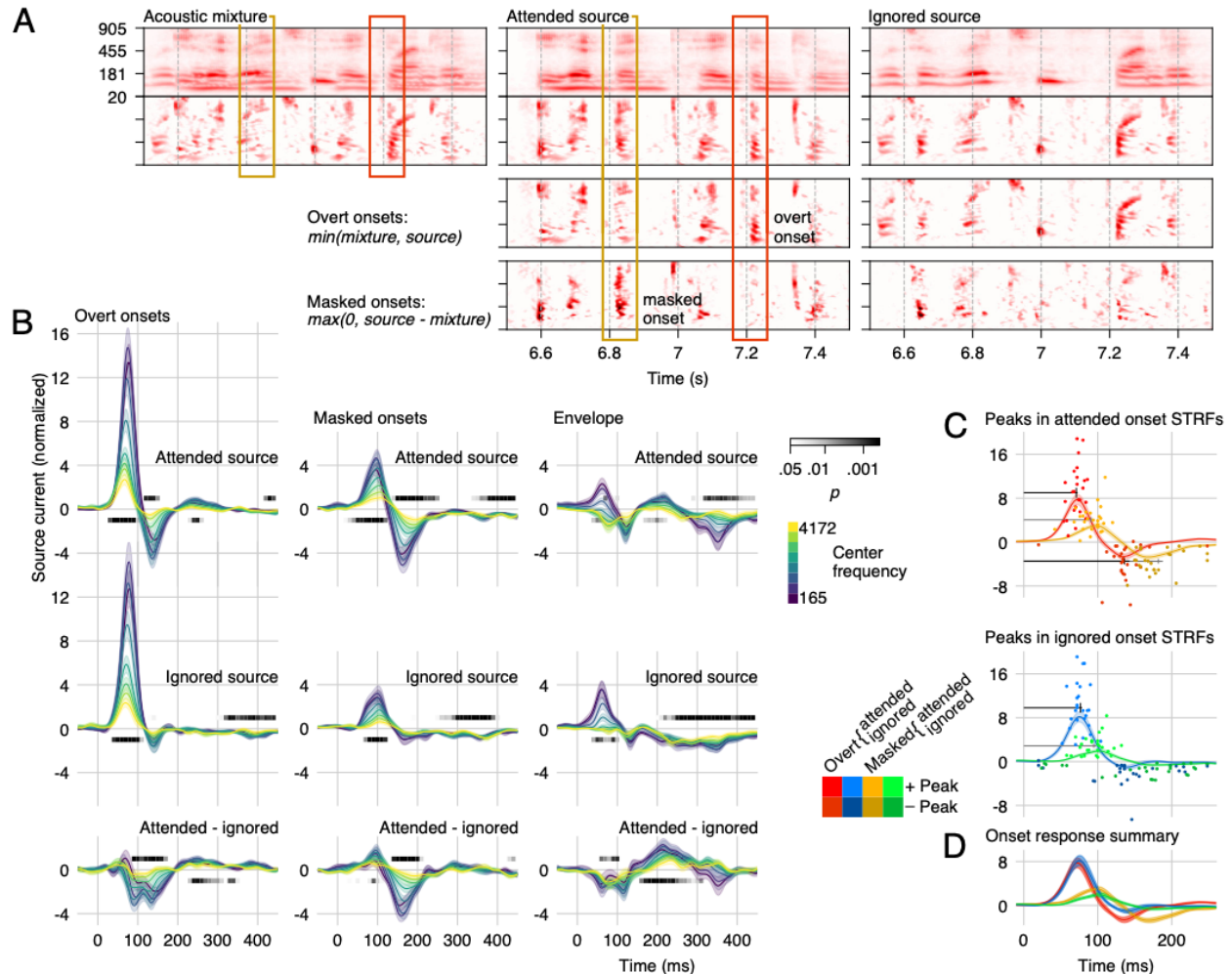
228 The results using these stream-based predictors suggest that the auditory cortex represents
229 acoustic onsets in both speech sources separately, in addition to onsets in the acoustic mixture.
230 This suggests a marked degree of abstraction from the acoustic input, involving early
231 reconstruction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

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

244 above the corresponding onset in the mixture, i.e., $\max(0, \max(\textit{attended}, \textit{ignored}) - \textit{mixture})$.
245 This additional predictor improved predictions of brain responses in the ROI bilaterally ($t_{\textit{max}} =$
246 $8.12, p < .001$), suggesting that responses in the auditory cortex indeed differentiate between
247 overt and masked onsets.

248 **Masked onsets are processed with a delay**

249 Model comparison thus indicates that the neural representation of masked onsets differs from
250 that of overt onsets. This implies that the influence of attention should also be assessed
251 separately for overt and masked onsets. The previously used predictors do not allow this in a
252 straight-forward manner, however, because the speech sources were modeled as unified
253 streams, combining overt and masked onsets. To separate effects of masking and attention, the
254 information from the previously used onset predictors was recombined to generate a new set of
255 predictors (Figure 4-A). Specifically, for each speech source, the new “overt onsets” predictor
256 models frequency- and time-points in which an onset in the source is also accompanied by an
257 onset in the mixture (element-wise $\min(\textit{mixture}, \textit{source})$), and the “masked onset” predictor
258 models the degree to which an onset in the source is attenuated (masked) in the mixture ($\max(0,$
259 $\textit{source} - \textit{mixture})$). This model thus disentangles the effect of attention (attended vs ignored
260 source) from whether an onset is overt in the mixture or masked. All four predictors significantly
261 improved MEG response predictions ($t_{\textit{max}} \leq 4.87, p < .001$). In particular, this was also true for
262 masked onsets in the ignored source ($t_{\textit{max}} = 4.87, p < .001$), confirming that the auditory cortex
263 recovers masked onsets even when they occur in the ignored source.



264
 265 Figure 4. Response functions to overt and masked onsets. A) Spectrograms were transformed
 266 using element-wise operations to distinguish between overt onset, i.e., onsets in a source that
 267 are apparent in the mixture, and masked onsets, i.e., onsets in a source that are masked by the
 268 other source. Two examples are marked by rectangles: The yellow rectangle marks a region with
 269 a masked onset, i.e., an onset in the attended source which is not apparent in the mixture. The
 270 red square marks an overt onset, with an onset in the attended source that also corresponds to
 271 an onset in the mixture. B) STRFs exhibited the previously described positive-negative two
 272 peaked structure. For overt onsets, only the second, negative peak was modulated by attention.
 273 For obscured onsets, even the first peak exhibited a small degree of attentional modulation. C)
 274 Responses to masked onsets were consistently delayed compared to responses to overt onsets.
 275 Details are analogous to Figure 3-D, except that the time window for finding peaks was extended
 276 to 20 – 250 ms to account for the longer latency of masked onset response functions. D) Direct
 277 comparison of the onset STRFs, averaged across frequency, ± 1 SE.

278 The STRFs to each stream's overt onsets exhibited an early positive peak at ~ 74 ms that did not
 279 differentiate between onsets originating from the attended and unattended source, followed by
 280 a negative peak at ~ 140 ms with increased amplitude for the attended source (Figure 4-B, left
 281 column). This suggests that the cortical processing stage corresponding to the first peak

282 represents onsets in the acoustic mixture without regard to their acoustic source (4). By the time
283 of the second peak, however, the cortical representations distinguish between the two sources,
284 with onsets in the attended source being represented more reliably than onsets in the ignored
285 source.

286 STRFs to masked onsets exhibited a similar positive-negative pattern as STRFs to overt onsets,
287 but now with a consistent temporal delay of approximately 20 ms (Figure 4-C). The delay was
288 significant for both streams' positive peak (attended overt: 71 ms, masked: 91 ms, $t_{25} = 6.77$, $p <$
289 $.001$; unattended overt: 77 ms, masked: 95 ms, $t_{25} = 7.23$, $p < .001$), as well as for the negative
290 peak to attended onsets (overt: 136 ms, masked: 182 ms; $t_{25} = 4.72$, $p < .001$). For masked
291 onsets in the ignored source, there is no evidence for a consistent negative peak at all, as can be
292 seen in Figure 4-C where data points are spread throughout the time window. Even the earlier,
293 positive peak was significantly larger for attended compared to ignored onsets. Thus, auditory
294 cortex not only represents masked onsets, but these representations are substantively affected
295 by whether the onset belongs to the attended or the ignored source. While this might indicate
296 that the two sources are segregated at this level, it does not necessarily mean that both sources
297 are represented as individuated streams. Another explanation could be that masked onsets are
298 evaluated early on, based on some available features, as to their likelihood of belonging to the
299 attended source. Onsets that are more likely to belong to the attended source might then be
300 represented more strongly, without yet being ascribed to one or the other source exclusively.
301 Overall, the difference between the attended and ignored source suggests that information from
302 the ignored source is represented to a lesser degree than information from the attended source.
303 This is consistent with evidence from psychophysics suggesting that the auditory background is
304 not as fully elaborated as the attended foreground (21).

305 [Increasing abstraction over time](#)

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

322 Attentive processing is not strictly time-locked

323 While the response magnitude to overt and masked onsets thus seems to be adjusted at
324 subsequent processing stages, the response latency was not. Representations of masked onsets
325 were consistently delayed compared to those of overt onsets by approximately 20 ms (see
326 Figure 4-D). Previous research found that the latency of the representation of speech increased
327 with increasing levels of stationary noise (22), suggesting a processing cost to recovering acoustic
328 source information from noise. Our results suggest that this is not a uniform delay for a given
329 perceptual stream, but that the delay varies by whether an acoustic element is overt or locally
330 masked by the acoustic background. The delay might thus arise from a variable processing cost
331 that depends on the local acoustic environment.

332 This latency difference between representations of overt and masked onsets entails that
333 upstream speech processing mechanisms may receive different packages of information about
334 the attended speech source with some temporal desynchronization. While this could imply a
335 need for a higher order corrective mechanism, it is also possible that upstream mechanisms are
336 tolerant to this small temporal distortion. A misalignment of 20 ms is small compared to the
337 normal temporal variability encountered in speech (although there do exist phonetic contrasts
338 where a distortion of a few tens of milliseconds would be relevant). Indeed, in audio-visual
339 speech perception, temporal misalignment between auditory and visual input can actually be
340 tolerated up to more than 100 ms (23).

341 Processing of “ignored” acoustic sources

342 The interference in speech perception from a second talker can be very different from the
343 interference caused by non-speech sounds. Music is cortically segregated from speech even
344 when both signals are unattended, consistent with a more automatic segregation, possibly due
345 to distinctive differences in acoustic signal properties (24). At moderate signal to noise ratios
346 (SNRs), a second talker causes much more interference with speech perception than a
347 comparable non-speech masker and, interestingly, this interference manifests not just in the
348 inability to hear attended words, but in intrusions of words from the ignored talker (25). The
349 latter fact in particular has been interpreted as evidence that ignored speech might be
350 segregated and processed to a relatively high level. On the other hand, listeners seem to be
351 unable to access words in more than one speech source at a time, even when the sources are
352 spatially separated (26). Demonstrations of semantic processing of ignored speech are rare and
353 usually associated with specific perceptual conditions such as dichotic presentation (27).
354 Consistent with this, recent EEG/MEG evidence suggests that unattended speech is not
355 processed in a time-locked fashion at the lexical (6) or semantic (7) level. The results presented
356 here, showing systematic recovery of acoustic features from the ignored speech source, suggest
357 a potential explanation for the increased interference from speech as opposed to other maskers.
358 Representing onsets in two sources could be expected to increase cognitive load compared to
359 detecting onsets of a single source in stationary noise. These representations of ignored speech
360 might also act as bottom-up cues and cause the tendency for intrusions from the ignored talker.
361 They might even explain why a salient and overlearned word, such as one’s own name (28),
362 might sometimes capture attention, which could happen based on acoustic rather than lexical
363 analysis (29). Finally, at very low SNRs this behavioral pattern can invert, and a background talker
364 can be associated with better performance than stationary noise maskers (25). In such

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

367 Conclusions

368 How do listeners succeed in selectively listening to one of two concurrent talkers? Our results
369 suggest that representations of acoustic onsets play a critical role. Early responses in the
370 auditory cortex represent not only overt acoustic onsets, but also reconstruct acoustic onsets in
371 the speech sources that are masked in the mixture. This recovery of masked onsets seems to be
372 a cognitively costly process, reflected in a temporal delay of about 20 ms compared to overt
373 onsets. Given the importance of temporal coherence for identifying auditory objects (31), it is
374 likely that the onset representations play a key role in linking concurrent onsets at different
375 frequency regions, and thus in segregating elements from the two auditory sources. While
376 acoustic onsets are themselves relevant features for some phonetic contrasts, they also often
377 precede informative regions in the spectrogram, such as the spectral detail of voiced segments.
378 The onsets might thus also serve as cues to spectral regions in which relevant information is
379 more likely to occur subsequently (10). Onsets might thus be used to decide which spectro-
380 temporal features to group into an auditory object, and to further analyze as a perceptual entity.
381 In our analysis, responses to these spectro-temporal features subsequent to onsets was modeled
382 in the envelope predictors. If onsets are used to group features and allocate attention to
383 information in the envelope, then this might explain why responses to the envelope predictors
384 were affected by attention so early on.

385 Materials and Methods

386 Participants

387 The data analyzed here have been previously used in an unrelated analysis (6). MEG responses
388 were recorded from 28 native speakers of English, recruited by media advertisements from the
389 Baltimore area. Participants with medical, psychiatric or neurological illnesses, head injury, and
390 substance dependence or abuse were excluded. All subjects provided informed consent in
391 accordance with the University of Maryland Baltimore Institutional Review Board and were paid
392 for their participation. Data from two participants were excluded, one due to corrupted localizer
393 measurements, and one due to excessive magnetic artifacts associated with dental work,
394 resulting in a final sample of 18 male and 8 female participants with mean age 45.2 (range 22 -
395 61).

396 Stimuli

397 Two chapters were selected from an audiobook recording of A Child's History of England by
398 Charles Dickens, one chapter read by a male and one by a female speaker ([https://librivox.org/a-
399 childs-history-of-england-by-charles-dickens/](https://librivox.org/a-childs-history-of-england-by-charles-dickens/), chapters 3 and 8). Four 1 minute long segments
400 were extracted from each chapters (referred to as male-1 through 4 and female 1 through 4).
401 Pauses longer than 300 ms were shortened to an interval randomly chosen between 250 and
402 300 ms, and loudness was matched perceptually. Two-talker stimuli were generated by
403 additively combining two segments, one from each speaker, with an initial 1 s period containing
404 only the to-be attended speaker (mix-1 through 4 were constructed by mixing male-1 and
405 female-1, through 4).

406 Procedure

407 During MEG data acquisition, participants lay supine and were instructed to keep their eyes
408 closed to minimize ocular artifacts and head movement. Stimuli were delivered through foam
409 pad earphones inserted into the ear canal at a comfortably loud listening level.

410 Participants listened four times to mix-1, while attending to one speaker and ignoring the other
411 (which speaker they attended to was counterbalanced across subject), then 4 times to mix-2
412 while attending to the other speaker. After each segment, participants answered a question
413 relating to the content of the attended stimulus. Then, the four segments just heard were all
414 presented once each, as single talkers. The same procedure was repeated for stimulus segments
415 3 and 4.

416 Data acquisition and preprocessing

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

424 Recordings were pre-processed using mne-python (32). Flat channels were automatically
425 detected and excluded. Extraneous artifacts were removed with temporal signal space
426 separation (33). Data were filtered between 1 and 40 Hz with a zero-phase FIR filter (mne-
427 python 0.15 default settings). Extended infomax independent component analysis (34) was then
428 used to remove ocular and cardiac artifacts. Responses time-locked to the onset of the speech
429 stimuli were extracted and downsampled to 100 Hz. For responses to the two-talker mixture, the
430 first second of data, in which only the to-be attended talker was heard, was discarded.

431 Five marker coils attached to subjects' head served to localize the head position with respect to
432 the MEG sensors. Two measurements, one at the beginning and one at the end of the recording
433 were averaged. The FreeSurfer (35) "fsaverage" template brain was coregistered to each
434 subject's digitized head shape (Polhemus 3SPACE FASTRAK) using rotation, translation, and
435 uniform scaling. A source space was generated using four-fold icosahedral subdivision of the
436 white matter surface, with source dipoles oriented perpendicularly to the cortical surface.
437 Minimum ℓ_2 norm current estimates (36, 37) were computed for all data. Initial analysis was
438 performed on the whole brain as identified by the FreeSurfer "cortex" label. Subsequent
439 analyses were restricted to sources in the STG and Heschl's gyrus as identified in the "aparc"
440 parcellation (38).

441 Predictor variables

442 Predictor variables were based on gammatone spectrograms sampled at 256 frequencies,
443 ranging from 20 to 5000 Hz in ERB space (39), resampled to 1 kHz and scaled with exponent 0.6
444 (40). At this point, different stimulus representations were computed. Spectrograms were then
445 binned into 8 frequency bands equally spaced in ERB space (omitting frequencies below 100 Hz

446 because the female speaker had little power below that frequency) and resampled to match the
447 MEG data.

448 Acoustic onset representations were computed by applying an auditory edge detection model
449 (17) independently to each frequency band of the spectrogram. The model was implemented
450 with a delay layer with 10 delays ranging from $\tau_2 = 3$ to 5 ms, a saturation scaling factor of $C =$
451 **30**, and a receptive field based on the derivative of a Gaussian window with $SD = 2$ ms.
452 Negative values in the resulting onset spectrogram were set to 0.

453 The linear dependence between different predictor variables (Figure 3-C) was estimated by
454 treating each predictor time series in turn as the dependent measure and predicting it from the
455 other predictors through a kernel with $T = [-500, \dots, 500]$ (see next section). For example,
456 segments [male-1, female-2, male-3, female-4] were combined, and each of the 8 bands in this
457 predictor were predicted from [[female-1, mix-1], [male-2, mix-2], ...] (including all 8 bands). The
458 same parameters were used as for fitting neural models, except that no temporal basis function
459 was used. The measure of interest was the proportion of the (ℓ_1) variability of the dependent
460 variable that could not be explained from a linear combination of the other variables.

461 Reverse correlation

462 Spectro-temporal response functions (STRFs) were computed independently for each virtual
463 current source (see 41). The neural response at time t , y_t was predicted from the sum of N
464 predictor variables x_n convolved with a corresponding response function h_n of length T :

$$465 \hat{y}_t = \sum_n^N \sum_{\tau}^T h_{n,\tau} \cdot x_{i,t-\tau}$$

466 STRFs were generated from a basis of 50 ms wide Hamming windows and were estimated using
467 an iterative coordinate descent algorithm (42) to minimize the ℓ_1 error. Early stopping was
468 based on 4-fold split of the data, freezing each h_n when it lead to an increase of error in the
469 testing data (see 43 for further details).

470 Model tests

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

$$473 MEG \sim o + e$$

474 Where o =onsets and e =envelope. Models were estimated with STRFs with $T = [0, \dots, 500]$ ms.
475 In order to test the predictive power of each predictor, three corresponding null models were
476 generated by temporally misaligning the predictor with the response by cyclically shifting the
477 predictor for each segment by 15, 30 and 45 seconds. Model quality was quantified as the
478 Pearson correlation between actual and predicted response. For each predictor, the model
479 quality of the full model was compared with the average model quality of the three
480 corresponding null models using a mass-univariate related measures t -test with threshold-free
481 cluster enhancement (44) and a null distribution based on 10,000 permutations (43 for further
482 details).

483 Initially, responses to speech in noise was predicted from:

484
$$MEG \sim o_{mix} + o_{att} + o_{ign} + e_{mix} + e_{att} + e_{ign}$$

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

487
$$MEG \sim o_{att,over} + o_{ign,over} + o_{att,mask} + o_{ign,mask} + e_{att} + e_{ign}$$

488

489 STRF tests

490 To evaluate STRFs, the corresponding model (only correctly aligned predictors) was refit with
491 $T = [-100, \dots, 500]$ ms to include an estimate of baseline activity (due to occasional edge
492 artifacts, STRFs are displayed between -50 to 450 ms).

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

502 In order to make TRFs more comparable across subjects, they were smoothed on the frequency
503 axis with a Hamming window of width 7. STRFs were statistically analyzed in the time range
504 $[0, \dots, 450]$ ms using mass-univariate *t*-tests and ANOVAs, with *p*-values calculated from null
505 distributions based on the maximum statistic (*t*, *F*) in 10,000 permutations (45).

506 Acknowledgements

507 This work was supported by a National Institutes of Health grant R01-DC014085 (to J.Z.S.) and by
508 a University of Maryland Seed Grant (to L.E.H. and J.Z.S.). We would like to thank Krishna
509 Puvvada for his assistance in designing and preparing the stimuli and Natalia Lapinskaya for her
510 help in collecting data and for excellent technical support.

511 References

- 512 1. D. S. Brungart, B. D. Simpson, The effects of spatial separation in distance on the
513 informational and energetic masking of a nearby speech signal. *J. Acoust. Soc. Am.* **112**,
514 664–676 (2002).
- 515 2. G. Kidd, *et al.*, Determining the energetic and informational components of speech-on-
516 speech masking. *J. Acoust. Soc. Am.* **140**, 132–144 (2016).
- 517 3. N. Ding, J. Z. Simon, Emergence of neural encoding of auditory objects while listening to
518 competing speakers. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 11854–9 (2012).

- 519 4. K. C. Puvvada, J. Z. Simon, Cortical Representations of Speech in a Multitalker Auditory
520 Scene. *J. Neurosci.* **37**, 9189–9196 (2017).
- 521 5. J. O’Sullivan, *et al.*, Hierarchical Encoding of Attended Auditory Objects in Multi-talker
522 Speech Perception. *Neuron*, S0896627319307809 (2019).
- 523 6. C. Brodbeck, L. E. Hong, J. Z. Simon, Rapid Transformation from Auditory to Linguistic
524 Representations of Continuous Speech. *Curr. Biol.* **28**, 3976–3983.e5 (2018).
- 525 7. M. P. Broderick, A. J. Anderson, G. M. D. Liberto, M. J. Crosse, E. C. Lalor,
526 Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of
527 Natural, Narrative Speech. *Curr. Biol.* **28**, 803–809.e3 (2018).
- 528 8. R. P. Carlyon, How the brain separates sounds. *Trends Cogn. Sci.* **8**, 465–471 (2004).
- 529 9. M. Elhilali, L. Ma, C. Micheyl, A. J. Oxenham, S. A. Shamma, Temporal Coherence in the
530 Perceptual Organization and Cortical Representation of Auditory Scenes. *Neuron* **61**, 317–
531 329 (2009).
- 532 10. A. S. Bregman, P. Ahad, J. Kim, L. Melnerich, Resetting the pitch-analysis system: 1. Effects
533 of rise times of tones in noise backgrounds or of harmonics in a complex tone. *Percept.*
534 *Psychophys.* **56**, 155–162 (1994).
- 535 11. A. S. Bregman, P. A. Ahad, J. Kim, Resetting the pitch-analysis system. 2. Role of sudden
536 onsets and offsets in the perception of individual components in a cluster of overlapping
537 tones. *J. Acoust. Soc. Am.* **96**, 2694–2703 (1994).
- 538 12. R. W. Hukin, C. J. Darwin, Comparison of the effect of onset asynchrony on auditory
539 grouping in pitch matching and vowel identification. *Percept. Psychophys.* **57**, 191–196
540 (1995).
- 541 13. L. S. Hamilton, E. Edwards, E. F. Chang, A Spatial Map of Onset and Sustained Responses to
542 Speech in the Human Superior Temporal Gyrus. *Curr. Biol.* **28**, 1860–1871.e4 (2018).
- 543 14. C. Daube, R. A. A. Ince, J. Gross, Simple Acoustic Features Can Explain Phoneme-Based
544 Predictions of Cortical Responses to Speech. *Curr. Biol.* **29**, 1924–1937.e9 (2019).
- 545 15. Y. Zhou, X. Wang, Cortical Processing of Dynamic Sound Envelope Transitions. *J. Neurosci.*
546 **30**, 16741–16754 (2010).
- 547 16. C. E. Stilp, K. R. Kluender, Cochlea-scaled entropy, not consonants, vowels, or time, best
548 predicts speech intelligibility. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 12387–12392 (2010).
- 549 17. A. Fishbach, I. Nelken, Y. Yeshurun, Auditory Edge Detection: A Neural Model for
550 Physiological and Psychoacoustical Responses to Amplitude Transients. *J. Neurophysiol.* **85**,
551 2303–2323 (2001).

- 552 18. G. R. Loftus, M. E. J. Masson, Using confidence intervals in within-subject designs. *Psychon.*
553 *Bull. Rev.* **1**, 476–490 (1994).
- 554 19. M. Moerel, F. De Martino, E. Formisano, Processing of Natural Sounds in Human Auditory
555 Cortex: Tonotopy, Spectral Tuning, and Relation to Voice Sensitivity. *J. Neurosci.* **32**, 14205–
556 14216 (2012).
- 557 20. P. W. Hullett, L. S. Hamilton, N. Mesgarani, C. E. Schreiner, E. F. Chang, Human Superior
558 Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech
559 Stimuli. *J. Neurosci.* **36**, 2014–2026 (2016).
- 560 21. B. G. Shinn-Cunningham, A. K. C. Lee, A. J. Oxenham, A sound element gets lost in
561 perceptual competition. *Proc. Natl. Acad. Sci.* **104**, 12223–12227 (2007).
- 562 22. N. Ding, J. Z. Simon, Adaptive Temporal Encoding Leads to a Background-Insensitive Cortical
563 Representation of Speech. *J. Neurosci.* **33**, 5728–5735 (2013).
- 564 23. V. van Wassenhove, K. W. Grant, D. Poeppel, Temporal window of integration in auditory-
565 visual speech perception. *Neuropsychologia* **45**, 598–607 (2007).
- 566 24. L. Hausfeld, L. Riecke, G. Valente, E. Formisano, Cortical tracking of multiple streams
567 outside the focus of attention in naturalistic auditory scenes. *NeuroImage* **181**, 617–626
568 (2018).
- 569 25. D. S. Brungart, Informational and energetic masking effects in the perception of two
570 simultaneous talkers. *J. Acoust. Soc. Am.* **109**, 1101–1109 (2001).
- 571 26. G. Kidd, T. L. Arbogast, C. R. Mason, F. J. Gallun, The advantage of knowing where to listen. *J*
572 *Acoust Soc Am* **118**, 12 (2005).
- 573 27. M. Rivenez, C. J. Darwin, A. Guillaume, Processing unattended speech. *J. Acoust. Soc. Am.*
574 **119**, 4027–4040 (2006).
- 575 28. N. Wood, N. Cowan, The cocktail party phenomenon revisited: How frequent are attention
576 shifts to one's name in an irrelevant auditory channel? *J. Exp. Psychol. Learn. Mem. Cogn.*
577 **21**, 255–260 (1995).
- 578 29. K. J. P. Woods, J. H. McDermott, Schema learning for the cocktail party problem. *Proc. Natl.*
579 *Acad. Sci.* **115**, E3313–E3322 (2018).
- 580 30. L. Fiedler, M. Wöstmann, S. K. Herbst, J. Obleser, Late cortical tracking of ignored speech
581 facilitates neural selectivity in acoustically challenging conditions. *NeuroImage* **186**, 33–42
582 (2019).
- 583 31. S. Teki, M. Chait, S. Kumar, S. Shamma, T. D. Griffiths, Segregation of complex acoustic
584 scenes based on temporal coherence. *eLife* **2** (2013).

- 585 32. A. Gramfort, *et al.*, MNE software for processing MEG and EEG data. *NeuroImage* **86**, 446–
586 460 (2014).
- 587 33. S. Taulu, J. Simola, Spatiotemporal signal space separation method for rejecting nearby
588 interference in MEG measurements. *Phys. Med. Biol.* **51**, 1759 (2006).
- 589 34. A. J. Bell, T. J. Sejnowski, An Information-Maximization Approach to Blind Separation and
590 Blind Deconvolution. *Neural Comput.* **7**, 1129–1159 (1995).
- 591 35. B. Fischl, FreeSurfer. *NeuroImage* **62**, 774–781 (2012).
- 592 36. M. S. Hämäläinen, R. J. Ilmoniemi, Interpreting magnetic fields of the brain: minimum norm
593 estimates. *Med. Biol. Eng. Comput.* **32**, 35–42 (1994).
- 594 37. A. M. Dale, M. I. Sereno, Improved Localizadon of Cortical Activity by Combining EEG and
595 MEG with MRI Cortical Surface Reconstruction: A Linear Approach. *J. Cogn. Neurosci.* **5**,
596 162–176 (1993).
- 597 38. R. S. Desikan, *et al.*, An automated labeling system for subdividing the human cerebral
598 cortex on MRI scans into gyral based regions of interest. *NeuroImage* **31**, 968–980 (2006).
- 599 39. J. Heeris, *Gammatone Filterbank Toolkit* (2018).
- 600 40. W. Biesmans, N. Das, T. Francart, A. Bertrand, Auditory-Inspired Speech Envelope Extraction
601 Methods for Improved EEG-Based Auditory Attention Detection in a Cocktail Party Scenario.
602 *IEEE Trans. Neural Syst. Rehabil. Eng.* **25**, 402–412 (2017).
- 603 41. C. Brodbeck, A. Presacco, J. Z. Simon, Neural source dynamics of brain responses to
604 continuous stimuli: Speech processing from acoustics to comprehension. *NeuroImage* **172**,
605 162–174 (2018).
- 606 42. S. V. David, N. Mesgarani, S. A. Shamma, Estimating sparse spectro-temporal receptive
607 fields with natural stimuli. *Netw. Comput. Neural Syst.* **18**, 191–212 (2007).
- 608 43. C. Brodbeck, T. L. Brooks, P. Das, S. Reddigari, *Eelbrain 0.30* (Zenodo, 2019)
609 <https://doi.org/10.5281/zenodo.2653785>.
- 610 44. S. M. Smith, T. E. Nichols, Threshold-free cluster enhancement: Addressing problems of
611 smoothing, threshold dependence and localisation in cluster inference. *NeuroImage* **44**,
612 83–98 (2009).
- 613 45. E. Maris, R. Oostenveld, Nonparametric statistical testing of EEG- and MEG-data. *J.*
614 *Neurosci. Methods* **164**, 177–190 (2007).