The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention

Antonio Elia Forte, Octave Etard and Tobias Reichenbach*

Department of Bioengineering and Centre for Neurotechnology, Imperial College London, South Kensington Campus, SW7 2AZ, London, U.K.

*To whom correspondence should be addressed (email: reichenbach@imperial.ac.uk)

Abstract

Humans excel at selectively listening to a target speaker in background noise such as competing voices. While the encoding of speech in the auditory cortex is modulated by selective attention, it remains debated whether such modulation occurs already in subcortical auditory structures. Investigating the contribution of the human brainstem to attention has, in particular, been hindered by the tiny amplitude of the brainstem response. Its measurement normally requires a large number of repetitions of the same short sound stimuli, which may lead to a loss of attention and to neural adaptation. Here we develop a mathematical method to measure the auditory brainstem response to running speech, an acoustic stimulus that does not repeat and that has a high ecological validity. We employ this method to assess the brainstem's activity when a subject listens to one of two competing speakers, and show that the brainstem response is consistently modulated by attention.

Introduction

It is well known that selective attention to one of several competing acoustic signals affects the encoding of sound in the auditory cortex (Shinn-Cunningham 2008; Hackley et al. 1990; Choi et al. 2013; Fritz et al. 2007b; Hillyard et al. 1973; Womelsdorf & Fries 2007; Fritz et al. 2007a; Näätänen et al. 2001). Because extensive auditory centrifugal pathways carry information from central to more peripheral levels of the auditory system (Winer 2006; Pickels 1988; Song et al. 2008; Bajo et al. 2010), neural activity in the subcortical structures may contribute to attention as well. Previous attempts to determine an attentional modulation from recording the auditory brainstem response through scalp electrodes have, however, yielded highly inconclusive results.

In particular, one investigation found that selective attention alters the brainstem's response to the fundamental frequency of a speech signal (Galbraith et al. 1998), while another study concluded that this response is modulated in an unsystematic but subject-specific manner (Lehmann & Schönwiesner 2014) and a third recent experiment did not find a significant attentional effect (Varghese et al. 2015). Results on the effects of attention on the auditory-brainstem response to short clicks or pure tones are similarly inconclusive (Brix 1984; Gregory et al. 1989; Hoormann et al. 2000; Galbraith et al. 2003). These inconsistencies may result from a main experimental limitation in these studies: because the brainstem response is tiny, its measurement requires hundred- to thousandfold repetition of the same sound. The large number of repetitions may lead to difficulties for subjects in sustaining selective attention, to adaptation in the nervous system, and to a reduction in efferent feedback (Lasky 1997; Neupane et al. 2014).

To overcome this limitation, we develop here a method to measure the auditory brainstem's response to natural running speech that does not repeat. We then use this method to assess the modulation of the auditory brainstem response to one of two competing speakers by selective attention.

Results

Assessing the brainstem's response to continuous non-repetitive speech does not allow to average over many repeated presentations of the same sound. Instead, we sought to quantify the brainstem's response to the fundamental frequency of speech. Neuronal activity in the brainstem, and in particular in the inferior colliculus, can indeed phase lock to the periodicity of voiced speech (Skoe & Kraus 2010). The fundamental frequency of running speech varies over time, however, compounding a direct read-out of the evoked brainstem response.

To overcome this difficulty, we employed empirical mode decomposition (EMD) of the speech stimuli to identify an empirical mode that, at each time instance, oscillates at the fundamental frequency of the speech signal (Huang & Pan 2006) (Methods). This mode is a nonlinear and

nonstationary oscillation with a temporally-varying amplitude and frequency that we refer to as the 'fundamental waveform' of the speech stimulus (Figure 1a).

We then recorded the brainstem response to running non-repetitive speech stimuli of several minutes in duration from human volunteers through scalp electrodes. We cross-correlated the obtained recording with the fundamental waveform of the speech signal (Figure 1b). Because the brainstem response may occur at a phase that is different from that of the fundamental waveform, we also correlated the neural signal to the Hilbert transform of the fundamental waveform that has a phase delay of 90˚. The two correlations can be viewed as the real and imaginary part of a complex correlation function that can trace the brainstem response at any phase delay. The amplitude of the complex correlation informs then on the strength of the brainstem response.

We found that the amplitude of the complex correlation peaked at a mean latency of 9.3 ± 0.7 ms, and our statistical analysis showed that this peak was significantly different from the noise in fourteen out of sixteen subjects ($p < 0.05$, Methods). The average value of the correlation at the peak was 0.015 ± 0.003. Moreover, the latency agrees with that found previously regarding the brainstem's response to short repeated speech stimuli (Skoe & Kraus 2010). We checked that the response does not contain a stimulus artifact or a contribution from the cochlear microphonic, and that the latency of the response is not affected by the processing of the speech signal or of the neural response (Methods; Figure 1–figure supplement 1). This demonstrates that the brainstem's response to continuous speech can be reliably extracted through the developed method, and the response can be characterized through the latency and amplitude of the correlation's peak.

Armed with the ability to quantify the brainstem's response to running non-repetitive speech, we sought to investigate if this neural activity is affected by selective attention. Employing a well-established paradigm of attention to one of two speakers (Ding & Simon 2012), we presented volunteers diotically with two concurrent speech streams of equal intensity, one by a male and another by a female voice. For parts of the speech presentation subjects attended the male voice and ignored the female voice, and *vice versa* for the remaining parts.

We quantified the brainstem's response to both the male and the female voice by extracting the fundamental waveforms of both speech signals and correlating the neural recording separately to both. We found that the latency of the response was unaffected by attention: the response to the unattended speaker occurred 0.8 ± 0.5 ms later than that to the attended speaker, which was not statistically significant ($p = 0.2$; average over the responses to the male and the female voice as well as all subjects).

In contrast, all subjects showed a larger response of the auditory brainstem to the male voice when attending rather than ignoring it (Figure 2a). The difference in the responses was statistically

significant in nine of the fourteen subjects ($p < 0.05$). The brainstem's response to the attended female speaker similarly exceeded that to the unattended female voice in all but one subject, with eight subjects showing a statistically-significant difference ($p < 0.05$; Figure 2b). The ratio of the brainstem's response to attended and to ignored speech, averaged over all subjects, was 1.5 ± 0.1 and 1.6 ± 0.2 for the male and for the female speaker, respectively. Both ratios were significantly different from unity ($p < 0.001$, male voice; $p < 0.01$, female voice). The male and the female voice elicited a comparable attentional modulation: the difference between the corresponding ratios was insignificant ($p = 0.7$). The magnitude of the brainstem's response was hence significantly enhanced through attention, and consistently so across subjects and speakers.

Discussion

Our results show that the human auditory brainstem response to continuous speech is larger when attending than when ignoring a speech signal, and consistently so across different subjects and speakers. In particular, the strength of the phase locking of the neural activity to the pitch structure of speech is larger for an attended than for an unattended speech stream. In contrast, we did not observe a difference in the latency of this activity.

The fundamental waveform of speech that we have obtained from EMD has a temporally varying frequency and amplitude and is therefore not a simple component of Fourier analysis. While it may be obtained from short-time Fourier transform or wavelet analysis, both methods suffer from an inherently limited time-frequency resolution that makes them inferior to the EMD analysis (Huang & Pan 2006).

Because we have employed a diotic stimulus presentation in which the same acoustical stimulus was presented to each ear, the attentional modulation cannot result from a general modulation of the brainstem's activity to acoustic stimuli between the two hemispheres. Moreover, although the fundamental frequencies of the two competing speakers differ at most time points, their spectra largely overlap. The attentional modulation can therefore not result from a broad-band modulation of the neural activity either. Instead, the attentional effect must result from a modulation of the brainstem's response to the specific pitch structure of a speech stimulus.

The brainstem response to the pitch of continuous speech that we have measured can reflect a response both to the fundamental frequency of speech as well as to higher harmonics. Indeed, previous studies have found that the brainstem responds at the fundamental frequency of a speech stimulus even when that frequency itself is removed from the acoustic signal (Galbraith & Doan 1995), or when it cancels out due to presentation of stimuli with opposite polarities and averaging of the obtained responses (Aiken & Picton 2008). The attentional modulation of the brainstem response can thus reflect a modulation of the response to the fundamental frequency itself or to higher

126 harmonics. Moreover, attentional modulation of higher harmonics may depend on frequency as shown
127 recently in recordings of otoacoustic emissions from the inner ear (Maison et al. 2001).

128 The attentional modulation of the brainstem's response to the pitch of a speaker may result
129 from an enhancement of the neural response to an attended speech signal, from the suppression of the
130 response to an ignored speech stimulus, or from both. Further investigation into this issue may
131 compare brainstem responses to speech when attending to the acoustical signal and when attending to
132 a visual stimulus (Woods et al. 1992; Karns & Knight 2009; Saupe et al. 2009).

133 The response at the fundamental frequency of speech can result from multiple sites in the
134 brainstem (Chandrasekaran & Kraus 2010). However, we observed a single peak with a width of a
135 few ms in the correlation of the neural signal to the fundamental waveform of speech. The brainstem
136 response to running speech that we have measured here can therefore only reflect neural sources
137 whose latencies vary by a few ms or less from the peak latency. . The neural delay of about 9 ms as
138 well as the similarity of the speech-evoked brainstem response to the frequency-following response
139 suggest that the main neural source may be in the inferior colliculus (Sohmer et al. 1977). The
140 attentional effect that we have observed may then result from the multiple feedback loops between the
141 inferior colliculus, the medial geniculate body and the auditory cortex (Huffman & Henson 1990).

142 Our study provides the mathematical tools to analyse the brainstem response to complex, real
143 world stimuli such as speech. Since our method does not require artificial and repeated stimuli, it
144 fosters sustained attention and avoids potential neural adaptation. This method can therefore pave the
145 way to further explore how the brainstem contributes to the processing of complex real-world acoustic
146 environments. It may also be relevant for better understanding and diagnosing the recently discovered
147 cochlear neuropathy or 'hidden hearing loss' (Kujawa & Liberman 2009). Because the latter alters the
148 brainstem's activity (Schaette & McAlpine 2011; Mehraei et al. 2016), assessing the auditory
149 brainstem response to speech as well as its modulation by attention may further clarify the origin,
150 prevalence and consequences of such poorly understood supra-threshold hearing loss.

151

152 ## Methods

153 **Participants.** 16 healthy adult volunteers aged 18 to 32, eight of which were female, participated in
154 the study. All subjects were native English speakers and had no history of hearing or neurological
155 impairments. All participants had pure-tone hearing thresholds better than 20 dB hearing loss in both
156 ears at octave frequencies between 250 Hz and 8 kHz. Each subject provided written informed
157 consent. All experimental procedures were approved by the Imperial College Research Ethics
158 Committee.

Auditory brainstem recordings to running speech. Samples of continuous speech from a male and a female speaker were obtained from publicly available audiobooks (https://librivox.org). All samples had a duration of at least two minutes and ten seconds; some were slightly longer to end upon completion of a sentence. To construct speech samples with two competing speakers, samples from the male and from the female speaker were normalized to the same root-mean-square amplitude and then superimposed.

Participants were placed in a comfortable chair in an acoustically and electrically insulated room (IAC Acoustics, U.K.). A personal computer outside the room controlled audio presentation and data acquisition. Speech stimuli were presented at a sampling frequency of 44.1 kHz through a high-performance sound card (Xonar Essence STX, Asus, U.S.A.). Stimuli were delivered diotically through insert earphones (ER-3C, Etymotic, U.S.A.) at a level of 78 dB(C) SPL (C-weighted frequency response). Sound intensity was calibrated with an ear simulator (Type 4157, Brüel & Kjaer, Denmark). All subjects reported that the stimulus level was comfortable.

The response from the auditory brainstem was measured through five passive Ag/AgCl electrodes (Multitrode, BrainProducts, Germany). Two electrodes were positioned at the cranial vertex (Cz), two further electrodes were placed on the left and right mastoid processes, and the remaining electrode was positioned on the forehead to measure the ground. The impedance between each electrode and the skin was reduced to below 5 kΩ using abrasive electrolyte-gel (Abralyt HiCl, Easycap, Germany). The electrode on the left mastoid, at the cranial vertex and the ground electrode were connected to a bipolar amplifier with low-level noise and a gain of 50 (EP-PreAmp, BrainProducts, Germany). The remaining two electrodes were connected to a second identical bipolar amplifier. The output from both bipolar amplifiers was fed into an integrated amplifier (actiCHamp, BrainProducts, Germany) where it was low-pass filtered through a hardware anti-aliasing filter with a corner frequency of 4.9 kHz and sampled at 25 kHz. The audio signals were measured by the integrated amplifier as well through an acoustic adapter (Acoustical Stimulator Adapter and StimTrak, BrainProducts, Germany). The electrophysiological data were acquired through PyCorder (BrainProducts, Germany). The simultaneous measurement of the audio signal and the brainstem response from the integrated amplifier was employed to temporally align both signals to a precision of less than 40 µs, the inverse of the sampling rate (25 kHz).

Experimental design. In the first part of the experiment, each volunteer listened to four speech samples of the female speaker only. Comprehension questions were asked at the end of each part in order to verify the subject's attention to the story.

The second part of the experiment employed eight samples of speech that contained both a male and a female voice. During the presentation of the first four samples, subjects were asked to attend either the male or the female speaker. Volunteers were then presented with the next four speech

194 samples and asked to attend to the speaker that they had ignored earlier. Whether the subject was
195 asked to attend first to the male or to the female voice was determined randomly for every subject.
196 Comprehension questions were asked after each sample.

197 **Computation of the fundamental waveform of speech.** The fundamental waveform of each speech
198 sample with a single speaker was computed through a custom-written Matlab program (code available
199 on Github; Forte 2017). The fundamental waveform of a speech sample with two speakers followed
200 from the two corresponding samples with a single speaker only.

201 First, each speech signal was downsampled to 8,820 Hz, low-pass filtered at 1,500 Hz (FIR,
202 transition band 1,500 – 1,650 Hz, stopband attenuation -80 dB, passband ripple 1 dB, order 296) and
203 time-shifted to compensate for the filter delay. Silent parts between words were identified by
204 computing the envelope of the speech signal. Each part where the envelope was less than 10% of the
205 maximal value found in the speech was considered silent, and the speech signal there was set to zero.

206 Second, the instantaneous fundamental frequency of the voiced parts of the speech signal was
207 detected through the autocorrelation method, employing rectangular windows of 50 ms duration with
208 a successive overlap of 49 ms. Speech segments that yielded a fundamental frequency outside the
209 range of 60 Hz to 400 Hz, or in which the fundamental frequency varied by more than 10 Hz between
210 two successive windows were considered voiceless. The speech segments that corresponded to voiced
211 speech, as well as their fundamental frequency, were thus obtained. The fundamental frequency of
212 each segment was interpolated through a cubic spline, and varied between 100 and 300 Hz in each
213 segment. Note that this method yields the fundamental frequency but not by itself the fundamental
214 wavemode.

215 Third, the voiced speech segments where analysed through the Hilbert-Huang transform. The
216 latter is an adaptive signal processing based on empirical basis functions and can thus be better suited
217 for analysing nonlinear and nonstationary signals such as speech than Fourier analysis (Huang & Pan
218 2006). The transform consists of two parts. First, empirical mode decomposition extracts intrinsic
219 mode functions (IMFs) that satisfy two properties: (i) the numbers of extrema and zero crossings are
220 either equal or differ by one; (ii) the mean of the upper and lower envelope vanishes. The signal
221 follows as the linear superposition of the IMFs. Second, the Hilbert spectrum of each IMF is
222 determined, which yields, in particular, the mode's instantaneous frequency. This analysis was
223 performed for each short segment of voiced speech, that is, for each part of voiced speech that was
224 preceded and followed by a pause or voiceless speech.

225 Fourth, the fundamental frequency of each short speech segment was compared to the
226 instantaneous frequencies of the segment's IMFs at each individual time point. All IMFs with an
227 instantaneous frequency that differed by less than 20% from the segment's fundamental frequency
228 were determined, and the IMF with the largest amplitude was therefrom selected as the fundamental

wavemode of that segment and at that time point (Huang & Pan 2006). If no IMF had an instantaneous frequency within 20% of the fundamental frequency, or if a speech segment was unvoiced, that time point was assigned a fundamental waveform of zero. The fundamental waveforms obtained at the different time points were combined through cosine crossfading functions with a window width of 10 ms to obtain the fundamental waveform of the speech signal. The Hilbert transform of that fundamental waveform was computed as well.

To control for latency changes in the acoustic signal induced by the subsequent processing steps, and in particular by the involved frequency filtering, the cross-correlation between the original speech signal and the fundamental waveform as well as with its Hilbert transform was computed (Figure 1–figure supplement 1**a**). The cross-correlations show that the fundamental waveform has no latency change and no phase difference with respect to the original speech stimulus.

Analysis of the auditory-brainstem response. The brainstem responses from the two measurement channels were averaged. A frequency-domain regression technique (CleanLine, EEGLAB) was used to attenuate noise from the power line in the brainstem recording. Moreover, because a voltage amplitude above 20 mV cannot result from the brainstem but represents artefacts such as spurious muscle activity, the signal was set to zero during episodes of such high voltage. The electrophysiological recording was then filtered between 100 – 300 Hz since the fundamental frequency of the speech was in that range (high-pass filter: FIR, transition band from 90 –100 Hz, stopband attenuation -80 dB, passband ripple 1 dB, order 6862; low-pass filter: FIR, transition band 300 – 360 Hz, stopband attenuation -80 dB, passband ripple 1 dB, order 1054). In particular, the high-pass filter eliminated neural signals from the cerebral cortex that occur predominantly below 100 Hz. To avoid transient activity at the beginning of each speech sample, the first ten seconds of each brainstem recording in response to a speech sample were discarded. The following two minutes of data were divided into 40 epochs of a duration of 3 s each, and the remaining data were discarded, if any.

The processing of the neural signal did not induce a latency. This was confirmed by computing the cross-correlation between the processed neural response and the original signal, demonstrating a maximum correlation at zero temporal delay (Figure 1–figure supplement 1**b**).

As set out above, the first part of the experiment measured the brainstem response to running speech without background noise. For each subject and each epoch, the cross-correlation of the brainstem response with the corresponding segment of the fundamental waveform as well as with its Hilbert transform were computed. A delay of 1 ms of the acoustic signal produced by the earphones was taken into account. The two cross-correlation functions were interpreted as the real and the imaginary part of a complex correlation function. For each individual subject, the average of the

263 complex cross-correlation over all epochs was then computed, and the latency at which the amplitude
264 peaked was determined.

265 The obtained latencies of about 9 ms affirmed that the signal resulted from the auditory
266 brainstem and not from the cerebral cortex, whose latencies exceed 20 ms. The latency also evidenced
267 that the signal resulted neither from stimulus artifacts nor from the cochlear microphonic, which
268 would occur at or near zero delay (Skoe & Kraus 2010). As an additional control, the brainstem
269 response was recorded when the earphones were near the ear, but not inserted into the ear canal, so
270 the subject could not hear the speech signals. The recording did then not yield a measurable brainstem
271 response (Figure 1–figure supplement 1**c**). Two presentations of the same speech stimulus, but with
272 opposite polarities, were employed as well, and the neural response to both presentations was
273 averaged before computing the correlation to the fundamental waveform. The correlation was
274 identical to that obtained by a single stimulus presentation, demonstrating the absence of a stimulus
275 artifact and of the cochlear microphonic (Figure 1–figure supplement 1**d**).

276 To determine whether the peak in the cross-correlation obtained from a given subject was
277 significant, the values of the complex cross-correlation from the individual epochs, and at the peak
278 latency, were analysed. Because each correlation value is an average of many measurements, it
279 follows from the Central Limit Theorem that the complex correlations from the different epochs
280 exhibit a two-dimensional normal distribution with a mean of zero if the measurements are randomly
281 distributed. A one-sample Hotelling's T-squared test was therefore used to assess the significance of
282 the complex correlation at the peak latency. Two subjects who did not show a significant correlation
283 ($p > 0.05$) were not included in the further analysis.

284 The population mean and standard error of the mean of the latency were computed from the
285 latencies of the individual subjects.

286 The brainstem responses to competing speakers were then analysed for each individual
287 subject. For each epoch, the complex cross-correlation between the brainstem response and the
288 fundamental waveform was computed, both for the fundamental waveform of the attended and for
289 that of the unattended speaker. The corresponding complex correlation functions were averaged
290 across epochs, and the amplitudes as well as latencies of the peaks were determined.

291 Statistical significance of the difference in latency of the brainstem responses to the attended
292 and the unattended speaker, obtained from the eight samples, was tested by computing population
293 mean as well as standard error of the mean for the differences in latencies obtained from individual
294 subjects. A two-tailed Student's t-test was employed to test if the difference was significantly different
295 from zero.

296 To control for differences in the voice of the male and the female speaker, differences in
297 amplitude of the brainstem response to the attended and ignored male speaker were determined
298 separately from differences in the amplitude of the brainstem response to the attended and ignored
299 female speaker. The amplitudes of the complex cross-correlations, at the peak latencies, were
300 computed for all epochs. A two-sample Student's t-test was then employed to test for a significant
301 difference between the amplitude in response to the attended and the ignored speaker.

302 The amplitude of the brainstem response to speech can vary widely between subjects (Figure 2), due
303 to variations such as in anatomy and scalp conductivity. The ratios of the amplitudes of the brainstem
304 responses to attended and ignored speech, rather than the differences, were thus computed for each
305 individual. The population mean and standard error of the mean were therefrom obtained. A one-
306 tailed Student's t-test assessed whether the population average of the ratio was significantly larger
307 than unity. A two-tailed two-sample Student's t-test was employed to assess whether the ratios
308 obtained from the responses to the male and to the female speaker were significantly different.

309

310 ## Acknowledgement

311 We thank Steve Bell, Karolina Kluk-de Kort, Patrick Naylor, David Simpson and Malcolm Slaney for
312 discussion as well as for comments on the manuscript. This research was supported by EPSRC grant
313 EP/M026728/1 to T.R. as well as in part by the National Science Foundation under Grant No. NSF
314 PHY-1125915.

315

316 ### Competing financial interests

317 The authors declare no competing financial interests.

318

319 ## References

320 Aiken, S.J. & Picton, T.W., 2008. Envelope and spectral frequency-following responses to vowel
321 sounds. *Hear. Res.*, 245(1–2), pp.35–47.

322 Bajo, V.M., Nodal, F.R., Moore, D.R. & King, A.J., 2010. The descending corticocollicular pathway
323 mediates learning-induced auditory plasticity. *Nat. Neurosci.*, 13(2), pp.253–260.

324 Brix, R., 1984. The influence of attention on the auditory brain stem evoked responses preliminary
325 report. *Acta Otolaryngol.*, 98(1–2), pp.89–92.

326 Chandrasekaran, B. & Kraus, N., 2010. The scalp-recorded brainstem response to speech: neural

origins and plasticity. *Psychophysiology*, 47(2), pp.236–246.

Choi, I., Rajaram, S., Varghese, L.A. & Shinn-Cunningham, B.G., 2013. Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography. *Front. Hum. Neurosci.*, 7, p.115.

Ding, N. & Simon, J.Z., 2012. Emergence of neural encoding of auditory objects while listening to competing speakers. *Proc. Natl. Acad. Sci. U. S. A.*, 109(29), pp.11854–9.

Forte, A.E., 2017. Fundamental_waveforms_extraction, GitHub, https://github.com/antn85/fundamental_waveforms_extraction, version 1.

Fritz, J.B., Elhilali, M., David, S. V. & Shamma, S.A., 2007a. Auditory attention - focusing the searchlight on sound. *Curr. Opin. Neurobiol.*, 17(4), pp.437–455.

Fritz, J.B., Elhilali, M., David, S. V. & Shamma, S.A., 2007b. Does attention play a role in dynamic receptive field adaptation to changing acoustic salience in a1? *Hear. Res.*, 229(1–2), pp.186–203.

Galbraith, G.C., Bhuta, S.M., Choate, A.K., Kitahara, J.M. & Mullen, T.A., 1998. Brain stem frequency-following response to dichotic vowels during attention. *Neuroreport*, 9(8), pp.1889–1893.

Galbraith, G.C. & Doan, B.Q., 1995. Brainstem frequency-following and behavioral responses during selective attention to pure tone and missing fundamental stimuli. *Int. J. Psychophysiol.*, 19(3), pp.203–214.

Galbraith, G.C., Olfman, D.M. & Huffman, T.M., 2003. Selective attention affects human brain stem frequency-following response. *Neuroreport*, 14(5), pp.735–8.

Gregory, S.D., Heath, J.A. & Rosenberg, M.E., 1989. Does selective attention influence the brain-stem auditory evoked potential? *Electroencephalogr. Clin. Neurophysiol.*, 73(6), pp.557–60.

Hackley, S.A., Woldorff, M. & Hillyard, S.A., 1990. Cross-modal selective attention effects on retinal, myogenic, brainstem, and cerebral evoked potentials. *Psychophysiology*, 27(2), pp.195–208.

Hillyard, S.A., Hink, R.F., Schwent, V.L. & Picton, T.W., 1973. Electrical signs of selective attention in the human brain. *Science*, 182(October), pp.177–180.

Hoormann, J., Falkenstein, M. & Hohnsbein, J., 2000. Early attention effects in human auditory-evoked potentials. *Psychophysiology*, 37(1), pp.29–42.

Huang, H. & Pan, J., 2006. Speech pitch determination based on hilbert-huang transform. *Signal*

Process., 86(4), pp.792–803.

Huffman, R.F. & Henson, O.W., 1990. The descending auditory pathway and acousticomotor systems: connections with the inferior colliculus. *Brain Res. Rev.*, 15(3), pp.295–323.

Karns, C.M. & Knight, R.T., 2009. Intermodal auditory, visual, and tactile attention modulates early stages of neural processing. *J. Cogn. Neurosci.*, 21(4), pp.669–83.

Kujawa, S.G. & Liberman, M.C., 2009. Adding insult to injurt: cochlear nerve degeneration after "temporary" noise- induced hearing loss. *J. Neurosci.*, 29(45), p.14077-.

Lasky, R.E., 1997. Rate and adaptation effects on the auditory evoked brainstem response in human newborns and adults. *Hear. Res.*, 111(1–2), pp.165–176.

Lehmann, A. & Schönwiesner, M., 2014. Selective attention modulates human auditory brainstem responses: relative contributions of frequency and spatial cues. *PLoS One*, 9(1), pp.1–10.

Maison, S., Micheyl, C. & Collet, L., 2001. Influence of focused auditory attention on cochlear activity in humans. *Psychophysiology*, 38(1), pp.35–40.

Mehraei, G., Hickox, A.E., Bharadwaj, H.M., Goldberg, H., Verhulst, S., Liberman, M.C., Barbara, X. & Shinn-Cunningham, G., 2016. Auditory brainstem response latency in noise as a marker of cochlear synaptopathy. *J. Neurosci.*, 36(13), pp.3755–3764.

Näätänen, R., Tervaniemi, M., Sussman, E., Paavilainen, P. & Winkler, I., 2001. "Primitive intelligence" in the auditory cortex. *Trends Neurosci.*, 24(5), pp.283–288.

Neupane, A.K., Gururaj, K., Mehta, G. & Sinha, S.K., 2014. Effect of repetition rate on speech evoked auditory brainstem response in younger and middle aged individuals. *Audiol. Res.*, 4(1), pp.21–27.

Pickels, J.O., 1988. *An introduction to the physiology of hearing*, Emerald.

Saupe, K., Widmann, A., Bendixen, A., Müller, M.M. & Schröger, E., 2009. Effects of intermodal attention on the auditory steady-state response and the event-related potential. *Psychophysiology*, 46(2), pp.321–327.

Schaette, R. & McAlpine, D., 2011. Tinnitus with a normal audiogram: physiological evidence for hidden hearing loss and computational model. *J. Neurosci.*, 31(38), pp.13452–13457.

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

Skoe, E. & Kraus, N., 2010. Auditory brain stem response to complex sounds: a tutorial. *Ear Hear.*, 31(3), pp.302–324.

389 Sohmer, H., Pratt, H. & Kinarti, R., 1977. Sources of frequency following responses (ffr) in man.
390 *Electroencephalogr. Clin. Neurophysiol.*, 42(5), pp.656–664.

391 Song, J.H., Skoe, E., Wong, P.C.M. & Kraus, N., 2008. Plasticity in the adult human auditory
392 brainstem following short-term linguistic training. *J. Cogn. Neurosci.*, 20(10), pp.1892–902.

393 Varghese, L., Bharadwaj, H.M. & Shinn-Cunningham, B.G., 2015. Evidence against attentional state
394 modulating scalp-recorded auditory brainstem steady-state responses. *Brain Res.*, 1626, pp.146–
395 164.

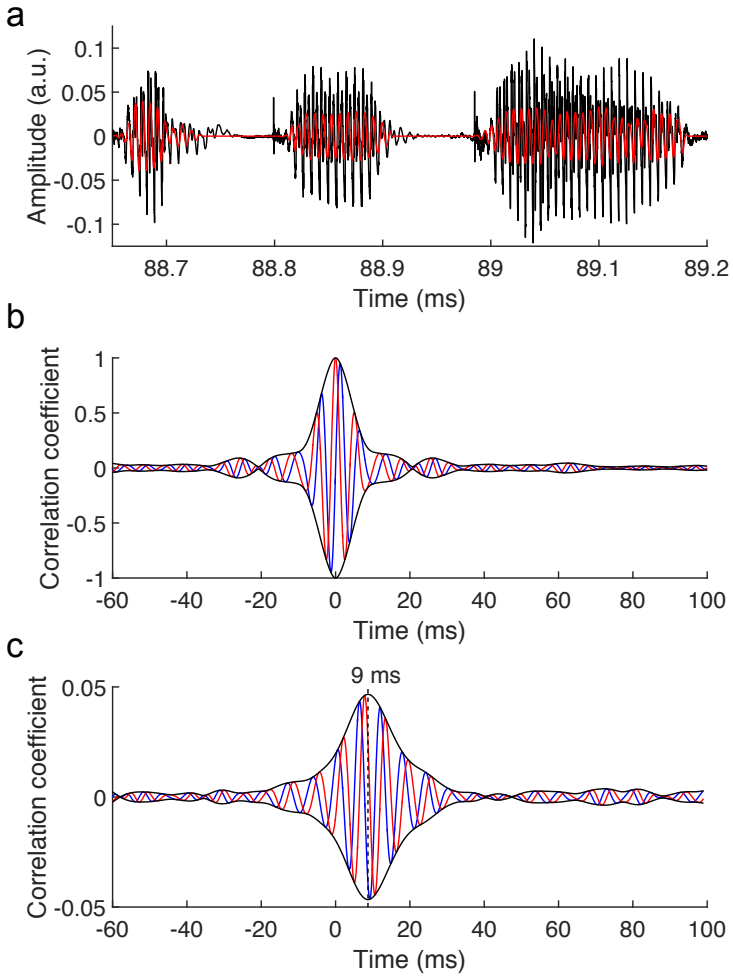
396 Winer, J.A., 2006. Decoding the auditory corticofugal systems. *Hear. Res.*, 212(1–2), pp.1–8.

397 Womelsdorf, T. & Fries, P., 2007. The role of neuronal synchronization in selective attention. *Curr.*
398 *Opin. Neurobiol.*, 17(2), pp.154–160.

399 Woods, D.L., Alho, K. & Algazi, A., 1992. Intermodal selective attention. i. effects on event-related
400 potentials to lateralized auditory and visual stimuli. *Electroencephalogr. Clin. Neurophysiol.*,
401 82(5), pp.341–355.

402

403 **Figures**



404

405

406 **Figure 1** The brainstem response to running speech. (**a**) Speech (black) contains voiced parts with
407 irregular oscillations at a time-varying fundamental frequency and higher harmonics. We extract a
408 fundamental waveform (red) that oscillates nonlinearly at the fundamental frequency. (**b**) The
409 autocorrelation of the fundamental waveform (red) peaks when the delay vanishes and oscillates at the
410 average fundamental frequency. The cross-correlation of the fundamental waveform with its Hilbert
411 transform (blue) can be seen as an imaginary part of the autocorrelation. The amplitude of the
412 resulting complex cross-correlation (black) shows a life-time of a few ms. (**c**) The correlation of the
413 speech-evoked brainstem response, recorded from one subject, to the fundamental waveform of the
414 speech signal (red) as well as to its Hilbert transform (blue) can serve as real and imaginary parts of a
415 complex correlation function. Its amplitude (black) peaks at a latency of 9 ms. The latency of the
416 correlation is not altered by the processing of the speech signal or of the neural recording, and
417 contains neither a stimulus artifact nor the cochlear microphonic (**Figure 1–figure supplement 1**).

418

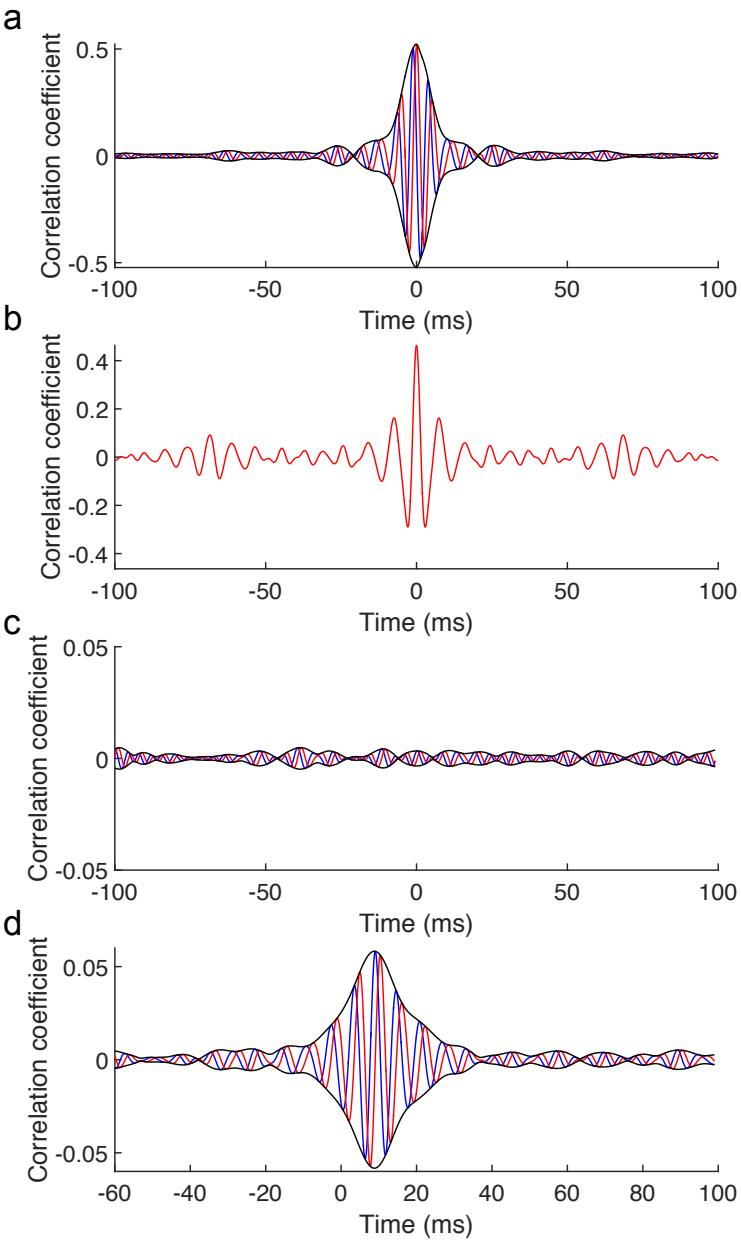
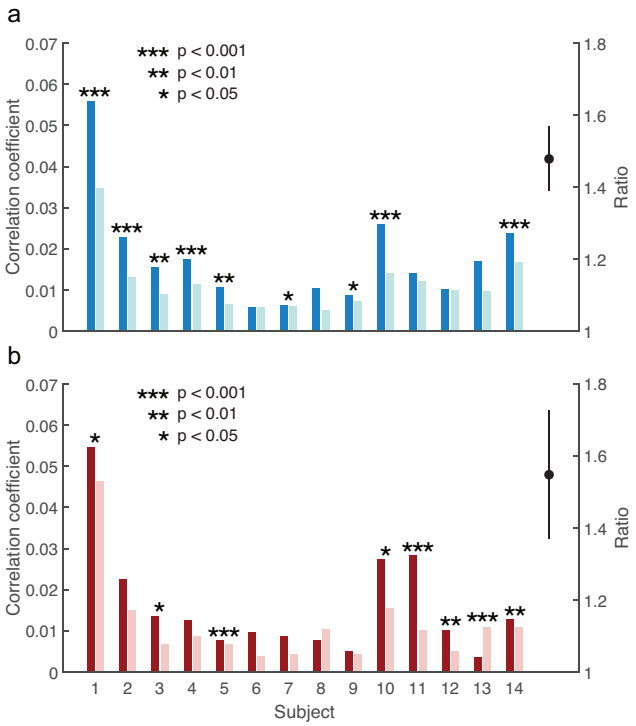


Figure 1–figure supplement 1. Controls for latencies induced by signal processing as well as for the source of the measured brainstem response to running speech. (**a**) The cross-correlation between the original speech signal with the fundamental waveform (red) as well as with its Hilbert transform (blue) and the resulting amplitude (black) show a peak at 0 ms and no phase delay. The processing of the acoustic signal does accordingly not change the latency or phase of that signal. (**b**) The computation of the cross-correlation of the fundamental waveform to the neural recording involved processing of the neural signal such as through filtering. However, the cross-correlation between the recorded neural signal and the filtered version shows a peak at vanishing latency. The processing of the neural signal did therefore not alter the latency. (**c**) When the earphones are placed close to the ears, but not inside the ear canal, preventing a subject from hearing the speech signal, the cross-correlation between the recorded neural signal and the fundamental waveform of speech (red) as well as its Hilbert transform (blue) do not yield a measurable peak. The amplitude of the resulting complex correlation function (black) does not peak either, demonstrating the absence of a stimulus artifact. (**d**)

When a subject listened to a speech signal and then to the same signal with reversed polarity, and when the average over the neural recordings to both stimulus presentations was employed for the analysis, the complex cross-correlation showed the same structure as when it was computed using the neural response to one stimulus only. This shows the absence of a stimulus artifact as well as the absence of the cochlear microphonic in the measured response. To enable comparison, all recordings were obtained from the same subject for whom we report the exemplary recording in Figure 1 (**c**).



443

444

445 **Figure 2** Modulation of the brainstem response to speech by selective attention. (**a**) The brainstem's
446 response to the male speaker is larger for each subject when attending the speaker (dark blue) than
447 when ignoring it (light blue). The average ratio of the brainstem responses to the attended and to the
448 ignored male speaker is significantly larger than 1 (black, mean and standard error of the mean). (**b**)
449 With the exception of subject 13, the neural response to the female voice is also larger when subjects
450 attend to it (dark red) instead of ignoring it (light red). The average ratio of the brainstem responses to
451 the attended and to the ignored female speaker is significantly larger than 1 as well (black, mean and
452 standard error of the mean).

453