

 Open access • Posted Content • DOI:10.1101/238642

## **Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions** — [Source link](#)

Lorenz Fiedler, Malte Woestmann, Sophie K. Herbst, Jonas Obleser

**Institutions:** University of Lübeck

**Published on:** 25 Jul 2018 - bioRxiv (Cold Spring Harbor Laboratory)

Related papers:

- [Late cortical tracking of ignored speech facilitates neural selectivity in acoustically challenging conditions.](#)
- [Unfolding of a noise-invariant neural representation of attended speech](#)
- [Neural speech restoration at the cocktail party: Auditory cortex recovers masked speech of both attended and ignored speakers](#)
- [Auditory Cortex Tracks Masked Acoustic Onsets in Background Speech: Evidence for Early Cortical Stream Segregation](#)
- [Effect of Task and Attention on Neural Tracking of Speech.](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/late-cortical-tracking-of-ignored-speech-facilitates-neural-294vdpc57s>

1     **Late cortical tracking of ignored speech facilitates neural selectivity**  
2                     **in acoustically challenging conditions**

3  
4                     Lorenz Fiedler,  
5                     Malte Wöstmann,  
6                     Sophie K. Herbst,  
7                     & Jonas Obleser

8  
9             Department of Psychology, University of Lübeck, Lübeck, Germany

10    *Running title: Late cortical tracking of ignored speech*

11    Keywords: attention, EEG, forward encoding models, speech, auditory  
12    cortex, fronto-parietal attention network, SNR

13

14    Author correspondence:

15    Lorenz Fiedler & Jonas Obleser

16    Department of Psychology, University of Lübeck

17    Maria-Goeppert Straße 9a

18    23562 Lübeck, Germany

19    [lorenz.fiedler@uni-luebeck.de](mailto:lorenz.fiedler@uni-luebeck.de); [jonas.obleser@uni-luebeck.de](mailto:jonas.obleser@uni-luebeck.de)

20

21    *Number of figures: 4*

22    *Number of tables: 0*

23    *Number of words: 7,778*

24    Abstract (150), Introduction (684), Discussion (1340)

25

26    *Acknowledgments:* Research was supported by the European Research  
27    Council (ERC-CoG-2014 646696 to JO) and the Oticon Foundation (NEURO-  
28    CHAT).

29 **Abstract**

30 Listening requires selective neural processing of the incoming sound  
31 mixture, which in humans is borne out by a surprisingly clean  
32 representation of attended-only speech in auditory cortex. How this neural  
33 selectivity is achieved even at negative signal-to-noise ratios (SNR) remains  
34 unclear. We show that, under such conditions, a late cortical representation  
35 (i.e., neural tracking) of the ignored acoustic signal is key to successful  
36 separation of attended and distracting talkers (i.e., neural selectivity). We  
37 recorded and modelled the electroencephalographic response of 18  
38 participants who attended to one of two simultaneously presented stories,  
39 while the SNR between the two talkers varied dynamically. The neural  
40 tracking showed an increasing early-to-late attention-biased selectivity.  
41 Importantly, acoustically dominant ignored talkers were tracked neurally  
42 by late involvement of fronto-parietal regions, which contributed to  
43 enhanced neural selectivity. This neural selectivity by way of representing  
44 the ignored talker poses a mechanistic neural account of attention under  
45 real-life acoustic conditions.

## 46 **Introduction**

47 Human listeners comprehend speech surprisingly well in the presence of  
48 distracting sound sources (Cherry, 1953). The ubiquitous question is how  
49 competing acoustic events capture bottom-up attention (e.g., by being  
50 dominant, that is, louder than the background), and how in turn top-down  
51 selective attention can overcome this dominance (e.g., listening to a certain  
52 talker against varying levels of competing talkers or noise; Kaya and Elhilali,  
53 2017).

54 Auditory selective neural processing has been mainly attributed to auditory  
55 cortex regions. It is by now well-established that the auditory cortical  
56 system selectively represents the (spectro-)temporal envelope of attended,  
57 but not ignored speech (i.e., neural phase-locking; Magneto-  
58 encephalography: Ding and Simon, 2012; Electroencephalography: Kerlin  
59 at al., 2010; Power et al., 2012; Horton et al., 2013; O'Sullivan et al., 2014).  
60 Accordingly, auditory cortical responses allow for a reconstruction of the  
61 spectrogram of speech and to detect the attended talker (e.g., Mesgarani  
62 and Chang, 2012; Zion Golumbic et al., 2013). In sum, selective neural  
63 processing in auditory cortices establishes an isolated and distraction-  
64 invariant spectro-temporal representation of the attended talker.

65 However, as has been shown, degradations of the acoustic signals  
66 attenuate the neural phase-locking to speech. Experimental degradations  
67 have included artificial transformations of temporal fine structure (Ding et  
68 al., 2014; Kong et al., 2015), or rhythmicity (Kayser et al., 2015), reverberation  
69 (Fuglsang et al., 2017) or decreased signal-to-noise ratio (SNR; Kong et al.,  
70 2014; Ding and Simon, 2013; Giordano et al., 2017). Not least, neural  
71 selection of speech appears weakened in people with hearing loss  
72 (Petersen et al., 2016). In sum, those studies suggest that the strength of  
73 neural phase-locking indicates behavioral performance such as speech  
74 comprehension.

75 Additionally, higher order non-auditory neural mechanisms facilitate  
76 speech comprehension as well. The supra-modal, fronto-parietal attention  
77 network is a candidate to be involved in top-down selective neural

78 processing during demanding listening tasks (Woolgar et al., 2016). Beyond  
79 the phase-locking in lower frequency bands (i.e., ~1 – 8 Hz; Wang et al 2018,  
80 Pomper and Chait 2017), top-down selective neural processing has also  
81 been associated with changes in the power of induced alpha-oscillations  
82 (i.e., ~8 – 12 Hz; Obleser and Weisz 2012; Kayser et al. 2015, Wöstmann et al.  
83 2016). Specifically, increased parietal alpha-power is related to enhanced  
84 suppression of the distracting input (Wöstmann et al., 2017). This reflects  
85 that, besides the neural spectro-temporal enhancement of the attended  
86 talker, a crucial role in top-down neural selective processing was attributed  
87 to the suppression of the ignored talker.

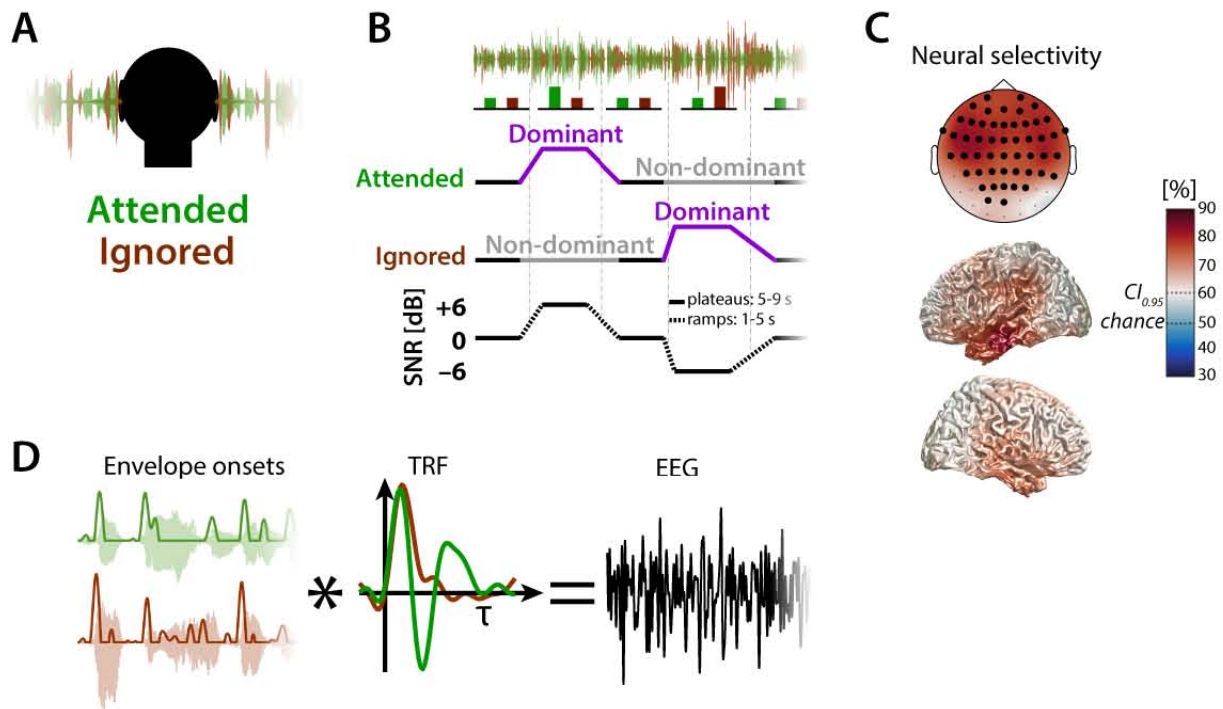
88 Neural signatures of suppression can be two-fold. First, suppression can  
89 attenuate the neural response to an ignored talker compared to an  
90 attended talker, like it was found in neural phase-locking from latencies of  
91 around 100 ms (Ding and Simon, 2012; Wang et al., 2018). Second, active  
92 suppression can add or increase components in the neural response to the  
93 ignored talker, given that the response is dissociable from the response to  
94 the attended talker (e.g.; a louder ignored talker evoking a stronger neural  
95 response anti-polar to the response to a louder attended talker). Here we  
96 asked, how the components of the phase-locked neural response are  
97 affected by selective attention under varying signal-to-noise ratio (SNR).

98 The phase-locked neural response to broad-band continuous speech can  
99 be obtained from EEG by estimating the (delayed) covariance of the  
100 temporal speech envelope and the EEG, which results in a linear model of  
101 the cortical response; a temporal response function (TRF; Lalor et al., 2009;  
102 Crosse et al., 2016). Analogous to the event-related potential (ERP), the  
103 components of the TRF can be interpreted as reflecting a sequence of  
104 neural processing stages where later components reflect higher order  
105 processes within the hierarchy of the auditory system (Davis and  
106 Johnsrude, 2003; Picton et al., 2013; Di Liberto et al., 2015).

107 Here, we use a listening scenario in which two concurrent talkers undergo  
108 continuous SNR variation. Our results demonstrate differential effects of  
109 bottom-up acoustics vs. top-down selective neural processing on earlier vs.  
110 later neural response components, respectively. Source localization reveals

111 that not only auditory cortex regions are involved in the selective neural  
112 processing of concurrent speech, but that a fronto-parietal attention  
113 network contributes to selective neural processing through late  
114 suppression of the ignored talker.

115



**Figure 1: Experimental design, forward model, and neural selectivity.** **A)** Two mixed talkers (female & male) were presented on both ears without spatial segregation (diotic). **B)** The signal-to-noise ratio (SNR) between attended (signal) and ignored (noise) talker was varied between -6, 0 and +6 dB by either raising the level of the attended talker or the ignored talker. Length of ramps and plateaus were drawn from uniform distributions. **C)** Neural selectivity here expressed as classification accuracy in detection of the attended and ignored talker averaged across subjects. Shown here is accuracy as obtained by prediction of EEG signals (Fiedler et al., 2017) at single EEG channels and single voxels in source space, respectively. Highlighted channels of topographic maps indicate that the lower bound of the confidence interval (bootstrapped mean on the group level) was greater than the 95%-confidence bound of a binomial distribution ( $CI_{0.95} = 60\%$ ). **D)** Temporal response functions (TRF) to the attended and ignored talker were extracted by a forward (encoding) regression model based on the assumption that the measured EEG signal is the superposition (convolution) of the envelope onsets (of the attended and ignored talkers) and the TRFs, respectively. TRFs reflect the neural response evoked by a single envelope onset.

116

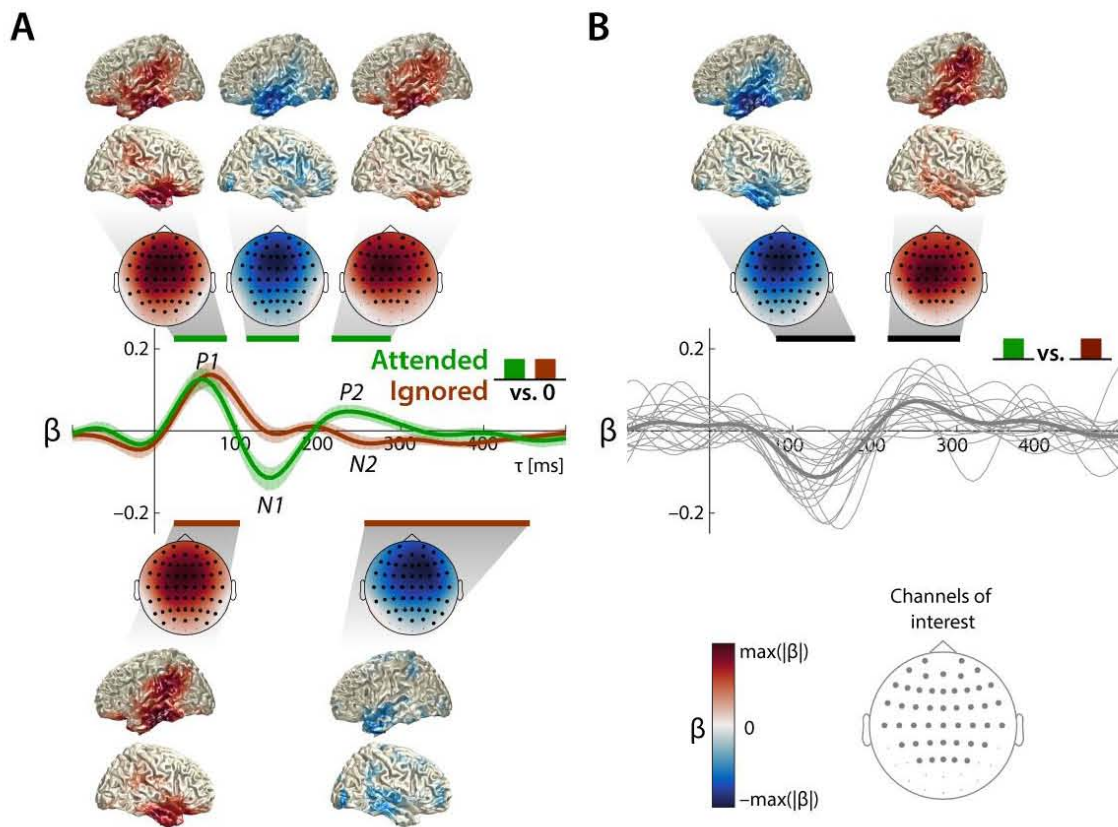
## 117 **Results**

118 We asked participants to listen to one of two simultaneously presented  
119 audiobooks under varying signal-to-noise ratio (Fig. 1A & B; -6 to +6 dB  
120 SNR). After each of twelve five-minute blocks, subjects were asked to rate  
121 the difficulty of listening to the to-be-attended talker on a color bar ranging  
122 from red (difficult = 1) to green (easy = 10). The average difficulty ratings  
123 strongly varied between subjects (mean: 5.2, SD: 2.2, range: 2.3–8.9). No  
124 difference in difficulty ratings for listening to the female versus the male  
125 talker was found (one-sample t-test,  $t_{17} = 1.17$ ,  $p = 0.26$ ).

126 To test their successful attending, participants were asked to answer four  
127 multiple-choice questions on the content of the to-be-attended audiobook  
128 after each five-minute block. The percentage of correctly answered  
129 questions was far above chance (25%) for all participants (mean: 81%, SEM:  
130 2%, range: 60–96%). All participants were thus able to follow the to-be-  
131 attended talker.

## 132 **Neural selectivity**

133 To obtain a general estimate of which EEG channels and which voxels  
134 reveal signatures of *neural selectivity*, we identified the attended (and the  
135 ignored) talker by forward prediction of EEG signals based on one-minute  
136 parts of the EEG and envelope onsets (see methods). Overall *neural*  
137 *selectivity* was highest (up to 80%) at fronto-central electrodes and  
138 respective temporal cortex regions in source-space (Fig. 1C).



**Figure 2: Temporal response functions (TRF) to continuous speech of concurrent talkers under balanced SNR (0 dB).** TRF  $\beta$ -weights depict average across subjects and average across channels of interest. Confidence bands (95%) were obtained by bootstrapping the mean across subjects. Horizontal lines indicate time ranges of significant difference from zero obtained from a cluster-based permutation test at the group level. Topographic maps show  $\beta$ -weights of clusters averaged across the cluster time range. Highlighted channels are part of the significant clusters. Source localizations show the 20% most strongly contributing voxels. **A)** Response to the attended talker (green, upper topographic maps) clearly show a cascade of three components (P1<sub>TRF</sub>-N1<sub>TRF</sub>-P2<sub>TRF</sub>). Response to the ignored talker (red, lower topographic maps) only show a P1<sub>TRF</sub>, whereas the N1<sub>TRF</sub> and P2<sub>TRF</sub> are suppressed. **B)** Significant differences between neural responses to the attended and ignored talker are present in the N1<sub>TRF</sub> and P2<sub>TRF</sub>-time range. Thin grey lines show single subject TRFs averaged across channels of interest.

### 139 Attention modulates neural responses to concurrent speech

140 Next, we assessed in greater detail the unfolding of attentional selection of  
 141 to-be-attended speech in time. To this end, we estimated the TRFs from the  
 142 balanced SNR trials of 0 dB (i.e. independent of the SNR manipulation) and  
 143 assessed the most prominent response components and their modulation  
 144 by attention. We inspected both the TRFs to the attended and ignored  
 145 talker individually (Fig. 2A), as well as the difference between the TRFs to  
 146 the attended and ignored talker (Fig. 2B) to examine signatures of *neural*  
 147 *selectivity*.



148 First, an early positive component (termed  $P1_{TRF}$ ) appeared in the TRFs to  
149 the attended (Fig. 2A, 24–88 ms,  $p = 2 \times 10^{-4}$ ) and ignored (Fig. 2A, 24–112  
150 ms,  $p = 2 \times 10^{-4}$ ) talkers, but without any attention-related difference (Fig.  
151 2B). Latency, polarity, and topography of this component compared well to  
152 a P1 as found in auditory evoked potentials (AEPs).

153 Second, a later negative deflection (termed  $N1_{TRF}$ ) was only present in the  
154 TRF to the attended talker (Fig. 2A; 112–176 ms,  $p = 5 \times 10^{-4}$ ). This  
155 component was significantly increased in magnitude (i.e., more negative)  
156 for the attended versus the ignored talker (Fig. 2B, 80–176 ms,  $p = 5 \times 10^{-4}$ ;  
157 see also Fig. S3). Noteworthy, the significant attentional modulation of this  
158 component (attended–ignored) started already at a time lag of 80 ms,  
159 when both the TRF to the attended and to the ignored talkers were still in  
160 positive deflection (see Fig. 2A).

161 Third, a positive deflection between 200 and 300 ms (termed  $P2_{TRF}$ ; Fig. 2A,  
162 216–304 ms,  $p = 5 \times 10^{-4}$ ), was again only present in the TRF to the attended  
163 talker. This component mainly drove the significant difference between the  
164 responses to the attended and ignored talker (Fig. 2B,  $p = 2 \times 10^{-4}$ ).

165 Interestingly, in the same time interval, a negative deflection was found in  
166 the TRF to the ignored talker (termed  $N2_{TRF}$ ; Fig. 2B, 248–424 ms,  $p = 2 \times 10^{-4}$ ).  
167 While at earlier stages, TRFs to the attended and the ignored talker showed  
168 the same polarity ( $P1_{TRF}$ ), at the stage of the  $P2_{TRF}$  we see an anti-polar  
169 relationship. Effectively, this also enhanced the late, attended–ignored  
170 difference in the  $P2_{TRF}$  time range (Fig. 2B).

171 In sum, three prominent components ( $P1_{TRF}$ ,  $N1_{TRF}$ ,  $P2_{TRF}$ ; Fig. 2A) were  
172 identifiable with notable consistency across individual subjects. The latter  
173 two components were absent in the TRF to the ignored talker and thus  
174 indicated *neural selectivity*. All three components ( $P1_{TRF}$ ,  $N1_{TRF}$ ,  $P2_{TRF}$ ) mainly  
175 localized to superior and inferior temporal regions (Fig. 2A). Note that the  
176 source localizations of the two latter components ( $N1_{TRF}$ ,  $P2_{TRF}$ ) compared

177 well to the sources of enhanced neural selectivity between attended and  
178 un-attended talkers (Fig. 1C).

### 179 **Late representation of ignored talker enhances towards more** 180 **detrimental SNRs**

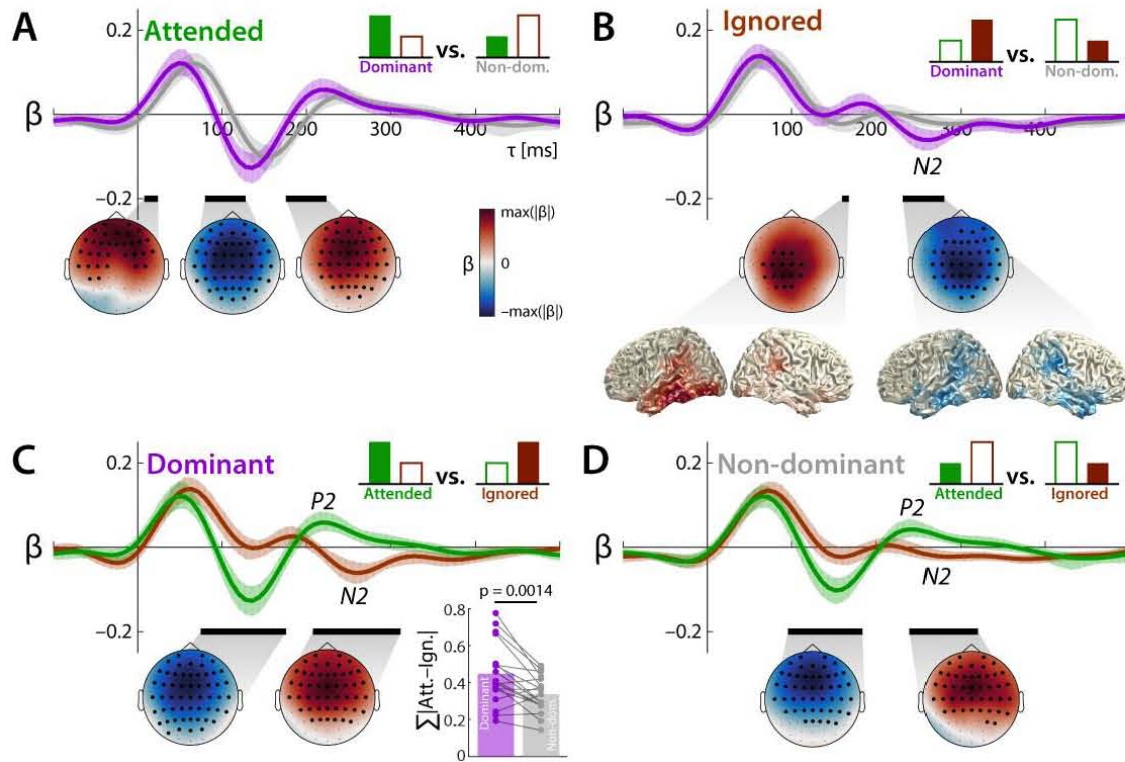
181 Next, we analyzed the impact of a varying SNR on the Temporal response  
182 functions (TRFs). To this end, we first contrasted the TRFs of the two  
183 extreme conditions (SNRs  $-6$  vs.  $+6$  dB; Fig. 3A&B). Second, we contrasted  
184 TRFs across SNRs matched for the acoustic properties of being either the  
185 louder or the quieter talker (Fig. 3C&D), such that the occurring differences  
186 between the TRFs to the attended and the ignored talker can solely be  
187 related to top-down attending versus ignoring. For simplicity, we will use  
188 the terms *dominant* (attended talker under  $+6$  dB SNR, ignored talker under  
189  $-6$  dB SNR) and *non-dominant* (attended talker under  $-6$  dB SNR, ignored  
190 talker under  $+6$  dB SNR). We observed an SNR-dependent latency shift  
191 which hindered time-lag-wise attended–ignored contrasts within SNRs  
192 (Fig. 3A&B, see appendix for more details).

193 Importantly, two later additional components appeared whenever the  
194 ignored talker was dominant (Fig. 3B): the first (160–178 ms,  $p = 0.04$ )  
195 localized to temporal regions, while the second extended markedly into  
196 parietal regions (232–280 ms,  $p = 0.001$ ). The enhanced involvement of  
197 parietal regions differentiated this detrimental-SNR, ignored-speech  
198 component from all others. Visual inspection of the TRFs to dominant  
199 talkers (Fig. 3C) highlights the additional late N2 component in the TRF to  
200 the ignored talker, which appears to be anti-polar to the  $P2_{\text{TRF}}$  to the  
201 attended talker.

202 In contrast, TRFs to *non-dominant* talkers (Fig. 3D) suggest that the  
203 observed attention-related differences are decreased (cf., Fig. 3C) due to  
204 smaller deflections of the  $N1_{\text{TRF}}$  and  $P2_{\text{TRF}}$  to the *non-dominant* attended  
205 talker and the lack of the anti-polar  $N2_{\text{TRF}}$  to the *non-dominant* ignored  
206 talker. We summed the magnitude of the attended–ignored difference  
207 across all time lags, which revealed a smaller attended–ignored difference  
208 for *non-dominant* versus *dominant* talkers ( $t_{17} = 3.80$ ,  $p = 0.0014$ ). Thus, the  
209 neural response to a *dominant* ignored talker does not resemble the neural

210 response to a dominant attended talker by capturing bottom-up attention.  
211 Instead, dominant ignored speech retains a distinct “ignored” neural  
212 signature, most likely to due to top-down neural signaling of its to-be-  
213 ignored status.

214 In sum, our findings indicate that, when a talker is *dominant*, neural  
215 signatures of selective processing are enhanced (compared to *non-*  
216 *dominant*). Importantly, this enhancement is not only affecting the  
217 representation of the attended talker, but an important contribution to this  
218 enhanced top-down processing can be attributed to an additional late  
219 component ( $N2_{TRF}$ ) in the neural response to the ignored talker. To further  
220 disentangle the contribution of the selective processing of the attended  
221 and ignored talker, we established the time lag and talker resolved  
222 measures *neural tracking* and *neural selectivity*, which will be discussed in  
223 the following section.



**Figure 3: Temporal response functions (TRF) to continuous speech of concurrent talkers contrasted as dominant vs. non-dominant talkers and attended vs. ignored talkers, respectively.** TRF  $\beta$ -weights depict average across ( $N = 18$ ) subjects and average across channels of interest. Confidence bands (95%) were obtained by bootstrapping the mean across subjects. Schematic bar graphs indicate the investigated contrast. Black horizontal lines indicate time ranges of significant difference obtained from a cluster-based permutation test at the group level. Topographic maps show  $\beta$ -weight differences of clusters averaged across the cluster time range. Highlighted channels are part of the significant clusters. Source localizations show the 20% most strongly contributing voxels with full opacity. **A)** Responses to the non-dominant attended talker are delayed compared to the dominant attended talker. **B)** A late component appeared in the response to the dominant ignored talker, which involved parietal regions. **C)** Late negative response ( $N2_{TRF}$ ) to the dominant ignored talker appears anti-polar to the response to the dominant attended talker. Inset: Magnitude of the attended-ignored TRF difference summed across all time lags for *dominant* and *non-dominant* talkers. **D)** Non-dominant talkers show significant but decreased attention-related differences.

224 **Neural selectivity increases by way of a late cortical representation of**  
 225 **ignored speech**

226 We established two measures to quantify the encoding and the selective  
 227 neural processing of the talkers during the unfolding of the neural response  
 228 reflected in the TRFs. First, *neural tracking* is a measure of how strongly a  
 229 single talker is represented (i.e., encoded) in the EEG. Second, *neural*  
 230 *selectivity* quantifies how accurately an attended talker can be identified as  
 231 attended and an ignored talker as ignored, respectively.

232 Parallel inspection of *neural tracking* and *neural selectivity* allowed us to  
233 disentangle the effects of bottom-up and top-down attention on the TRFs.  
234 For example, the increased sound pressure level of a talker may increase its  
235 saliency and thus bottom-up pull attention towards it. This would result in  
236 enhanced *neural tracking* of the ignored talker and the neural response  
237 would become less distinct from the respective response to a *dominant*, but  
238 intentionally attended talker. However, if there exists a counter-acting, top-  
239 down process that enhances and maintains a neural-response  
240 differentiation between the attended and the ignored talker, *neural*  
241 *selectivity* would increase at the same time.

242 To get a total estimate of *neural tracking* of the two talkers, we first used all  
243 time lags of the TRFs (i.e., -100-500 ms). Fig 4A shows the *neural tracking* of  
244 the attended, the ignored as well as the overall tracking of the two talkers  
245 (attended & ignored). The overall tracking was found to be well above zero  
246 for all participants as well as the tracking of the two talkers separately (Fig.  
247 4A, bottom).

248 In a next step, we estimated the time-lag- and channel-dependent  
249 unfolding of *neural tracking*. As expected, we found enhanced *neural*  
250 *tracking* of the attended talker compared to the ignored talker between 144  
251 and 288 ms under the balanced SNR of 0 dB (Fig. S1 A), driven by fronto-  
252 central channels. This is congruent with the time range and topography of  
253 the N1<sub>TRF</sub> and P2<sub>TRF</sub>, which were found to be non-present in the TRF to the  
254 ignored talker.

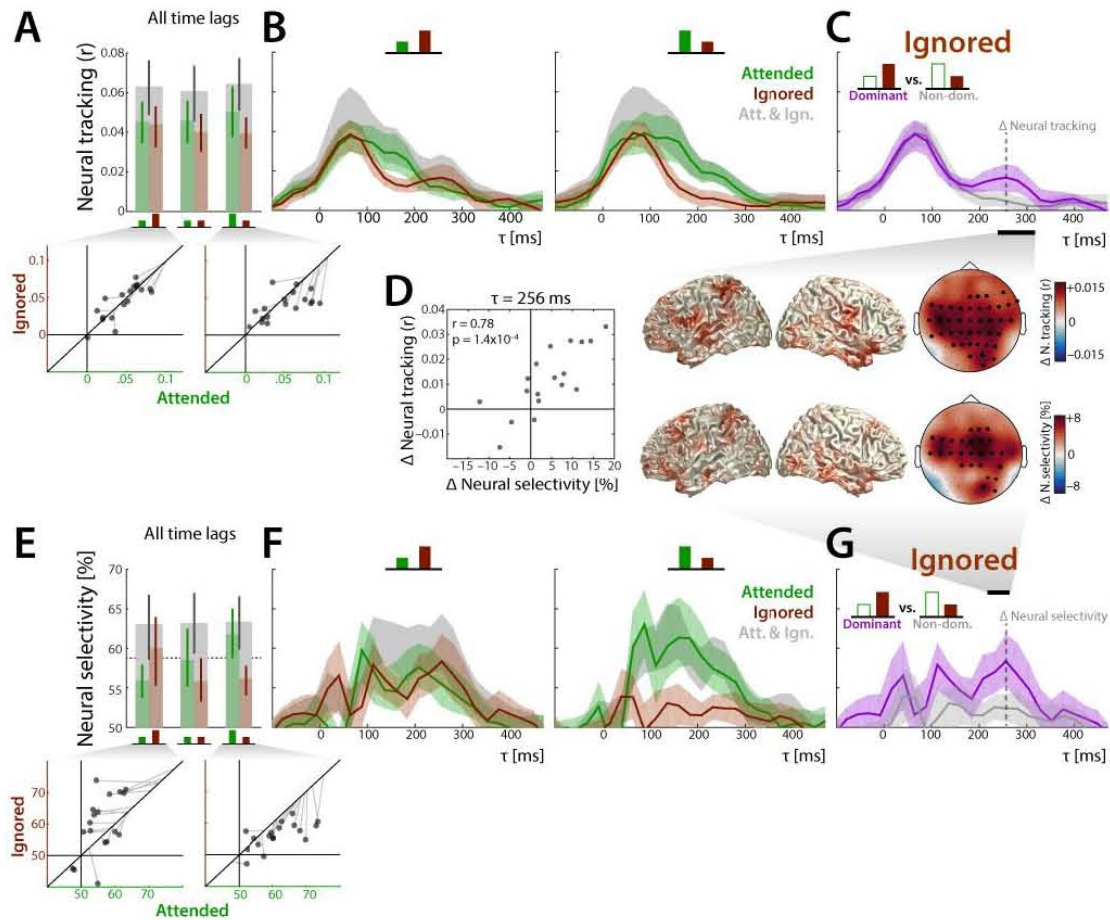
255 Interestingly, towards more adverse SNRs (dominant ignored talker), the  
256 late enhanced *neural tracking* of the attended talker compared to the  
257 ignored talker seems to shrink (Fig. 4B). Visual inspection of the time-lag  
258 resolved *neural tracking* suggests that this shrinkage is due to an additional  
259 late cortical representation of the ignored talker that appears whenever the  
260 ignored talker is *dominant*. The contrast of the *neural tracking* of the  
261 dominant and the non-dominant ignored talker confirmed such a late  
262 cortical representation (Fig. 4C, 240–312 ms,  $p = 1.5 \times 10^{-3}$ ) originating  
263 mainly from fronto-parietal as well as temporal regions.

264 Importantly, the overall *neural selectivity* is not affected by adverse  
265 conditions (Fig. 4E, grey bars,  $-6$  vs  $+6$  dB, one-sample t-test,  $t_{17} = 0.24$ ,  $p =$   
266  $0.81$ ). However, the relative contribution of the neural selectivity of the  
267 attended talker and ignored talker changes across SNRs ( $-6$  vs  $+6$  dB; one-  
268 sample t-test; attended:  $t_{17} = -4.6$ ,  $p = 2.77 \times 10^{-4}$ ; ignored:  $t_{17} = 2.18$ ,  $p =$   
269  $0.044$ ): Towards more adverse SNRs, the *neural selectivity* of the ignored  
270 talker increases, while the *neural selectivity* of the attended talker decreases  
271 (Fig. 4E, top). This is also discernible in single subjects (Fig. 4E, bottom),  
272 where *neural selectivity* of the attended talker is stronger under an SNR of  
273  $+6$  dB (right, 16 of 18 subjects) and stronger for the ignored talker under an  
274 SNR of  $-6$  dB (left, 11 of 18 subjects).

275 If the increased *neural tracking* of the *dominant* ignored talker at later stages  
276 (Fig. 4C) is solely driven by its increased saliency (i.e., higher dominance  
277 evoking a stronger response), we would expect no concomitant increase in  
278 neural selectivity (see above). However, we found a late increase in *neural*  
279 *selectivity* for the dominant compared to the *non-dominant* ignored talker  
280 (Fig. 4G, 216–264 ms,  $2.5 \times 10^{-3}$ ). Neural sources compared well to the  
281 increased fronto-parietal *neural tracking* of the dominant ignored talker  
282 (see Fig. 4C&G).

283 Furthermore, *neural tracking* and *neural selectivity* (for *dominant* vs *non-*  
284 *dominant* ignored speech) were positively correlated (Fig. 4D,  $r = 0.78$ ,  $p =$   
285  $0.014 \times 10^{-2}$ ): If a listener's neural tracking was relatively strong for the  
286 dominant versus non-dominant ignored talker, the neural response  
287 allowed more accurate identification of the ignored talker as ignored.

288 In sum, at later stages, not only increased selective neural processing of the  
289 attended talker but also the selective neural processing of the ignored  
290 talker facilitates input segregation under adverse listening conditions.



**Figure 4: Unfolding of neural tracking and neural selectivity reveals late neural selective processing of the ignored talker.** Neural tracking and neural selectivity were estimated based on the extracted TRFs to the attended (green), the ignored (red), as well as both talkers (grey). Confidence bands (95%) were obtained by bootstrapping. Highlighted channels (topographic maps) are part of a significant cluster. Source localizations show the 20% most strongly contributing voxels with full opacity. **A)** Neural tracking across all time lags (-100-500 ms). Scatterplots (bottom) show single-subject data averaged across channels of interest. Grey lines indicate overall neural tracking of both talkers at the 45°-line. **B)** Unfolding of neural tracking across time lags under SNR of -6 (left) and +6 dB (right). **C)** Contrast of neural tracking between the dominant and non-dominant ignored talker. **D)** Correlation of change in neural tracking and change of neural selectivity at  $\tau = 256$  ms. **E)** Neural selectivity across all time lags (-100-500 ms). Scatterplots (bottom) show single-subject data averaged across channels of interest. Grey lines indicate overall neural tracking of both talkers at the 45°-line. **F)** Unfolding of neural selectivity across time lags under SNR of -6 (left) and +6 dB (right). **G)** Contrast of neural selectivity between the dominant and non-dominant ignored talker.

291

292

## 293 **Discussion**

294 In the present study, human listeners attended to one of two concurrent  
295 talkers under continuously varying signal-to-noise ratio (SNR). We asked to  
296 what extent a late cortical representation (i.e., neural tracking) of the  
297 ignored acoustic signal is key to the successful separation of to-be-  
298 attended and distracting talkers (i.e., neural selectivity) under such  
299 demanding listening conditions.

300 Forward modeling of the EEG response revealed neural responses to the  
301 temporal envelopes of individual talkers and their modulation by both, top-  
302 down attentional set, and bottom-up SNR. Critically, towards more adverse  
303 SNRs, an additional late negative component occurred in the neural  
304 response to the ignored talker. Under adverse conditions, this component  
305 was found to be accompanied by enhanced selective neural processing  
306 (*neural selectivity*), emerging primarily from fronto-parietal brain regions.

307 The present result suggests that irrelevant, to-be-ignored acoustic inputs  
308 are not simply absent from the late cortical response but become actively  
309 suppressed in regions beyond auditory cortex.

### 310 **Early and late neural signatures of selective neural processing**

311 Generally, we replicated previous results that showed that attention-  
312 ignored differences in the neural response can mainly be found at time lags  
313 > 80 ms, which were mainly attributed to stronger neural tracking caused  
314 by enhanced N1 and P2 components in the response to the attended vs.  
315 ignored talker (Horton et al. 2013; O'Sullivan et al., 2014; Ding & Simon,  
316 2012). Here we show that a P2-counter-acting response to the ignored  
317 talker enhances the attended-ignored difference as well.

318 While earlier studies showed that selective neural processing in auditory  
319 cortices is mainly working out a clean representation of the attended talker  
320 (Mesgarani and Chang, 2012; Zion Golumbic, 2013), we show that a late  
321 neural representation of a distracting auditory input is accompanied with  
322 enhanced selective neural processing in a cocktail-party scenario as well.  
323 This additional late neural representation was revealed by going beyond



324 strictly matched sound pressure levels of attended versus ignored speech  
325 (cf., Horton et al., 2013; O’Sullivan et al., 2014, Ding & Simon, 2012; Mirkovic  
326 et al., 2015; Biesmanns et al., 2016), by presenting speech signals both as  
327 target and distractor (cf., Ding & Simon, 2013) and by applying SNR-  
328 variation symmetrically around 0 dB (cf., Kong et al., 2014). In sum, our  
329 design allowed us to draw conclusions on the neural selective processing  
330 of real-world listening scenarios of dynamically varying listening demand.

331 Our investigation of concurrent speech under varying SNR helps  
332 disentangle neural mechanisms of early and late selection (Treisman 1964).  
333 Since the ignored talker predominantly masks the attended talker under  
334 adverse listening conditions (i.e., negative SNRs, which we have labelled  
335 *dominant*), early neural filters tuned to the spectro-temporal properties of  
336 the attended talker might not be sufficient (i.e., neural gain, Willmore et al.,  
337 2014).

338 Thus, a later filter on the ignored signal must actively suppress distracting  
339 inputs. We found such a neural filter mechanism (Fig 4C&G) active in a time  
340 range which was previously attributed to processing of phonological (Di  
341 Liberto et al. 2016, Brodbeck et al. 2018) as well as semantic features  
342 (Broderick et al. 2018), which both go beyond basic acoustic properties of  
343 speech (Obleser and Eisner 2009). One suggestion of our results is that  
344 when phonemes (or even words) of the dominant ignored talker pull  
345 bottom-up attention, their representation is actively suppressed at a late  
346 stage in order not to impair linguistic representation of the attended  
347 talker’s speech.

#### 348 **Late distractor suppression in a non-auditory, fronto-parietal** 349 **attention network**

350 Previously, it has been shown that neural selective processing of concurring  
351 auditory stimuli is mainly accomplished in auditory cortex, resulting in a  
352 ‘clean’ and distraction-invariant representation of the attended talker  
353 (Mesgarani and Chang 2012; Zion Golumbic 2013).

354 Critically, under the adverse SNR of  $-6$  dB, our analysis revealed an  
355 enhanced response to the ignored talker in a later time range (i.e., 200–300

356 ms) consisting of a positive and a negative component (Fig. 3B). The latter  
357 is anti-polar to the P2<sub>TRF</sub>(to the attended talker). This additional component,  
358 which we interpret as a signature of active suppression of the ignored  
359 talker, involved non-auditory regions, which are part of the fronto-parietal  
360 attention or global-demand network (Woolgar et al., 2016), where we  
361 found enhanced neural selective processing of the ignored talker.

362 Under the assumption that such active suppression is costly to the  
363 cognitive system, it has been suggested that it is only deployed if necessary  
364 (Chait et al., 2010). Neural signatures for active suppression of irrelevant  
365 signals during late (~200 ms) AEPs have been examined before (Melara et  
366 al., 2002; Chait et al., 2010). Pomper and Chait (2017) related enhanced  
367 centro-parietal activity in the theta band (4–7 Hz) to enhanced top-down  
368 control. Parietal activity in the theta-band was also found to be inversely  
369 related to the delta-band auditory entrainment in superior temporal gyrus  
370 (Keitel et al., 2017). Here we show how late top-down, fronto-parietal neural  
371 processing of the distracting auditory input is unfolding in time and might  
372 facilitate overall selective neural processing.

373 In earlier studies, researchers highlighted the predominant tracking of the  
374 attended talker (Mesgarani and Chang, 2012; Ding & Simon, 2012, Zion  
375 Golumbic, 2013, O’Sullivan et al. 2014), emphasizing that a clean  
376 representation of the attended talker is key to successful listening. In some  
377 contrast to this, previous results shed light on the neural processing of the  
378 ignored talker (see also Wöstmann et al., 2017, Olguin et al., 2018). We have  
379 shown here that the overall neural selective processing is surprisingly  
380 robust against such demanding listening conditions (Ding & Simon, 2013),  
381 and that a ‘clean’ or isolated tracking of the ignored talker is at least as  
382 essential.

383 This finding invites some speculation on the neural implementation of  
384 attentional filters more generally. On the one hand, a selective neural filter  
385 can be solely optimized to let pass relevant features of attended signals. On  
386 the other hand, it can be optimized to let pass features of the ignored talker,  
387 which might be relevant for suppression at a later stage. In line with earlier  
388 studies, we found that the *neural tracking* was dominated by the attended

389 talker (speaking for the first strategy). However, under most demanding  
390 listening conditions (i.e., negative SNR), *neural selectivity* was dominated by  
391 the ignored talker.

392 Neural filter mechanisms might thus adapt depending on the listening  
393 demand. Follow-up studies should investigate the relationship of such filter  
394 adaptation to the concept of listening effort (Rönnberg et al., 2013;  
395 McGarrigle et al., 2014): Additional tracking of the ignored talker leads to  
396 higher neuro-computational load and might also be related to working  
397 memory performance (Rudner et al. 2011).

398 Within our design, we can only draw limited conclusions on the behavioral  
399 impact of the late neural tracking of the ignored talker. This is due to the  
400 tradeoff between sufficient behavioral data (e.g., trial-based design) and  
401 ecological validity (e.g., presentation of continuous speech; Hamilton and  
402 Huth, 2018). Following studies should acquire more fine-grained  
403 behavioral data, ideally without losing much of the ecological validity.

404 Our results show that, within the hierarchy of the central auditory  
405 pathways, the cocktail-party problem might look solved or settled at the  
406 stage of secondary auditory cortex (Mesgarani & Chang, 2012), but higher-  
407 order, attentional networks and their dedicated processing of distracting  
408 speech appear key to this solution.

## 409 **Conclusions**

410 The present data show how components of the unfolding temporal  
411 response function as identified in a forward encoding model of the  
412 electroencephalographic signal can reflect distinct neural stages of  
413 attentional filtering. These stages contain the initial, attention-  
414 independent encoding of acoustic signals; the extraction and amplification  
415 of relevant features; and lastly a robust, purely attention-driven selective  
416 response to the attended and ignored acoustic signals.

417 Most consequential to our thinking about attentional filtering in the central  
418 auditory system, an active-suppression response to ignored acoustic  
419 signals originates from non-auditory, fronto-parietal attentional networks.

420 In sum, with a design closer to real-life listening scenarios, our study  
421 provides insight into how selective neural processing of attended speech  
422 unfolds and is upheld not only by auditory cortices. Instead, establishing a  
423 clean cortical representation of the attended talker as suggested previously  
424 hinges on achieving a late suppression of ignored signals, with  
425 contributions by regions of the fronto-parietal attention network.

## 426 **Methods**

### 427 **Participants**

428 Eighteen native speakers of German (9 females) were invited from the  
429 participant database of the Department of Psychology, University of  
430 Lübeck, Germany. We recruited participants who were between 23 and 68  
431 years old at the time of testing (mean: 49, SD: 17), to allow valid conclusions  
432 from such a challenging listening scenario to middle-aged and older adults.  
433 All reported normal hearing and no histories of neurological disorders.  
434 Incomplete data due to recording hardware failure was obtained in four  
435 more, initially invited participants. All participants gave informed consent  
436 and received payment of 8 €/hour. The study was approved by the local  
437 ethics committee of the University of Lübeck.

### 438 **Stimuli**

439 The goal of this study was to investigate the selective neural processing of  
440 one of two talkers under a continuously varying signal-to-noise ratio (SNR).  
441 Here, the signal is a to-be-attended talker and the noise is a to-be-ignored  
442 talker. Our study was conducted in a within subject 2 by 3 design (attention  
443 by SNR (three levels)).

444 We selected two audiobooks read by native German speakers, one female  
445 (Elke Heidenreich, 'Nero Corleone kehrt zurück', read by Elke Heidenreich)  
446 and one male (Yuval Noah Harari, 'Eine kurze Geschichte der Menschheit',  
447 read by Jürgen Holdorf). The following steps of stimulus preparation were  
448 done using custom code written in MATLAB (Version 2017a; *Mathworks Inc.*,  
449 *Natick, MA*). Sequences of silence longer than 500 ms were truncated to 500  
450 ms to avoid long parts of silence (O'Sullivan et al., 2014). The first hour of  
451 each audiobook was selected for further preparation. The first 30 minutes  
452 of each audiobook served as the to-be-attended and the rest served as the  
453 to-be-ignored speech, such that all subjects could attend both stories from

454 the beginning and attended (and ignored) both the female and the male  
455 voice the same amount of time.

456 The identical mixture of the attended and ignored talker was presented on  
457 both ears, resulting in a concurrent listening scenario without any spatial  
458 cue (i.e. diotic, Fig. 1A). Hence, the only cues available for talker segregation  
459 consisted in the spectro-temporal features of the talkers, such as pitch,  
460 formants, and amplitude modulation.

461 The SNR was modulated symmetrically around 0 dB. An SNR of 0 dB refers  
462 to concurrent talker signals with a matched long-term root-mean-square  
463 (rms) amplitude as used previously in numerous studies (e.g. Power et al.,  
464 2012; O'Sullivan et al., 2014; Mirkovic et al., 2015). Coming from an SNR of  
465 0dB, the SNR was either increased to +6 dB by raising the sound pressure  
466 level (SPL) of the to-be-attended talker by 6 dB or decreased to -6 dB by  
467 raising the SPL of the to-be-ignored talker by 6 dB. Thus, the talkers were  
468 either *balanced* (Fig. 1B, black) or one of the talkers was *dominant* (Fig. 1B,  
469 purple) and the other was *non-dominant* (Fig. 1B, grey). The particular SNR-  
470 range (-6 to +6 dB) was chosen to create a challenging but at the same time  
471 solvable listening task. Even if an SNR of -6 dB is rare in real-life listening  
472 scenarios (Smeds et al., 2015), the neural tracking of attended speech has  
473 been reported as intact at SNRs as low as -6 dB (Ding and Simon, 2013).  
474 However, speech perception (number of words repeated correctly) of  
475 normal hearing subjects starts to suffer around an SNR < 0 dB and the  
476 speech-reception threshold (i.e. 50% correct) usually lies between -5 and 0  
477 dB (Pichora-Fuller et al., 1995, Bentler et al., 2004).

478 As building blocks for SNR modulation, we created a sample of plateaus  
479 (i.e., constant SNR of -6, 0 or +6 dB) and ramps (i.e., transition between  
480 plateaus). The length of plateaus was uniformly distributed between 5 and  
481 9 seconds in discrete steps of one second. The ramps were linear  
482 interpolations between SNRs with the length uniformly distributed  
483 between 1 and 5 seconds in discrete steps of one second. The length  
484 distributions of plateaus and ramps were kept uniform within each talker  
485 and within their assignments as being attended or ignored. We  
486 concatenated plateaus via ramps such that a 0 dB plateau was either

487 followed by a +6 dB or a –6 dB, whereas a +6 dB or a –6 dB plateau were  
488 always followed by a 0 dB plateau via a respective ramp. Randomly varying  
489 SNR time courses were created for each subject individually in order to  
490 avoid systematic overlap between the SNR modulation and the  
491 audiobooks. Stimulus material was cut into twelve blocks, which resulted  
492 in an average block length of five minutes. Sound files were created with a  
493 sampling rate of 44.1 kHz and a 16-bit resolution. The experiment was  
494 implemented in the software *Presentation (Neurobehavioural Systems)*.  
495 Stimuli were presented via headphones (Sennheiser HD25).

#### 496 **Task**

497 The twelve blocks were presented such that subjects were instructed to  
498 attend to the female or to the male talker in an alternating fashion. After  
499 instruction before each block (i.e. attend to female or attend to male),  
500 subjects were asked to start the stimulus presentation by a button press,  
501 which enabled the participants to take a break between blocks. During  
502 listening, subjects were asked to fixate a cross presented on the screen in  
503 order to reduce eye movement.

504 Every other block, the stories picked up at the point it ended two blocks  
505 before. After each block, subjects were asked to rate the difficulty of  
506 maintaining attention by mouse-clicking on a continuous color bar ranging  
507 from red (difficult) to green (easy). For later analysis, the continuous color  
508 bar was discretized into ten segments (1 = difficult, 10 = easy).  
509 Subsequently, participants were asked to answer four multiple-choice  
510 questions concerning the content of the to-be-attended audiobook. The  
511 average rating of difficulty was neither significantly correlated with the  
512 number of questions correctly answered (Pearson's  $r = 0.1$ ,  $p = 0.73$ ), nor  
513 with participants' age (Pearson's  $r = -0.17$ ,  $p = 0.51$ ). Furthermore, we found  
514 no significant correlation of the number of correctly answered questions  
515 with age (Pearson's  $r = -0.11$ ,  $p = 0.65$ ).

#### 516 **Data acquisition and preprocessing**

517 EEG was recorded with 64 electrodes *Acticap (Easycap, Herrsching,*  
518 *Germany)* connected to an *ActiChamp (Brain Products, Gilching, Germany)*

519 amplifier. EEG signals were recorded with the software *BrainVision Recorder*  
520 (*Brain Products*) at a sampling rate of 1 kHz. Impedances were kept below  
521 10 k $\Omega$ . Electrode TP9 (left mastoid) served as reference during recording.

522 The EEG data were pre-processed in *MATLAB (2017a)* using both the  
523 *Fieldtrip*-toolbox (version: 20170321; Oostenveld et al., 2011) and custom  
524 written code. The EEG data were re-referenced to the average of the  
525 electrodes TP9 and TP10 (left and right mastoids) and resampled to  $f_s = 125$   
526 Hz. The continuous EEG data were highpass-filtered at  $f_c = 1$  Hz and  
527 lowpass-filtered at  $f_c = 30$  Hz (two-pass Hamming window FIR, filter order:  
528  $3f_s/f_c$ ).

529 From the continuous EEG data, we extracted the parts during which the  
530 twelve blocks of audiobooks were presented (see above). For every subject,  
531 we applied independent component analysis (ICA; Makeig et al., 2004) on  
532 the concatenated data of the twelve blocks and manually rejected  
533 components that were clearly related to eye movements, eye blinks,  
534 muscle artifacts, heartbeat as well as single-channel noise. On average, 26  
535 of 62 components (SD: 7.3) were rejected.

536 For further analysis, we lowpass-filtered the data again at  $f_c = 10$  Hz (two-  
537 pass Hamming window FIR, filter order:  $3f_s/f_c$ ), which assured that the  
538 amplitudes at all frequencies up to 8 Hz were not reduced. Previously,  
539 neural activity phase-locked to the envelope was only found up to a  
540 frequency of approximately 8 Hz (Zion Golumbic et al., 2013; Ding et al.,  
541 2014). We could confirm this finding by incrementally raising the cutoff  
542 frequency, which didn't change the morphology of the TRFs (see below)  
543 but only decreased the prediction accuracy due to the interference of non-  
544 phase-locked neural activity and external noise in higher frequencies.

#### 545 **Extraction of envelope onsets**

546 A temporal representation of the acoustic onsets, further called envelope  
547 onsets, was extracted from the presented speech signals (Fiedler et al.,  
548 2017). Those representations later served as regressors to model neural  
549 responses to the talkers (see below). First, we extracted an auditory  
550 spectrogram containing 128 spectrally resolved sub-band envelopes of the



551 speech signals logarithmically spaced between approximately 90 and 4000  
552 Hz using the *NSL* toolbox (Chi et al., 2005). Second, the auditory  
553 spectrogram was summed up across frequencies, which resulted in broad-  
554 band temporal envelopes of the audiobooks. Taking the derivative of the  
555 envelope and zeroing all values smaller than zero (Hertrich et al., 2012)  
556 returned the envelope onsets, which only contain positive values at time  
557 periods of an increasing envelope, as can be found at acoustic onsets (Fig.  
558 1C).

559 Using the envelope onsets as regressor does not imply that we only  
560 modeled the encoding of acoustic onsets. Every onset is followed by a peak  
561 in the speech envelope (Fig. 1C), which is then again followed by an offset  
562 and the next onset and so forth, resulting in a high autocorrelation between  
563 those features. Nevertheless, onsets are the earliest feature that could  
564 possibly evoke a neural response (Picton, 2013). The latency of modeled  
565 responses to envelope onsets (compared to envelopes) was found to be  
566 most similar to conventional ERPs (Fiedler et al., 2017, supplemental  
567 material).

## 568 **Estimation of temporal response functions**

569 We applied an established method to estimate a linear forward (encoding)  
570 model (Lalor et al., 2009; Crosse et al., 2016). The model contains temporal  
571 response functions (TRFs), which are estimations of the neural response to  
572 a continuously varying stimulus feature. In our case, this stimulus feature is  
573 the envelope onsets (see above) of both, the attended and the ignored  
574 talker. Based on the assumption that every sample in the EEG signal  $r(t)$  is  
575 the superposition of neural responses to past onsets and thus can be  
576 expressed for one talker by a convolution operation:

$$r(t) = s * TRF = \sum_{\tau} [s(t - \tau) \cdot TRF(\tau)] \quad (1)$$

577 where  $s(t)$  is the envelope onsets, TRF is the temporal response function  
578 that describes the relationship between  $s$  and  $r$  over a range of time lags  $\tau$   
579 (Fig. 1C). The TRF contains a weight for each time lag  $\tau$ . We investigated  
580 time lags in the range from -100 to 500 ms. In order to obtain the  $\beta$ -weights

581 of the TRF to both talkers contained in the matrix  $G_{TRF}$ , ridge regression  
582 (Hoerl and Kennard, 1970) was applied, which can be expressed in the linear  
583 algebraic form:

$$G_{TRF} = (S^T S + \lambda m I)^{-1} S^T R \quad (2)$$

584 where  $S$  is matrix containing the onset envelopes of both the attended and  
585 ignored talker and its sample-wise time lagged replications,  $R$  contains the  
586 measured EEG signal,  $\lambda$  is the ridge parameter for regularization, the scalar  
587  $m$  is the mean of the trace of  $S^T S$  (Biesmans et al., 2016) and  $I$  is the identity  
588 matrix. The optimal ridge parameter  $\lambda$  was estimated according to Fiedler  
589 et al. (2017) and was set to  $\lambda = 10$ .

590 TRFs were estimated on a trial-by-trial basis, where trial refers to a part (e.g.  
591 a plateau of +6 dB) of certain length cut from the continuous stimulus and  
592 the respective EEG data. For the subsequent analysis, we subdivided the  
593 data in two ways: First, to get a general estimate of the model's ability to  
594 dissociate between attended and ignored talkers, we cut the data into one-  
595 minute trials, resulting in trial lengths comparable to previous studies  
596 (O'Sullivan et al., 2014; Mirkovic et al., 2015; Biesmans et al., 2016; Fiedler et  
597 al., 2017). This resulted in 60 trials per subject. Second, we cut the data  
598 based on the applied SNR modulation, which resulted in three groups of  
599 trials: -6 dB, 0 dB and +6 dB. To use the entire recording, the data were cut  
600 at the time points where ramps of the SNR time courses either crossed -3  
601 dB or +3 dB (Fig. 1B). This resulted in 180 trials of 0 dB and 90 trials of -6 and  
602 +6 dB, respectively. The average length of those trials was 10 seconds (i.e.  
603 average length of a plateau (7 seconds) and average length of two halves  
604 of a ramp (2x1.5 seconds)). In order to balance the number of trials across  
605 SNRs, 90 trials from 0 dB were randomly drawn from the 180 trials of every  
606 subject. During the analysis, we contrasted TRFs not only within conditions,  
607 but also contrasted the TRFs to the talkers within their role of being  
608 dominant (Fig 2B, purple; attended under SNR = +6 dB, ignored under SNR  
609 = -6 dB) or non-dominant (Fig 2B, grey; attended under SNR = -6 dB,  
610 ignored under SNR = +6 dB). We will use those terms and schematic bar  
611 graphs (Fig. 1B) throughout the entire article.

## 612 **Statistical analysis on temporal response functions**

613 To extract significant spatio-temporal deflections in the TRFs at an SNR of 0  
614 dB, we applied a two-level statistical analysis (two-level cluster-test; e.g.  
615 Obleser et al., 2012). At the single-subject level, we used one-sample t-tests  
616 to test the TRF to the attended, the ignored as well as the attended-ignored  
617 difference against zero. Resulting t-values were transformed to z-scores. At  
618 the group level, the deflection of z-scores from zero was tested by a cluster-  
619 based permutation one-sample t-test (Maris and Oostenveld, 2007), which  
620 clusters t-values with p-values < 0.001 of adjacent time-electrode bins (with  
621 a minimum of 4 neighboring electrodes). The extracted cluster is compared  
622 to 4,000 clusters drawn randomly from the data by permuting condition  
623 labels. The resulting cluster p-value reflects the relative number of Monte  
624 Carlo iterations in which the summed t-statistic of the observed cluster is  
625 exceeded. This contrast indicates how components of the TRF are generally  
626 affected by attention under balanced conditions.

627 In a second step, the identical cluster-based permutation test was applied  
628 to obtain significant differences between the TRFs depending on whether  
629 a talker was dominant or non-dominant. This contrast was separately  
630 computed for the attended an ignored talker and it indicates, how the TRFs  
631 are affected by changing SNR.

632 In a third step, the difference between the TRFs to the attended and  
633 ignored talker were contrasted separately for *dominant* and *non-dominant*  
634 talkers. This contrast describes how attention affects the TRF to a dominant  
635 talker (easy-to-attend, hard to ignore) or a non-dominant talker (hard-to-  
636 attend, easy-to-ignore), respectively.

637 For illustration of the neural responses, we averaged single-subject TRF  $\beta$ -  
638 weights across channels of interest. Channels of interest were defined as  
639 the channels being part of both significant clusters found in the attended-  
640 ignored difference between TRFs under a balanced SNR of 0 dB (Fig. 2B).  
641 The 95%-confidence-bands were obtained by bootstrapping (Efron, 1979)  
642 across the averaged TRFs of all subjects, using 4,000 iterations.

## 643 **Neural tracking and neural selectivity**

644 To disentangle bottom-up and top-down effects, we investigated the TRFs  
645 based on two measures: *neural tracking* and *neural selectivity*. While *neural*  
646 *tracking* is a measure of how strongly a talker is encoded in the EEG  
647 (irrespective of attention), *neural selectivity* is a measure of how differential  
648 (i.e., attended vs. ignored) those representations are due to the impact of  
649 selective attention.

650 As a base for those two measures, we followed the forward method of  
651 predicting EEG signals and comparing those to the measured EEG signal, as  
652 described in detail by in Fiedler et al. (2017). In a leave-one-out fashion, we  
653 predicted EEG signals of a single trial contained in  $\hat{R}$  following the equation:

$$\hat{R} = SG_{TRF}, \quad (3)$$

654 where S is the matrix containing the onset envelopes and  $G_{TRF}$  is the matrix  
655 containing the trained TRFs.

656 *Neural tracking* was defined as the Pearson-correlation coefficient between  
657 the predicted and recorded EEG signals using the estimated TRFs (see  
658 above).

659 *Neural selectivity* was defined as the percentage of trials the TRFs could  
660 successfully identify a talker as being attended or ignored. Therefore, two  
661 different EEG signals were predicted per trial (Eq. 3), the first representing a  
662 talker being attended and the second representing the same talker being  
663 ignored. While one of the EEG signals is representing the task instruction  
664 (i.e., attend the to-be-attended talker; ignore the to-be-ignored talker), the  
665 other EEG signal represents the alternative (i.e. attending the to-be-ignored  
666 talker; ignoring the to-be-attended talker). We calculated the Pearson  
667 correlations for both predicted EEG signals with the measured EEG signal  
668 (Fiedler et al., 2017). Talker identification was successful if the EEG signal  
669 referring to the task instruction yielded higher correlation. Note that during  
670 unbalanced SNRs (i.e., -6 dB & +6 dB), the alternative EEG signal was  
671 predicted based on the TRFs estimated on the opposite SNR (e.g., under an

672 SNR of +6 dB, the alternative to attending the to-be-attended talker  
673 (*dominant*) is ignoring the to-be-ignored talker under an SNR of -6 dB).

674 Since this is a forward model approach, *neural tracking* and *neural selectivity*  
675 were obtained at every single EEG channel (Crosse et al., 2016). Likewise,  
676 both measures were obtained at the source level at every single voxel. We  
677 split up the prediction by either using only the prediction of the to-be-  
678 attended, only the prediction of the to-be-ignored or the sum of both  
679 predictions, such that the talker-specific contribution to *neural tracking*  
680 (*neural selectivity*) could be compared to the overall *neural tracking* (*neural*  
681 *selectivity*).

682 In order to evaluate the unfolding of *neural tracking* and *neural selectivity*  
683 over TRF time lags, we used a sliding-window of time lags (size: 48 ms, 6  
684 samples) with an overlap of 24 ms (3 samples) for the prediction. For every  
685 position of the window, *neural tracking* and *neural selectivity* were  
686 calculated (see above).

687 In advance of any arithmetic operation on *neural tracking*, the underlying  
688 Pearson-correlation coefficients were fisher-z transformed. Accordingly,  
689 *neural selectivity* (i.e., percentage correct) was logit-transformed.

## 690 **Source localization**

691 To further trace the origin of effects observed in sensor space, we applied  
692 LCMV-beamforming (Drongelen et al., 1994; Van Veen et al., 1997) to obtain  
693 source-activity time courses in single voxels of the brain. Using a standard  
694 template brain from Fieldtrip/SPM (Montreal Neurological Institute)  
695 together with the *Acticap* electrode layout, leadfields were calculated with  
696 a grid resolution of 10 mm. Individual LCMV-filter weights were obtained  
697 using 5% regularization. The continuous time-domain EEG data were  
698 projected to source space, resulting in three source activity time courses (X-  
699 Y-Z) per voxel. In order to obtain a single time course for each voxel, the  
700 direction of highest variance was determined by principal component  
701 analysis and used for further analysis. All further processing steps in source  
702 space were done analogously to sensor space EEG data.

703 **References**

- 704 Bentler RA, Palmer C, Dittbener AB. 2004. Hearing-in-Noise: Comparison of  
705 Listeners with Normal and (Aided) Impaired Hearing. *J Am Acad Audiol.* 15:216–  
706 225.
- 707 Biesmans W, Das N, Francart T, Bertrand A. 2016. Auditory-inspired speech  
708 envelope extraction methods for improved EEG-based auditory attention  
709 detection in a cocktail party scenario. *{IEEE} Trans Neural Syst Rehabil Eng.* 4320.
- 710 Brodbeck C, Hong LE, Simon JZ. 2018. Transformation from auditory to linguistic  
711 representations across auditory cortex is rapid and attention dependent for  
712 continuous speech. *bioRxiv.*
- 713 Broderick MP, Anderson AJ, Liberto GM Di, Crosse MJ, Edmund C. 2018.  
714 Electrophysiological correlates of semantic dissimilarity reflect the comprehension  
715 of natural, narrative speech. *Curr Biol.* 28:803–809.
- 716 Chait M, Cheveigné A De, Poeppel D, Simon JZ. 2010. *Neuropsychologia* Neural  
717 dynamics of attending and ignoring in human auditory cortex. *Neuropsychologia.*  
718 48:3262–3271.
- 719 Cherry EC. 1953. Some Experiments on the Recognition of Speech, with One and  
720 with Two Ears. *J Acoust Soc Am.* 25:975–979.
- 721 Chi T, Ru P, Shamma SA. 2005. Multiresolution Spectrotemporal Analysis of  
722 Complex Sounds. *J Acoust Soc Am.* 118:887–906.
- 723 Combrisson E, Jerbi K. 2015. Exceeding chance level by chance: The caveat of  
724 theoretical chance levels in brain signal classification and statistical assessment of  
725 decoding accuracy. *J Neurosci Methods.* 1–11.
- 726 Crosse MJ, Di Liberto GM, Bednar A, Lalor EC. 2016. The Multivariate Temporal  
727 Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals  
728 to Continuous Stimuli. *Front Hum Neurosci.* 10:604.
- 729 Davis MH, Johnsrude IS. 2003. Hierarchical Processing in Spoken Language  
730 Comprehension. 23:3423–3431.
- 731 Di Liberto GM, O’Sullivan JA, Lalor EC. 2015. Low-frequency cortical entrainment to  
732 speech reflects phoneme-level processing. *Curr Biol.* 25:2457–2465.

- 733 Ding N, Chatterjee M, Simon JZ. 2014. Robust cortical entrainment to the speech  
734 envelope relies on the spectro-temporal fine structure. *Neuroimage*. 88:41–46.
- 735 Ding N, Simon JZ. 2012. Neural coding of continuous speech in auditory cortex  
736 during monaural and dichotic listening. *J Neurophysiol*. 107:78–89.
- 737 Ding N, Simon JZ. 2013. Adaptive temporal encoding leads to a background-  
738 insensitive cortical representation of speech. *J Neurosci*. 33:5728–5735.
- 739 Drongelen W Van, Yuchtman M, Veen BD Van, Huffelen AC Van. 1994. A Spatial  
740 Filtering Technique to Detect and Localize Multiple Sources in the Brain. *Brain*  
741 *Topogr*. 9:39–49.
- 742 Efron B. 1979. Bootstrap methods: Another look at the jackknife. *Ann Stat*. 7:1–26.
- 743 Fiedler L, Wöstmann M, Graverson C, Brandmeyer A, Lunner T, Obleser J. 2017.  
744 Single-channel in-ear-EEG detects the focus of auditory attention to concurrent  
745 tone streams and mixed speech. *J Neural Eng*. 14.
- 746 Fuglsang SA, Dau T, Hjortkjær J. 2017. *NeuroImage* Noise-robust cortical tracking  
747 of attended speech in real-world acoustic scenes. *Neuroimage*. 156:435–444.
- 748 Giordano BL, Ince RAA, Gross J, Schyns PG, Panzeri S, Kayser C. 2017. Contributions  
749 of local speech encoding and functional connectivity to audio-visual speech  
750 perception. *Elife*. 6:1–27.
- 751 Hamilton LS, Huth AG. 2018. The revolution will not be controlled: natural stimuli  
752 in speech neuroscience. *Lang Cogn Neurosci*. 1–10.
- 753 Hertrich I, Dietrich S, Trouvain J, Moos A, Ackermann H. 2012. Magnetic brain  
754 activity phase-locked to the envelope, the syllable onsets, and the fundamental  
755 frequency of a perceived speech signal. *Psychophysiology*. 49:322–334.
- 756 Hoerl AE, Kennard RW. 1970. Ridge Regression: Biased Estimation for  
757 Nonorthogonal Problems. *Technometrics*. 12:55–67.
- 758 Horton C, Zmura MD, Srinivasan R. 2013. Suppression of competing speech  
759 through entrainment of cortical oscillations. *J Neurophysiol*. 109:3082–3093.
- 760 Kaya EM, Elhilali M. 2017. Modelling auditory attention. *Phil Trans R Soc B*. 372.

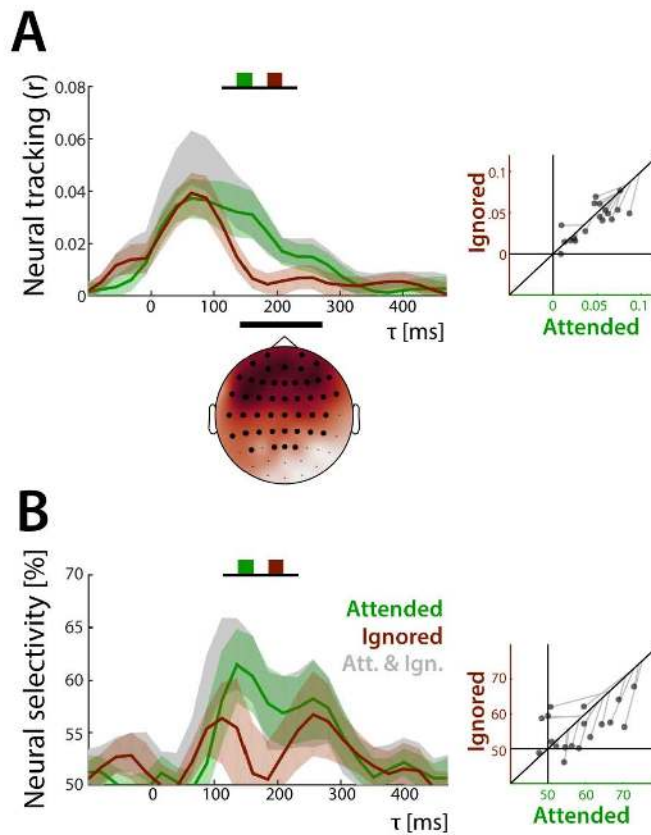
- 761 Kayser SJ, Ince RAA, Gross J, Kayser C. 2015. Irregular Speech Rate Dissociates  
762 Auditory Cortical Entrainment, Evoked Responses, and Frontal Alpha. *J Neurosci.*  
763 35:14691–14701.
- 764 Keitel A, Ince RAA, Gross J, Kayser C. 2017. NeuroImage Auditory cortical delta-  
765 entrainment interacts with oscillatory power in multiple fronto-parietal networks.  
766 *Neuroimage.* 147:32–42.
- 767 Kong Y-Y, Somarowthu A, Ding N. 2015. Effects of Spectral Degradation on  
768 Attentional Modulation of Cortical Auditory Responses to Continuous Speech. *J*  
769 *Assoc Res Otolaryngol.* 16:783–796.
- 770 Kong YY, Mullangi A, Ding N. 2014. Differential modulation of auditory responses  
771 to attended and unattended speech in different listening conditions. *Hear Res.*  
772 316:73–81.
- 773 Lalor EC, Power AJ, Reilly RB, Foxe JJ. 2009. Resolving Precise Temporal Processing  
774 Properties of the Auditory System Using Continuous Stimuli. *J Neurophysiol.*  
775 102:349–359.
- 776 Makeig S, Debener S, Onton J, Delorme A. 2004. Mining event-related brain  
777 dynamics. *Trends Cogn Sci.* 8:204–210.
- 778 Maris E, Oostenveld R. 2007. Nonparametric statistical testing of EEG- and MEG-  
779 data. *J Neurosci Methods.* 164:177–190.
- 780 Mcgarrigle R, Munro KJ, Dawes P, Stewart AJ, David R, Barry JG, Amitay S. 2014.  
781 Listening effort and fatigue: What exactly are we measuring? *Int J Audiol.* 53:433–  
782 445.
- 783 Melara RD, Rao A, Tong Y. 2002. The duality of selection: Excitatory and inhibitory  
784 processes in auditory selective attention. *J Exp Psychol.* 28:279–306.
- 785 Mesgarani N, Chang EF. 2012. Selective cortical representation of attended speaker  
786 in multi-talker speech perception. *Nature.* 485:233–236.
- 787 Mirkovic B, Debener S, Jaeger M, Vos M De. 2015. Decoding the attended speech  
788 stream with multi-channel EEG: implications for online, daily-life applications. *J*  
789 *Neural Eng.* 12:46007.



- 790 O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG,  
791 Slaney M, Shamma SA, Lalor EC. 2014. Attentional Selection in a Cocktail Party  
792 Environment Can Be Decoded from Single-Trial EEG. *Cereb Cortex*. 25:1697–1706.
- 793 Obleser J, Eisner F. 2009. Pre-lexical abstraction of speech in the auditory cortex.  
794 *Trends Cogn Sci*. 13:14–19.
- 795 Obleser J, Wöstmann M, Hellbernd N, Wilsch A, Maess B. 2012. Adverse Listening  
796 Conditions and Memory Load Drive a Common Alpha Oscillatory Network. *J*  
797 *Neurosci*. 32:12376–12383.
- 798 Olguin A, Bekinschtein TA, Bozic M. 2018. Neural Encoding of Attended Continuous  
799 Speech under Different Types of Interference. *J Cogn Neurosci*. (in Print).
- 800 Oostenveld R, Fries P, Maris E, Schoffelen J. 2011. FieldTrip: Open Source Software  
801 for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data.  
802 *Comput Intell Neurosci*. 2011.
- 803 Petersen EB, Wöstmann M, Obleser J, Lunner T. 2016. Neural tracking of attended  
804 versus ignored speech is differentially affected by hearing loss. *J Neurophysiol*.  
805 117:18–27.
- 806 Pichora-Fuller MK, Schneider BA, Daneman M. 1995. How young and old listen to  
807 and remember speech in noise. *J Acoust Soc Am*. 91:593–608.
- 808 Picton T. 2013. Hearing in Time: Evoked Potential Studies of Temporal Processing.  
809 *Ear Hear*. 34:385–401.
- 810 Pomper U, Chait M. 2017. The impact of visual gaze direction on auditory object  
811 tracking. *Sci Rep*. 7:1–16.
- 812 Power AJ, Foxe JJ, Forde EJ, Reilly RB, Lalor EC. 2012. At what time is the cocktail  
813 party? A late locus of selective attention to natural speech. *Eur J Neurosci*. 35:1497–  
814 1503.
- 815 Rönnerberg J. 2013. The Ease of Language Understanding (ELU) model: theoretical,  
816 empirical, and clinical advances. *Front Syst Neurosci*. 7:1–17.
- 817 Rudner M, Rönnerberg J, Lunner T. 2011. Working Memory Supports listening in  
818 Noise for Persons with Hearing Impairment. *J Am Acad Audiol*. 22:156–167.
- 819 Smeds K, Wolters F, Rung M. 2015. Estimation of Signal-to-Noise Ratios in Realistic  
820 Sound Scenarios. *J Am Acad Audiol*. 196:183–196.

- 821 Veen BD Van, Drongelen W Van, Yuchtman M, Suzuki A. 1997. Localization of Brain  
822 Electrical Activity via Linearly Constrained Minimum Variance Spatial Filtering. IEEE  
823 Trans Biomed Eng. 44:867–880.
- 824 Wang Y, Zhang J, Ding N, Zou J, Luo H. 2018. Prior Knowledge Guides Speech  
825 Segregation in Human Auditory Cortex. *Cereb Cortex*. bhy052:1–11.
- 826 Waschke L, Wöstmann M, Obleser J. 2017. States and traits of neural irregularity in  
827 the age-varying human brain. 1–12.
- 828 Willmore BDB, Cooke JE, King AJ. 2014. Hearing in noisy environments: noise  
829 invariance and contrast gain control. *J Physiol*. 16:3371–3381.
- 830 Woolgar A, Jackson J, Duncan J. 2016. Coding of Visual, Auditory, Rule, and  
831 Response Information in the Brain: 10 Years of Multivoxel Pattern Analysis. *J Cogn  
832 Neurosci*. 28:1433–1454.
- 833 Wöstmann M, Fiedler L, Obleser J. 2016. Tracking the signal, cracking the code:  
834 Speech and speech comprehension in non-invasive human electrophysiology.  
835 *Languange, Cogn Neurosci*. 1–15.
- 836 Wöstmann M, Lim S, Obleser J. 2017. The Human Neural Alpha Response to Speech  
837 is a Proxy of Attentional Control. *Cereb Cortex*. 27:3307–3317.
- 838 Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM,  
839 Goodman RR, Emerson R, Mehta AD, Simon JZ, Poeppel D, Schroeder CE. 2013.  
840 Mechanisms underlying selective neuronal tracking of attended speech at a  
841 “cocktail party.” *Neuron*. 77:980–991.
- 842

843 **Appendix**



**Figure S1: Unfolding of neural tracking and neural selectivity under the balanced SNR of 0 dB.** Neural tracking and neural selectivity were estimated based on the extracted TRFs to the attended (green), the ignored (red) as well as both talkers (grey). Confidence bands (95%) were obtained by bootstrapping. Highlighted channels (topographic maps) are part of a significant cluster. **A)** Neural tracking across all time lags (-100-500 ms). Scatterplots (bottom) show single-subject data averaged across channels of interest. Grey lines indicate overall neural tracking of both talkers at the 45°-line. **B)** Neural selectivity across all time lags (-100-500 ms). Scatterplots (bottom) show single-subject data averaged across channels of interest. Grey lines indicate overall neural selectivity of both talkers at the 45°-line.

844

845

846

847

#### 848 **Extraction of peak latencies and peak amplitude**

849 In order to disentangle amplitude- and latency-effects, we extracted peak  
850 latency and peak amplitude for every subject and every component.

851 Peak latencies were defined as the time lag of the maximum or minimum  
852 within a certain time interval ( $P1_{TRF}$ : 0–100 ms;  $N1_{TRF}$ : 100–200 ms;  $P2_{TRF}$ : 200–  
853 350 ms) of the subject- and SNR-specific TRF. The peak amplitudes were  
854 defined as the  $\beta$ -weights at the respective peak latencies. Please note that  
855 a reliable extraction of a P2 component was only possible in the TRF to the  
856 attended talker, whereas a reliable estimation of the N2 component was  
857 only achieved in the TRF to the dominant ignored talker under –6.

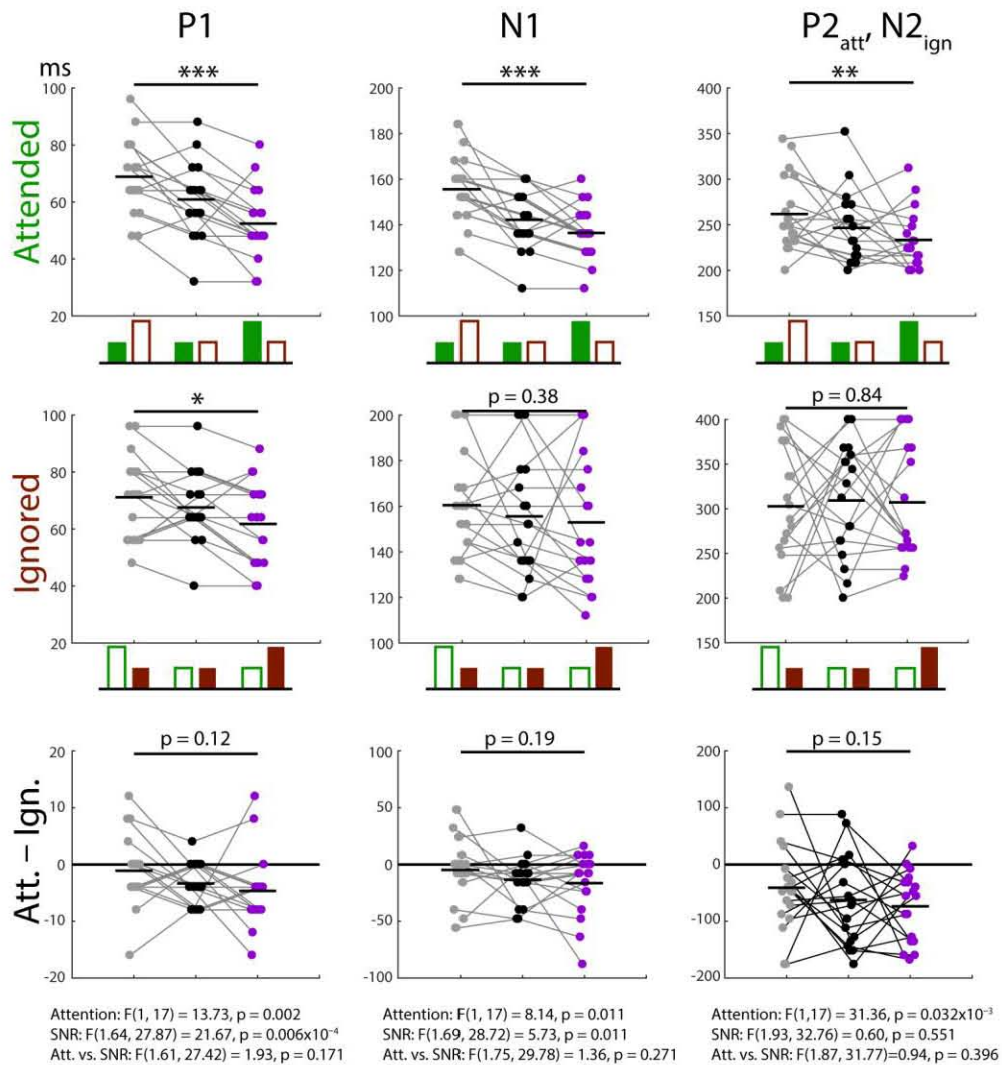
858 The main effects and interactions of attention and SNR on both the peak  
859 latency and peak amplitude were investigated by a repeated-measures  
860 ANOVA. Reported p-values were obtained with Greenhouse-Geisser-  
861 corrected degrees of freedom.

862

#### 863 **SNR-induced TRF latency shift**

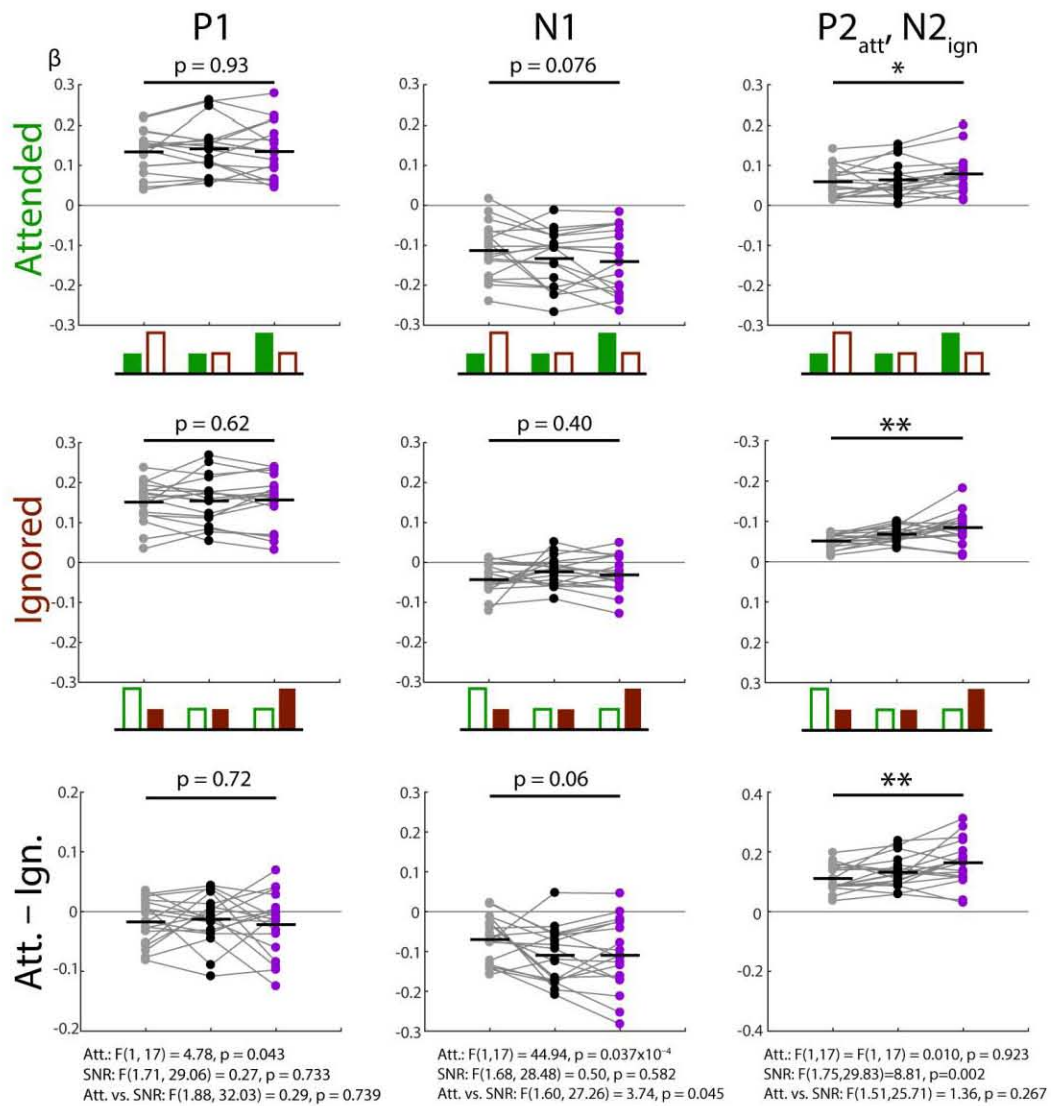
864 The contrasted TRFs between the *dominant* and *non-dominant* attended  
865 talker (Fig. 3A) showed significant differences during three time-lag  
866 intervals, first at around 20 ms (8–24 ms,  $p = 0.004$ ), a second around 100  
867 ms (80–128 ms,  $p = 2 \times 10^{-4}$ ) and a third around 200 ms (176–224 ms,  $p =$   
868 0.001). These differences occurred in the transition between components  
869 ( $P1_{TRF}$  to  $N1_{TRF}$ , and  $N1_{TRF}$  to  $P2_{TRF}$ ). This was consistent with the visual  
870 impression of the TRFs being similar in morphology yet delayed whenever  
871 a talker was less dominant. The TRFs to the ignored talker (Fig. 3B) also  
872 suggest such an SNR-related delay, even if no comparable significant  
873 differences were observed at the transitions between components.  
874 Nevertheless, the individual peak latencies showed a main effect of SNR for  
875 the  $P1_{TRF}$  (the more dominant the earlier; Fig. S2,  $F_{1,64,27.87} = 21.67$ ,  $p =$   
876  $0.006 \times 10^{-4}$ ), but also a main effect of attention (earlier if attended,  $F_{1,17} =$

877 13.73,  $p = 0.002$ ). No interaction between SNR and attention was found  
 878 ( $F_{1,61,27,42} = 1.93$ ,  $p = 0.171$ ; see appendix for more details).



**Figure S2: Peak latencies extracted from TRFs of single subjects for dominant (purple), balanced (black) and non-dominant talkers (attended and ignored).**

879



**Figure S3: Peak amplitudes extracted from TRFs of single subjects for dominant (purple), balanced (black) and non-dominant talkers (attended and ignored).**

880

881