1 EXPOSING DISTINCT SUBCORTICAL COMPONENTS OF THE AUDITORY BRAINSTEM RESPONSE

2 EVOKED BY CONTINUOUS NATURALISTIC SPEECH

3 Melissa J Polonenko[1, 3, 4], Ph.D. (ORCID: 0000-0003-1914-6117)

4 Ross K Maddox, Ph.D.[1, 2, 3, 4, *] (ORCID: 0000-0003-2668-0238)

5

6 [1] Department of Neuroscience, University of Rochester

7 [2] Department of Biomedical Engineering, University of Rochester

8 [3] Del Monte Institute for Neuroscience, University of Rochester

9 [4] Center for Visual Science, University of Rochester

10

11 * to whom correspondence should be addressed:

12 Ross Maddox

13 University of Rochester

14 Goergen Hall

15 Box 270168

16 Rochester, NY 14627

17 Email: ross.maddox@rochester.edu

18

ABSTRACT

Speech processing is built upon encoding by the auditory nerve and brainstem, yet we know very little about how these processes unfold in specific subcortical structures. These structures are deep and respond quickly, making them difficult to study during ongoing speech. Recent techniques begin to address this problem, but yield temporally broad responses with consequently ambiguous neural origins. Here we describe a method that pairs re-synthesized "peaky" speech with deconvolution analysis of EEG recordings. We show that in adults with normal hearing, the method quickly yields robust responses whose component waves reflect activity from distinct subcortical structures spanning auditory nerve to rostral brainstem. We further demonstrate the versatility of peaky speech by simultaneously measuring bilateral and ear-specific responses across different frequency bands, and discuss important practical considerations such as talker choice. The peaky speech method holds promise as a tool for investigating speech encoding and processing, and for clinical applications.

KEYWORDS

speech, auditory brainstem response, evoked potentials, electroencephalography, assessment

INTRODUCTION

Understanding speech is an important, complex process that spans the auditory system from cochlea to cortex. A temporally precise network transforms the strikingly dynamic fluctuations in amplitude and spectral content of natural, ongoing speech into meaningful information, and modifies that information based on attention or other priors (Mesgarani et al., 2009). Subcortical structures play a critical role in this process – they do not merely relay information from the periphery to the cortex, but also perform important functions for speech understanding, such as localizing sound (e.g., Grothe and Pecka, 2014) and encoding vowels across different levels and in background noise (e.g., Carney et al., 2015). Furthermore, subcortical structures receive descending information from the cortex through corticofugal pathways (Bajo et al., 2010; Bajo and King, 2012; Winer, 2005), suggesting they may also play an important role in modulating speech and auditory streaming. Given the complexity of speech processing, it is important to parse and understand contributions from different neural generators. However, these subcortical structures are deep and respond to stimuli with very short latencies, making them difficult to study during ecologically-salient stimuli such as continuous and naturalistic speech. We created a novel paradigm aimed at elucidating the contributions from distinct subcortical structures to ongoing, naturalistic speech.

Activity in deep brainstem structures can be "imaged" by the latency of waves in a surface electrical potential (electroencephalography, EEG) called the auditory brainstem response (ABR). The ABR's component waves have been attributed to activity in different subcortical structures with characteristic latencies: the auditory nerve contributes to waves I and II (~1.5–3 ms), the cochlear nucleus to wave III (~4 ms), the superior olivary complex and lateral lemniscus to wave IV (~ 5 ms), and the lateral lemniscus and inferior colliculus to wave V (~6 ms) (Møller and Jannetta, 1983; review by Moore, 1987; Starr and Hamilton, 1976). Waves I, III, and V are most often easily distinguished in the human response. Subcortical structures may also contribute to the earlier P_0 (12–14 ms) and N_a (15–25 ms) waves (Hashimoto, 1982; Kileny et al., 1987; Picton et al., 1974) of the middle latency response (MLR), which are then followed by thalamo-cortically generated waves P_a, N_b, and P_b/P_1 (Geisler et al., 1958; Goldstein and Rodman, 1967). ABR and MLR waves have a low signal-to-noise ratio (SNR) and require numerous stimulus repetitions to record a good response. Furthermore, they are quick and often occur before the stimulus has ended. Therefore, out of necessity, most human brainstem studies have focused on brief stimuli such as clicks, tone pips, or speech syllables, rather than more natural speech.

Recent analytical techniques have overcome limitations on stimuli, allowing continuous naturally uttered speech to be used. One such technique extracts the fundamental waveform from the speech stimulus and finds the envelope of the cross-correlation between that waveform and the recorded EEG data (Forte et al., 2017). The response has an average peak time of about 9 ms, with contributions primarily from the inferior colliculus (Saiz-Alia and Reichenbach, 2020). A second technique considers the rectified broadband speech waveform as the input to a linear system and the EEG data as the output, and uses deconvolution to compute the ABR waveform as the impulse response of the system (Maddox and Lee, 2018). The speech-derived ABR shows a wave V peak whose latency is highly correlated with the click response wave V across subjects, demonstrating that the component is generated in the rostral brainstem. A third technique averages responses to each chirp (click-like transients that quickly increase in frequency) in re-synthesized "cheech" stimuli (CHirp spEECH; Miller et al., 2017) that interleaves alternating octave frequency bands of speech and of chirps aligned with some glottal pulses (Backer et al., 2019). Brainstem responses to these stimuli also show a wave V, but do not show earlier waves unless presented monaurally over headphones (Backer et al., 2019; Miller et al., 2017). While these methods reflect subcortical activity, the first two provide temporally broad responses with a lack of specificity regarding underlying neural sources. None of the three methods shows the earlier canonical components such as waves I and III that would allow rostral brainstem activity to be distinguished from, for example, the auditory nerve. Such activity is important to assess, especially given the current interest in the potential

82 contributions of auditory nerve loss in disordered processing of speech in noise (Bramhall et al., 2019;
83 Liberman et al., 2016; Prendergast et al., 2017).

84   Although click responses can assess sound encoding (of clicks) at early stages of the auditory
85 system, a speech-evoked response with the same components would assess region-specific encoding
86 within the acoustical context of the dynamic spectrotemporal characteristics of speech – information that is
87 not possible to obtain from click responses. Furthermore, changes to the amplitudes and latencies of these
88 early components could inform our understanding of speech processing if deployed in experiments that
89 compare conditions requiring different states of processing, such as attended/unattended speech or
90 understood/foreign language. For example, if a wave I from the auditory nerve differed between speech
91 stimuli that were attended versus unattended, then this would add to our current understanding of the
92 brainstem's role in speech processing. Therefore, a click-like response that is evoked by speech stimuli
93 facilitates new investigations into speech encoding and processing.

94   Thus, we asked if we could further assess underlying speech encoding and processing in multiple
95 distinct early stages of the auditory system by 1) evoking additional waves than wave V of the canonical
96 ABR, and 2) measuring responses to different frequency ranges of speech (corresponding to different
97 places of origin on the cochlea). The ABR is strongest to very short stimuli such as clicks, so we created
98 "peaky" speech. The design goal of peaky speech is to re-synthesize natural speech so that its defining
99 spectrotemporal content is unaltered – maintaining the speech as intelligible and identifiable – but its
100 pressure waveform consists of maximally sharp peaks so that it drives the ABR as effectively as possible
101 (giving a very slight "buzzy" quality when listening under good headphones; Audio files 1–6). The results
102 show that peaky speech evokes canonical brainstem responses and frequency-specific responses, paving
103 the way for novel studies of subcortical contributions to speech processing.

104

105 **RESULTS**

106 <u>Broadband peaky speech yields more robust responses than unaltered speech</u>

107 *Broadband peaky speech elicits canonical brainstem responses*

108   In previous work, brainstem responses to natural, on-going speech exhibited a temporally broad
109 wave V but no earlier waves (Maddox and Lee, 2018). We re-synthesized speech to be "peaky" with the
110 primary aim to evoke additional, earlier waves of the ABR that identify different neural generators. Indeed,
111 Figure 1 shows that waves I, III, and V of the canonical ABR are clearly visible in the group average and in
112 the individual responses to broadband peaky speech. This means that broadband peaky speech, unlike the
113 unaltered speech, can be used to assess naturalistic speech processing at discrete parts of the subcortical
114 auditory system, from the auditory nerve to rostral brainstem. These responses represent weighted
115 averaged data from ~43 minutes of continuous speech (40 epochs of 64 s each), and were filtered at a
116 typical high-pass cutoff of 150 Hz to highlight the earlier ABR waves. Responses containing an equal
117 number of epochs from the first and second half of the recording had a median (interquartile range)
118 correlation coefficient of 0.96 (0.95–0.98) for the 0–15 ms lags, indicating good replication.

119   Morphology of the broadband peaky speech ABR was inspected and waves marked by a trained
120 audiologist (MJP) on 2 occasions that were 3 months apart. The intraclass correlation coefficients for
121 absolute agreement (ICC3) were ≥ 0.90 (lowest ICC3 95% confidence interval was 0.79–0.95 for wave I, p
122 < 0.01), indicating excellent reliability for chosen peak latencies. Waves I and V were identifiable in
123 responses from all subjects (N = 22), and wave III was identifiable in 19 of the 22 subjects. The numbers of
124 subjects with identifiable waves I and III in these peaky speech responses were similar to the 24 and 16 out
125 of 24 subjects for the click-evoked responses in Maddox and Lee (2018). These waves are marked on the
126 individual responses in Figure 1. Mean ± SEM peak latencies for ABR waves I, III, and V were 3.23 ± 0.09
127 ms, 5.51 ± 0.07 ms, and 7.22 ± 0.07 ms respectively. These mean peak latencies are shown superimposed

128   on the group average response in Figure 1 (bottom right). Inter-wave latencies were 2.24 ± 0.06 ms ($N =$

129   19) for I–III, 1.68 ± 0.05 ms ($N =$ 19) for III–V, and 4.00 ± 0.08 ($N =$ 22) for I–V. These peak inter-wave

130   latencies fall within a range expected for brainstem responses but the absolute peak latencies were later

131   than those reported for a click ABR at a level of 60 dB sensation level (SL) and rate between 50 to 100 Hz

132   (Burkard and Hecox, 1983; Chiappa et al., 1979; Don et al., 1977).
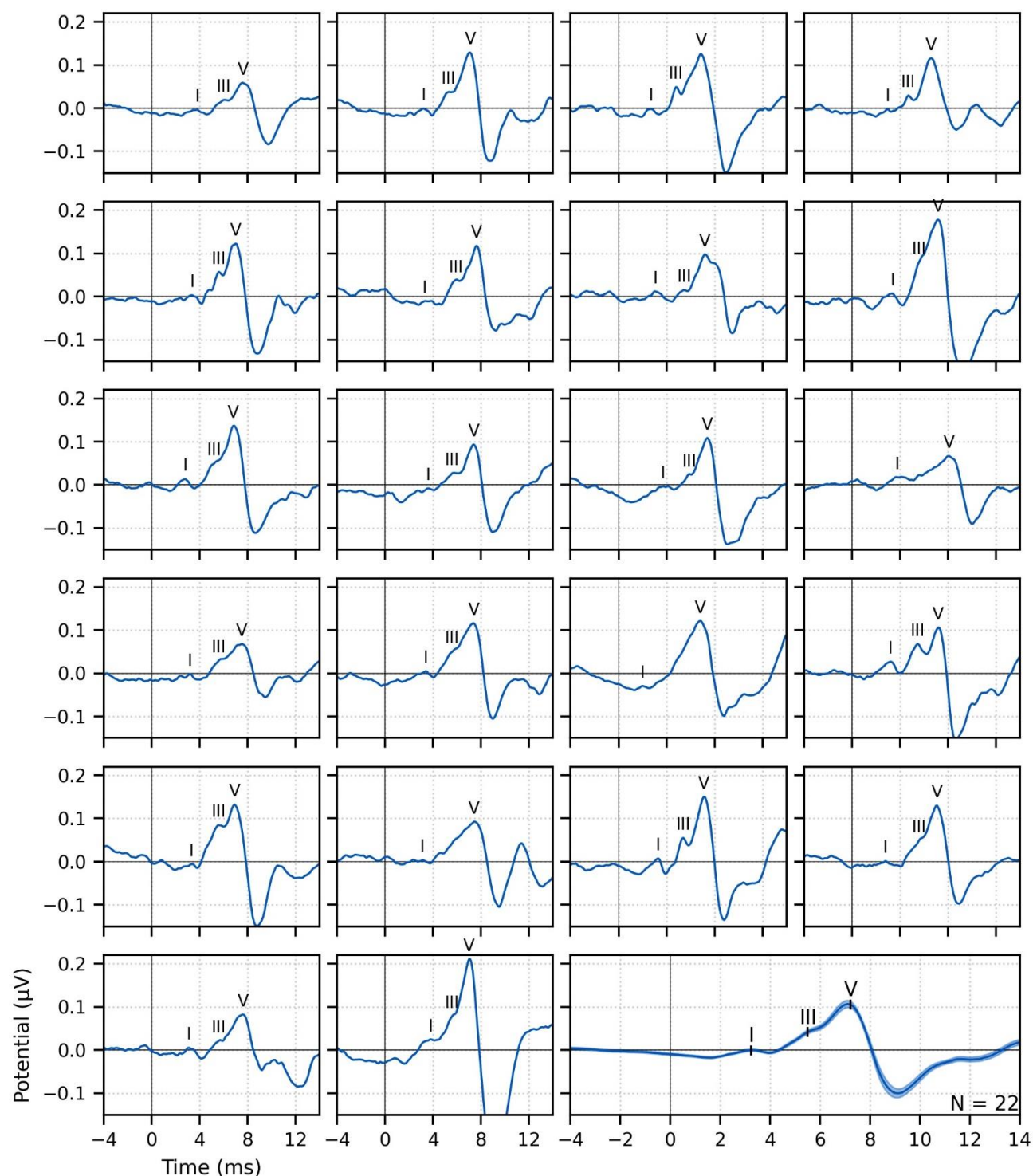
133

Figure 1. *Single subject and group average (bottom right) weighted-average auditory brainstem responses (ABR) to ~43 minutes of broadband peaky speech. Area for the group average shows ± 1 SEM. Responses were high-pass filtered at 150 Hz using a first order Butterworth filter. Waves I, III, and V of the canonical ABR are evident in most of the single subject responses (N = 22, 16, 22 respectively), and are marked by the average peak latencies on the average response.*

134 *More components of the ABR and MLR are present with broadband peaky than unaltered speech*

135   Having established that broadband peaky speech evokes robust canonical ABRs, we next
136 compared both ABR and MLR responses to those evoked by unaltered speech. To simultaneously
137 evaluate ABR and MLR components, a high-pass filter with a 30 Hz cutoff was used on the responses to
138 ~43 minutes of each type of speech. Figure 2A shows that overall there were morphological similarities
139 between responses to both types of speech; however, there were more early and late component waves to
140 broadband peaky speech. More specifically, whereas both types of speech evoked waves V, N_a and P_a,
141 broadband peaky speech also evoked waves I, often III (14–19 of 22 subjects depending if a 30 or 150 Hz
142 high-pass filter cutoff was used), and P_0. With a lower cutoff for the high-pass filter, wave III rode on the
143 slope of wave V and was less identifiable in the grand average shown in Figure 2A than that shown with a
144 higher cutoff in Figure 1. Wave V was more robust and sharper to broadband peaky speech but peaked
145 slightly later than the broader wave V to unaltered speech. For reasons unknown to us, the half-rectified
146 speech method missed the MLR wave P_0, and consequently had a broader and earlier N_a than the
147 broadband peaky speech method, though this missing P_0 was consistent with the results of Maddox and
148 Lee (2018). These waveforms indicate that broadband peaky speech is better than unaltered speech at
149 evoking canonical responses that distinguish activity from distinct subcortical and cortical neural
150 generators.

151   Peak latencies for the waves common to both types of speech are shown in Figure 2B. Again, there
152 was good agreement in peak wave choices for each type of speech, with ICC3 ≥ 0.95 (the lowest two ICC3
153 95% confidence intervals were 0.91–0.98 and 0.94–0.99 for waves V and N_a to unaltered speech
154 respectively). As suggested by the waveforms in Figure 2A, mean ± SEM peak latencies for waves V, N_a,
155 and P_a were longer for broadband peaky than unaltered speech by 0.87 ± 0.06 ms (independents t-test, $t_{(21)}$
156 = 14.7 $p < 0.01$, $d = 3.22$), 6.92 ± 0.44 ms ($t_{(21)} = 15.4$, $p < 0.01$, $d = 3.44$), and 1.07 ± 0.45 ms ($t_{(20)} = 2.3$, p
157 = 0.032, $d = 0.52$) respectively.

158   The response to broadband peaky speech showed a small but consistent response at negative and
159 early positive lags (i.e., pre-stimulus) when using the pulse train as a regressor in the deconvolution,
160 particularly when using the lower high-pass filter cutoff of 30 Hz (Figure 2A) compared to 150 Hz (Figure
161 1). For a perfectly linear causal system this would not be expected. To better understand the source of this
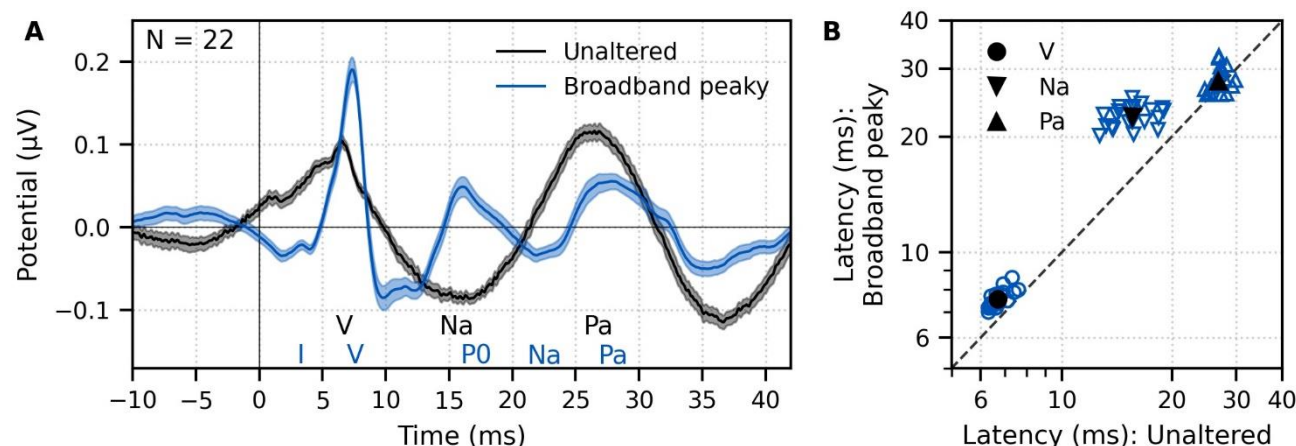

Figure 2. *Comparison of auditory brainstem (ABR) and middle latency responses (MLR) to ~43 minutes each of unaltered speech and broadband peaky speech. (A) The average waveform to broadband peaky speech (blue) shows additional, and sharper, waves of the canonical ABR and MLR than the broader average waveform to unaltered speech (black). Responses were high-pass filtered at 30 Hz with a first order Butterworth filter. Areas show ± 1 SEM. (B) Comparison of peak latencies for ABR wave V (circles) and MLR waves N_a (downward triangles) and P_a (upward triangles) that were common between responses to broadband peaky and unaltered speech. Blue symbols depict individual subjects and black symbols depict the mean.*

162 pre-stimulus component – and to determine whether later components were influencing the earliest
163 components – we completed two models: 1) a simple linear deconvolution model in which EEG for each
164 trial was simulated by convolving the rectified broadband peaky speech audio with an ABR kernel (the
165 average broadband peaky speech ABR was zero-padded to the beginning of wave I, filtered with a Hann
166 window and then normalized); and 2) a more nuanced and well-known model of the auditory nerve and
167 periphery that accounts for acoustics and some of the nonlinearities of the auditory system (Rudnicki et al.,
168 2015; Verhulst et al., 2018; Zilany et al., 2014) – we simulated EEG for waves I, III and V using the
169 framework described by Verhulst et al. (2018), but due to computation constraints, modified the
170 implementation to use the peripheral model by Zilany et al. (2014). The simulated EEG of each model was
171 then deconvolved with the pulse trains to derive the modeled responses, which are shown in comparison
172 with the measured response in Figure 3. The pre-stimulus component of the measured response was
173 present in both models, suggesting that there are nonlinear parts of the response that are not accounted
174 for in the deconvolution with the pulse train regressor. However, the pre-stimulus component was
175 temporally broad compared to the components representing activity from the auditory nerve and cochlear
176 nucleus (i.e., waves I and III), and could thus be dealt with by high-pass filtering. The pre-stimulus
177 component was reduced with a 150 Hz first order Butterworth high-pass filter (Figure 1), and was
178 minimized with a more aggressive 200 Hz second order Butterworth high-pass filter (Figure 3, bottom).
179 Therefore, when doing an experiment where the analysis needs to evaluate specific contributions to the
180 earliest ABR components, we recommend high-pass filtering to help mitigate the complex and time-varying
181 nonlinearities inherent in the auditory system, as well as potential influences by responses from later
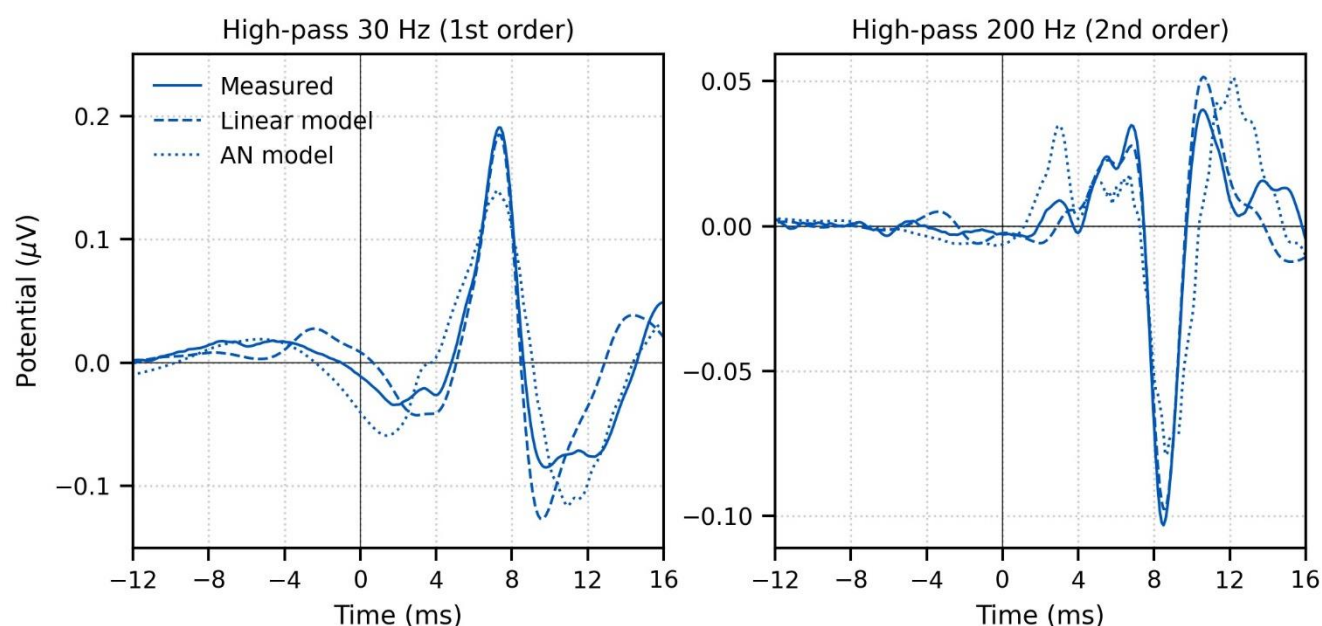182 sources.

183

Figure 3. Comparison of grand average (N = 22) measured and modeled responses to ~43 minutes of broadband peaky speech. Amplitudes of the linear (dashed line) and auditory nerve (AN; dotted line) modelled responses were in arbitrary units, and thus scaled to match the amplitude of the measured response (solid line) over the 0–20 ms lags. The pre-stimulus component was present in all three responses using a first order 30 Hz high-pass Butterworth filter (top row), but was minimized by aggressive high-pass filtering with a second order 200 Hz high-pass Butterworth filter (bottom row).

184   We verified that the EEG data collected in response to broadband peaky speech could be
185 regressed with the half-wave rectified speech to generate a response. Figure 4A shows that the derived
186 responses to unaltered and broadband peaky speech were similar in morphology when using the half-wave

187   rectified audio as the regressor. Correlation coefficients from the 22 subjects for the 0–40 ms lags had a
188   median (interquartile range) of 0.84 (0.72–0.91), which were similar to correlation coefficients obtained by
189   re-analyzing the data split into even/odd epochs that each contained an equal number of epochs with EEG
190   to unaltered speech and peaky broadband speech (0.86, interquartile range 0.80–0.91; Wilcoxon signed-
191   rank test $W_{(21)} = 89.0$, p = 0.235). This means that the same EEG collected to broadband peaky speech
192   can be flexibly used to generate the robust canonical brainstem response to the pulse train, as well as the
193   broader response to the half-wave rectified speech.

194   Although responses to broadband peaky speech can be derived from both the half-wave rectified
195   audio and pulse train regressors, the same cannot be said about unaltered speech. We also regressed the
196   EEG data collected in response to unaltered speech with the pulse train. Figure 4B shows that simply using
197   the pulse train as a regressor does not give a robust canonical response – the response contained a wave
198   that slightly resembled wave V of the response to broadband peaky speech, albeit at a later latency and
199   smaller amplitude, but there were no other earlier waves of the ABR or later waves of the MLR. The
200   correlation coefficients comparing the unaltered and broadband peaky speech responses to the pulse train
201   (median 0.20, interquartile range 0.05–0.39) were significantly poorer than that of even/odd splits of the
202   same EEG (median 0.58, interquartile range 0.25–0.78; $W_{(21)} = 25.0$, p < 0.001).The response morphology
203   was abnormal, with an acausal response at 0 ms and a smearing of the response in time, particularly at
204   lags of the MLR – these effects are consistent with some phases in the unaltered speech that come pre-
205   pulse, which differs from that of the peaky speech which has aligned phases at each glottal pulse. The
206   aligned phases of the broadband peaky response allow for the distinct waves of the canonical brainstem
207   and middle latency responses to be derived using the pulse train regressor.
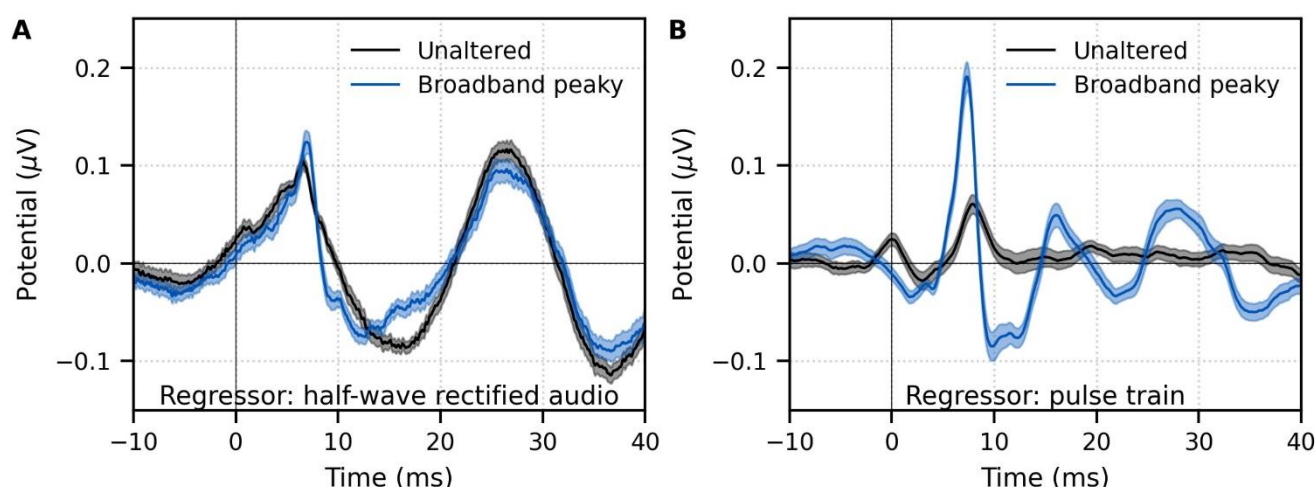
208

Figure 4. *Comparison of responses derived by using the same type of regressor in the deconvolution. Average waveforms (areas show ± 1 SEM) are shown for ~43 minutes each of unaltered speech (black) and broadband peaky speech (blue). EEG was regressed with the (A) half-wave rectified audio and (B) pulse train. Responses were high-pass filtered at 30 Hz using a first order Butterworth filter.*

209

210   Broadband peaky speech responses differ across talkers

211   We next sought to determine whether response morphologies depended on the talker identity.
212   Responses derived from unaltered speech show similar, but not identical, morphology between male and
213   female narrators, indicating some dependence on the stimulus (Maddox and Lee, 2018). To determine
214   what extent the morphology and robustness of peaky speech responses depend on a specific narrator's
215   voice and fundamental frequency – and therefore, rate of stimulation – we compared waveforms and peak
216   wave latencies for 32 minutes (30 epochs of 64 s each) each of male- and female-narrated broadband

217   peaky speech in 11 subjects. The average fundamental frequency was 115 Hz for the male narrator and
218   198 Hz for the female narrator.

219   The group average waveforms to female- and male-narrated broadband peaky speech showed
220   similar canonical morphologies but were smaller and later for female-narrated ABR responses (Figure 5A),
221   much as they would be for click stimuli presented at higher rates (e.g., Burkard et al., 1990; Burkard and
222   Hecox, 1983; Chiappa et al., 1979; Don et al., 1977; Jiang et al., 2009). All component waves of the ABR
223   and MLR were visible in the group average, although fewer subjects exhibited a clear wave III in the
224   female-narrated response (9 versus all 11 subjects). The median (interquartile range) male-female
225   correlation coefficients were 0.68 (0.56–0.78) for ABR lags of 0–15 ms with a 150 Hz high-pass filter, and
226   0.53 (0.50–0.60) for ABR/MLR lags of 0–40 ms with a 30 Hz high-pass filter (Figure 5B). This stimulus
227   dependence was significantly different than the variability introduced by averaging only half the epochs
228   (i.e., splitting by male- and female-narrated epochs) – the correlation coefficients for the data reanalyzed
229   into even and odd epochs (each of the even/odd splits contained the same number of male- and female-
230   narrated epochs), had a median (interquartile range) of 0.89 (0.79–0.95) for ABR lags and 0.66 (0.39–0.78)
231   for ABR/MLR lags. These odd-even coefficients were significantly higher than the male-female coefficients
232   for the ABR ($W_{(10)} = 0.0$, $p = 0.001$; Wilcoxon signed-rank test) but not when the response included all lags
233   of the MLR ($W_{(10)} = 17.0$, $p = 0.175$). which is indicative of the increased variability of these later waves.
234   These overall differences for the ABR indicate that the choice of narrator for using peaky speech impacts
235   the morphology of the early response.

236   As expected from the waveforms, peak latencies of component waves differed between male- and
237   female-narrated broadband peaky speech (Figure 5C). As before, ICC3 ≥ 0.90 indicated good agreement
238   in peak wave choices (the lowest two 95% confidence intervals were 0.79–0.95 for N_a and 0.93–0.99 for I).
239   Mean ± SEM peak latency differences (female – male) for wave I, III, and V of the ABR were 0.21 ± 0.13
240   ms ($t_{(10)} = 1.59$, $p = 0.144$, $d = 0.50$), 0.42 ± 0.11 ms ($t_{(9)} = 3.47$, $p = 0.007$, $d = 1.16$), and 0.54 ± 0.09 ms
241   ($t_{(10)} = 5.56$, $p < 0.001$, $d = 1.76$) respectively. Latency differences were not significant for MLR peaks (P_0:
242   −0.89 ± 0.40 ms, $t_{(9)} = -2.13$, $p = 0.062$, $d = -0.71$; N_a: −0.91 ± 0.55 ms, $t_{(8)} = -1.56$, $p = 0.158$, $d = -0.55$;
243   P_a: 0.09 ± 0.55 ms, $t_{(9)} = -0.16$, $p = 0.880$, $d = -0.05$).
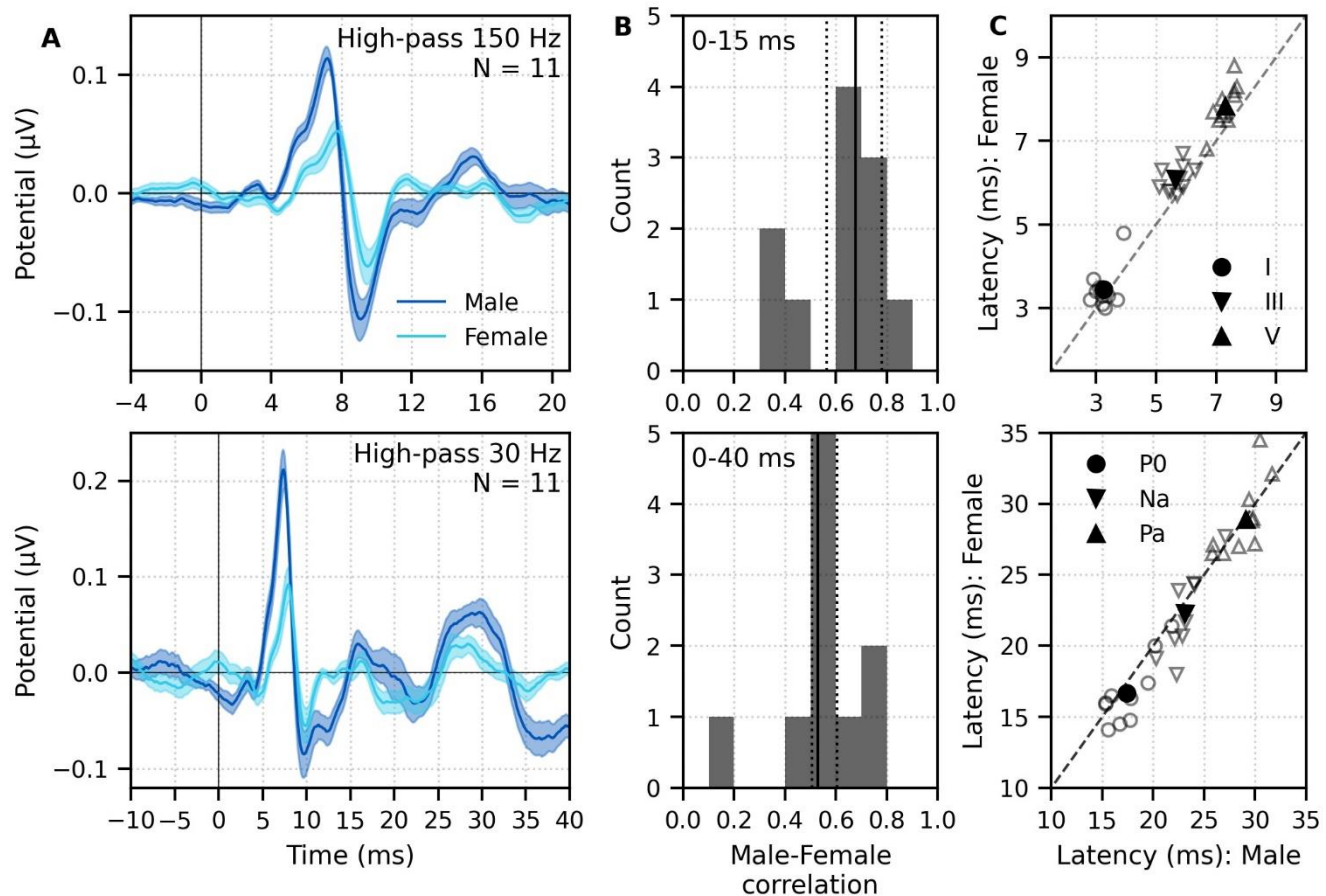
Figure 5. *Comparison of responses to 32 minutes each of male (dark blue) and female (light blue) narrated re-synthesized broadband peaky speech. (A) Average waveforms across subjects (areas show ± 1 SEM) are shown for auditory brainstem response (ABR) time lags with high-pass filtering at 150 Hz (top), and both ABR and middle latency response (MLR) time lags with a lower high-pass filtering cutoff of 30 Hz (bottom). (B) Histograms of the correlation coefficients between responses evoked by male- and female-narrated broadband peaky speech during ABR (top) and ABR/MLR (bottom) time lags. Solid lines denote the median and dotted lines the inter-quartile range. (C) Comparison of ABR (top) and MLR (bottom) wave peak latencies for individual subjects (gray) and the group mean (black). ABR and MLR responses were similar to both types of input but are smaller for female-narrated speech, which has a higher glottal pulse rate. Peak latencies for female-evoked speech were delayed during ABR time lags but faster for early MLR time lags.*

244

245   <u>Multiband peaky speech yields frequency-specific brainstem responses to speech</u>

246   *Frequency-specific responses show frequency-specific lags*

247   Broadband peaky speech gives new insights into subcortical processing of naturalistic speech. Not
248   only are brainstem responses used to evaluate processing at different stages of auditory processing, but
249   ABRs can also be used to assess hearing function across different frequencies. Traditionally, frequency-
250   specific ABRs are measured using clicks with high-pass masking noise or frequency-specific tone pips. We
251   tested the flexibility of using our new peaky speech technique to investigate how speech processing differs
252   across frequency regions, such as 0–1, 1–2, 2–4, and 4–8 kHz frequency bands. To do this, we created
253   new pulse trains with slightly different fundamental waveforms for each filtered frequency regions of
254   speech, and then combined those filtered frequency bands together as multiband speech (for details, see
255   the Multiband peaky speech subsection of Methods). Using this method, we took advantage of the fact that
256   over time, stimuli with slightly different fundamental frequencies will be independent, yielding independent
257   auditory brainstem responses. Therefore, the same EEG was regressed with each band's pulse train to
258   derive the ABR and MLR to each frequency band.

259   Mean ± SEM responses from 22 subjects to the 4 frequency bands (0–1, 1–2, 2–4, and 4–8 kHz) of
260   ~43 minutes of male-narrated multiband peaky speech are shown as colored waveforms with solid lines in

10

261 Figure 6A. A high-pass filter with a cutoff of 30 Hz was used. Each frequency band response comprises a
262 frequency-band-specific component as well as a band-independent common component, both of which are
263 due to spectral characteristics of the stimuli and neural activity. The pulse trains are independent over time
264 in the vocal frequency range – thereby allowing us to pull out responses to each different pulse train and
265 frequency band from the same EEG – but they became coherent at frequencies lower than 72 Hz for the
266 male-narrated speech and 126 Hz for the female speech (see Figure 15 in Methods). This coherence was
267 due to all pulse trains beginning and ending together at the onset and offset of voiced segments and was
268 the source of the low-frequency common component of each band's response. The way to remove the
269 common component is to calculate the common activity across the frequency band responses and subtract
270 this waveform from each of the frequency band responses. This common component was calculated by
271 regressing the EEG to multiband speech with 6 independent "fake" pulses trains – pulse trains with slightly
272 different fundamental frequencies that were not used to create the multiband peaky speech stimuli that
273 were presented during the experiment – and then averaging across these 6 responses. This common
274 component waveform is shown by the dot-dashed gray line, which is superimposed with each response to
275 the frequency bands in Figure 6A. The subtracted, frequency-specific waveforms to each frequency band
276 are shown by the solid lines in Figure 6B. Of course, the subtracted waveforms could also then be high-
277 pass filtered at 150 Hz to highlight earlier waves of the brainstem responses, as shown by the dashed lines
278 in Figure 6B. However, this method reduces the amplitude of the responses, which in turn affects response
279 SNR and detectability. In some scenarios, due to the low-pass nature of the common component, high-
280 passing the waveforms at a high enough frequency may obviate the need to formally subtract the common
281 component. For example, at least for the narrators used in these experiments, the common component
282 contained minimal energy above 150 Hz, so if the waveforms are already high-passed at 150 Hz to focus
283 on the early waves of the ABR, then the step of formally subtracting the common component may not be
284 necessary. But beyond the computational cost, there is no reason not to subtract the common component,
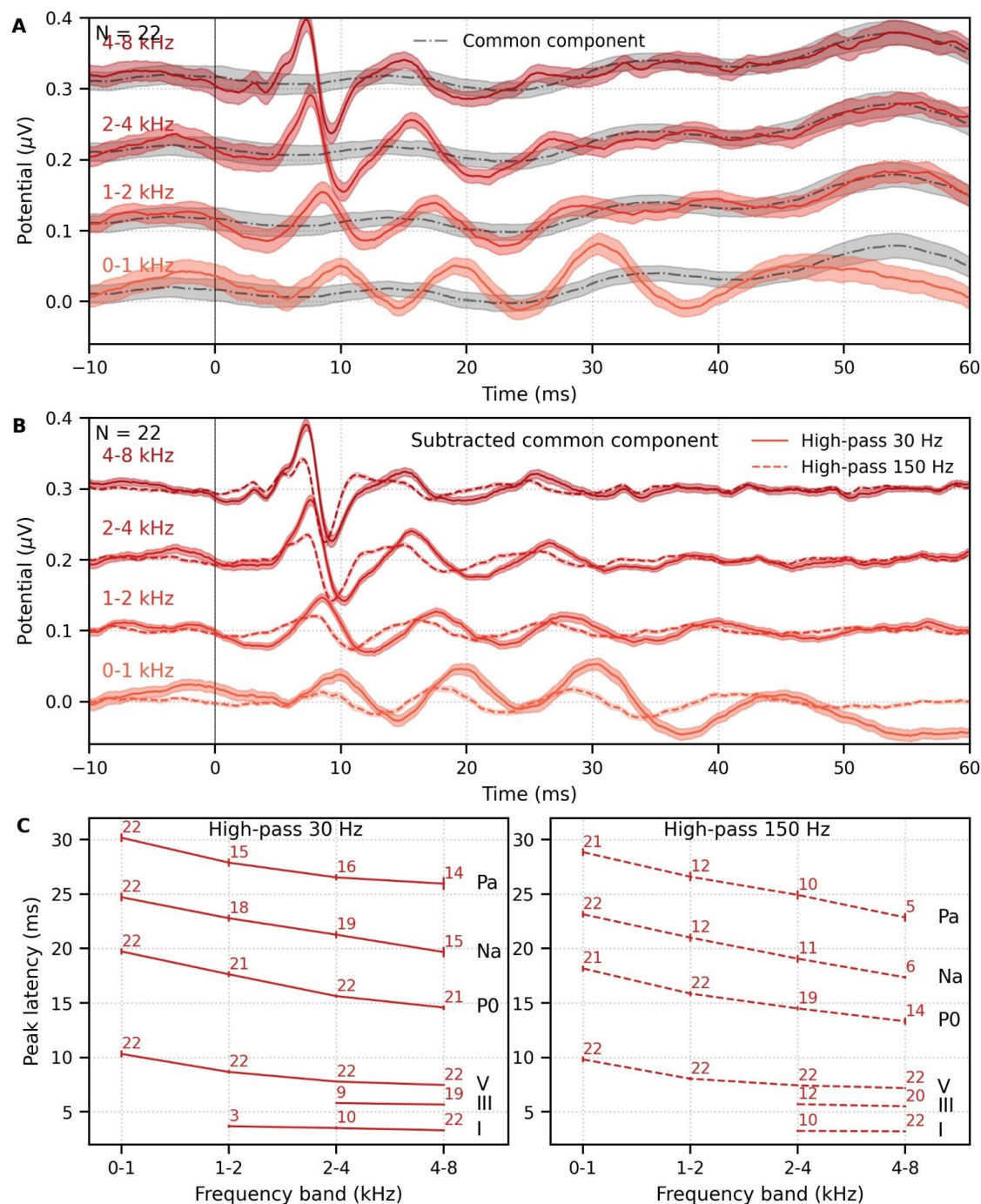285 and doing so allows lower filter cut-offs to be used.

Figure 6. *Comparison of responses to ~43 minutes of male-narrated multiband peaky speech. (A) Average waveforms across subjects (areas show ± 1 SEM) are shown for each band (colored solid lines) and for the common component (dot-dash gray line, same waveform replicated as a reference for each band), which was calculated using 6 false pulse trains. (B) The common component was subtracted from each band's response to give the frequency-specific waveforms (areas show ± 1 SEM), which are shown with high-pass filtering at 30 Hz (solid lines) and 150 Hz (dashed lines). (C) Mean ± SEM peak latencies for each wave decreased with increasing band frequency. Numbers of subjects with an identifiable wave are given for each wave and band. Details of the mixed effects models for (C) are provided in Supplementary File 1A.*

286

287   Overall, the frequency-specific responses showed characteristic ABR and MLR waves with longer
288 latencies for lower frequency bands, as would be expected from responses arising from different cochlear
289 regions. Also, waves I and III of the ABR were visible in the group average waveforms of the 2–4 kHz
290 (≥41% of subjects) and 4–8 kHz (≥86% of subjects) bands, whereas the MLR waves were more prominent
291 in the 0–1 kHz (≥95% of subjects) and 1–2 kHz (≥54% of subjects) bands.

292   These frequency-dependent latency changes for the frequency-specific responses are highlighted
293   further in Figure 6C, which shows mean ± SEM peak latencies and the number of subjects who had a
294   clearly identifiable wave. ICC3 ≥ 0.89 indicated good agreement in peak wave choices (lowest two 95%
295   confidence intervals were 0.84–0.93 for P_a and 0.90–0.95 for N_a). The nonlinear change in peak latency, τ,
296   with frequency band was modeled using power law regression according to the formula (Harte et al., 2009;
297   Neely et al., 1988; Rasetshwane et al., 2013; Strelcyk et al., 2009):

298
$$\tau(f) = a + bf^{-d}$$

299   where

300
$$a = \tau_{synaptic} + \tau_{I-V},$$

301   and where f is the band center frequency normalized to 1 kHz (i.e., divided by 1000), $\tau_{synaptic}$ is the
302   synaptic delay (assumed to be 0.8; Eggermont, 1979; Strelcyk et al., 2009), and τ_{I-V} is the I-V inter-wave
303   delay from the subjects' responses to broadband peaky speech. The power law model was completed for
304   each filter cutoff of 30 and 150 Hz in the log-log domain, according to the formula:

305
$$\log_{10}(\tau(f) - a) = \log_{10} b + (-d) \log_{10} f$$

306   and using linear mixed effects regression to estimate the terms d and b (calculated as $b = 10^{intercept}$). The
307   fixed effect of wave and its interaction with frequency were added to the model, along with random effects
308   of intercept and frequency for each subject. For the 30 and 150 Hz filter cutoffs, the mean ± SEM τ_{I-V} was
309   4.33 ± 0.08 and 4.00 ± 0.08 respectively, resulting in average estimates of 5.13 and 4.80 for a respectively
310   – both of which are similar to the post-cochlear delays of 4.7–5.0 ms reported for tone pips and derived-
311   bands from clicks (Neely et al., 1988; Rasetshwane et al., 2013; Strelcyk et al., 2009). There were
312   insufficient numbers of subjects with identifiable waves I and III for the 0–1 kHz and 1–2 kHz bands, so
313   these waves were not included in the full model. Details of each log-log mixed effects model are described
314   in Supplementary File 1A. For wave V, the estimated mean parameters were $b = 3.95$ ms and $d = -0.41$ for
315   a 30 Hz filter cutoff, and $b = 3.95$ ms and $d = -0.37$ for a 150 Hz filter cutoff, which also fell within the
316   corresponding ranges of 3.46–5.39 ms for b (calculated for a level of 65 dB ppeSPL using the models' level
317   dependent parameter) and 0.22–0.50 for $-d$ that were previously reported for tone pips and derived-bands
318   (Neely et al., 1988; Rasetshwane et al., 2013; Strelcyk et al., 2009). As expected, there were significantly
319   different latencies for each MLR wave P_0, N_a and P_a compared to the ABR wave V (all effects of wave on
320   the intercept $p < 0.001$, for each high-pass filter cutoff of 30 and 150 Hz). The significant decrease in
321   latency with frequency band (slope, d : $p < 0.001$ for 30 and 150 Hz) was shallower (i.e., less negative) for
322   MLR waves compared to the ABR wave V (all $p < 0.001$ for interactions between wave and frequency term
323   for 30 and 150 Hz).

324   Next, the frequency-specific responses (i.e., multiband responses with common component
325   subtracted) were summed and the common component added to derive the entire response to multiband
326   peaky speech. As shown in Figure 7, this summed multiband response was strikingly similar in morphology
327   to the broadband peaky speech. Both responses were high-passed filtered at 150 Hz and 30 Hz to
328   highlight the earlier ABR waves and later MLR waves, respectively. The median (interquartile range)
329   correlation coefficients from the 22 subjects were 0.90 (0.84–0.94) for 0–15 ms ABR lags, and 0.66 (0.55–
330   0.83) for 0–40 ms MLR lags. The similarity verifies that the frequency-dependent responses are
331   complementary to each other to the common component, such that these components add linearly into a
332   "whole" broadband response. If there were significant overlap in the cochlear regions, for example, the
333   summed response would not resemble the broadband response to such a degree, and would instead be
334   larger. The similarity also verified that the additional changes we made to create re-synthesized multiband
335   peaky speech did not significantly affect responses compared to broadband peaky speech.
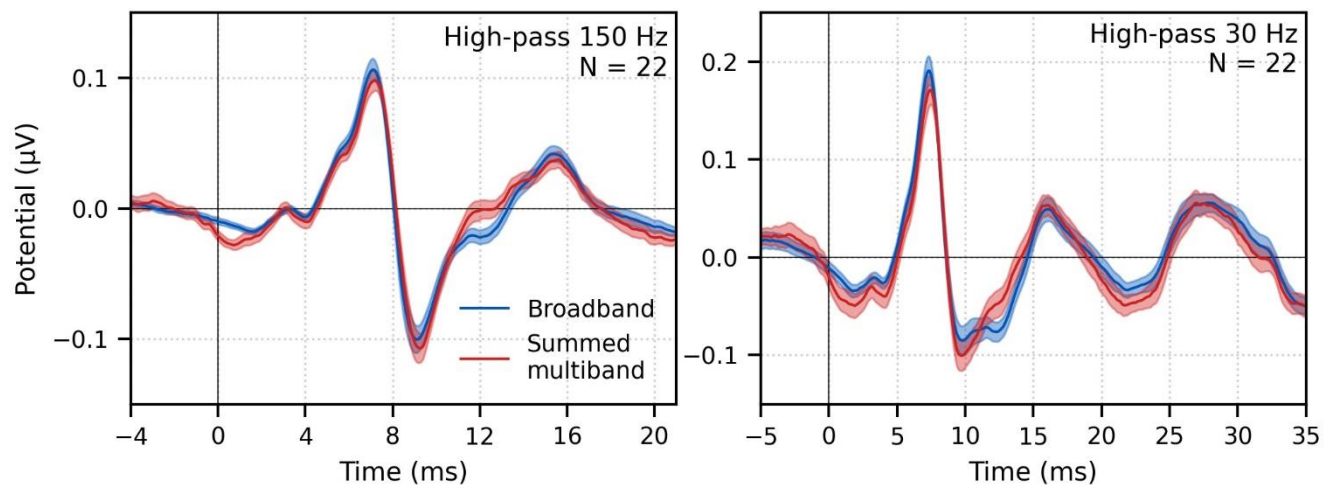
13

Figure 7. *Comparison of responses to ~43 minutes of male-narrated peaky speech in the same subjects. Average waveforms across subjects (areas show ± 1 SEM) are shown for broadband peaky speech (blue) and for the summed frequency-specific responses to multiband peaky speech with the common component added (red), high-pass filtered at 150 Hz (left) and 30 Hz (right). Regressors in the deconvolution were pulse trains.*

We also verified that the EEG data collected in response to multiband peaky speech could be regressed with the half-wave rectified speech to generate a response. Figure 4-figure supplement 1 shows that the derived responses to unaltered and multiband peaky speech were similar in morphology when using the half-wave audio as the regressor, although the multiband peaky speech response was broader in the earlier latencies. Correlation coefficients from the 22 subjects for the 0–40 ms lags had a median (interquartile range) of 0.83 (0.77–0.88), which were slightly smaller than the correlation coefficients obtained by re-analyzing the data split into even/odd epochs that each contained an equal number of epochs with EEG to unaltered speech and peaky broadband speech (0.89, interquartile range 0.83–0.94; Wilcoxon signed-rank test $W_{(21)} = 58.0$, p = 0.025). This means that the same EEG collected to multiband peaky speech can be flexibly used to generate the frequency-specific brainstem responses to the pulse train, as well as the broader response to the half-wave rectified speech.

Frequency-specific responses also differ by narrator

We also investigated the effects of narrator on multiband peaky speech by deriving responses to 32 minutes (30 epochs of 64 s each) each of male- and female-narrated multiband peaky speech in the same 11 subjects. As with broadband peaky speech, responses to both narrators showed similar morphology, but the responses were smaller and the MLR waves more variable for the female than male narrator (Figure 8A). Figure 8B shows the male-female correlation coefficients for responses between 0–40 ms with a high-pass filter of 30 Hz and between 0–15 ms with a high-pass filter of 150 Hz. The median (interquartile range) male-female correlation coefficients were better for higher frequency bands, ranging from 0.12 (−0.11–0.34) for the 1–2 kHz band to 0.44 (0.32–0.49) for the 4–8 kHz band for MLR lags (Figure 8B, left panel), and from 0.49 (0.30–0.69) for the 1–2 kHz band to 0.79 (0.64–0.81) for the 4–8 kHz band for ABR lags (Figure 8B, right panel). These male-female correlation coefficients were significantly weaker than those of the same EEG split into even and odd trials for all but the 2–4 kHz frequency band when responses were high-pass filtered at 30 Hz and correlated across 0–40 ms lags (2–4 kHz: $W_{(10)} = 17.0$, $p = 0.175$; other bands: $W_{(10)} \leq 5.0$, $p \leq 0.010$), but were similar to the even/odd trials for responses from all frequency bands high-pass filtered at 150 Hz ($W_{(10)} \geq 11.0$, $p \geq 0.054$). These results indicate that the specific narrator can affect the robustness of frequency-specific responses, particularly for the MLR waves.
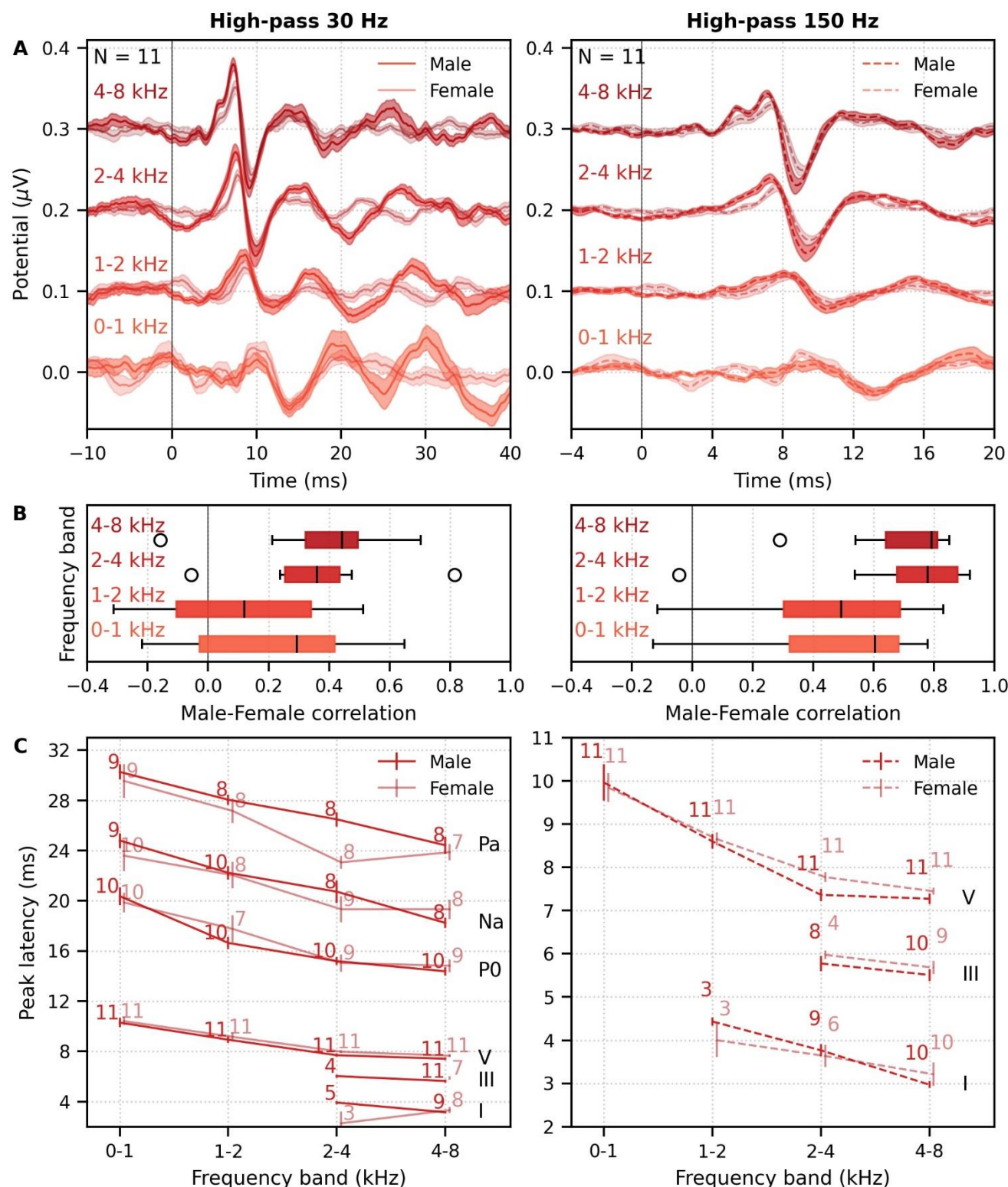
Figure 8. Comparison of responses to 32 minutes each of male- and female-narrated re-synthesized multiband peaky speech. *(A) Average frequency-specific waveforms across subjects (areas show ± 1 SEM; common component removed) are shown for each band in response to male- (dark red lines) and female-narrated (light red lines) speech. Responses were high-pass filtered at 30 Hz (left) and 150 Hz (right) to highlight the MLR and ABR respectively. (B) Correlation coefficients between responses evoked by male- and female-narrated multiband peaky speech during ABR/MLR (left) and ABR (right) time lags for each frequency band. Black lines denote the median. (C) Mean ± SEM peak latencies for male- (dark) and female- (light) narrated speech for each wave decreased with increasing frequency band. Numbers of subjects with an identifiable wave are given for each wave, band and narrator. Lines are given a slight horizontal offset to make the error bars easier to see. Details of the mixed effects models for (C) are provided in Supplementary File 1B.*

365

366   As expected from the grand average waveforms and male-female correlations, there were fewer
367 subjects who had identifiable waves across frequency bands for the female- than male-narrated speech.
368 These numbers are shown in Figure 8C, along with the mean ± SEM peak latencies for each wave,
369 frequency band and narrator. ICC3 ≥0.91 indicated good agreement in peak latency choices (the two
370 lowest 95% confidence intervals were 0.82–0.96 for wave III and 0.98–0.99 for P_a, both with a high-pass
371 filter cutoff of 30 Hz). Again, there were few numbers of subjects with identifiable waves I and III for the
372 lower frequency bands. Therefore, a power law model (see previous subsection for the formula) was
373 completed for waves V, P_0, N_a, and P_a of responses in the 4 frequency bands that were high-pass filtered at
374 30 Hz. The model was completed in the log-log domain with linear mixed effect regression to estimate the
375 terms d and b (calculated as $b = 10^{intercept}$). The mixed effects model included fixed effects of narrator,
376 wave, logged center frequency that was normalized to 1 kHz, the interaction between narrator and wave
377 and between narrator and frequency, and the interactions between wave and frequency, as well as a
378 random intercept and frequency term per subject. The mean ± SEM τ_{I-V} for the male and female narrators
379 was 4.26 ± 0.14 and 4.78 ± 0.12 respectively, resulting in average estimates of 5.06 and 5.58 for a
380 respectively. The value of a for the male narrator was similar to that obtained in the first cohort of 22
381 subjects and was consistent with previously reported values for tone pips and derived-bands from clicks
382 (Neely et al., 1988; Rasetshwane et al., 2013; Strelcyk et al., 2009). The a estimate for the female narrator
383 was slightly longer than the previously reported values, although a lower filter cut-off of 30 Hz was used
384 than the 100 Hz in the previous studies. Filtering the female-narrated responses at 150 Hz gave an
385 estimate for a of 5.18, which is closer to the corresponding range of 4.7–5.0 ms than 5.58. Details of the
386 log-log mixed effects model are described in Supplementary File 1B. For wave V, the estimated mean
387 parameters were $b = 4.42$ ms and $d = -0.45$ for the male narrator, and $b = 3.94$ ms and $d = -0.44$ for the
388 female narrator. The b, but not d, parameter was significantly different between narrators (narrator main
389 effect, $p = 0.001$; narrator interaction with frequency: $p = 0.732$), and the parameters were within the
390 ranges previously reported for tone pips and derived-bands at 65 dB ppeSPL (Neely et al., 1988;
391 Rasetshwane et al., 2013; Strelcyk et al., 2009). For those subjects with identifiable waves, peak latencies
392 shown in Figure 8C differed by wave ($p < 0.001$ for effects of each wave on the intercept), and latency
393 decreased with increasing frequency (slope, d: $p < 0.001$). This change with frequency was smaller (i.e.,
394 shallower slope) for each MLR wave compared to wave V ($p < 0.001$ for all interactions between wave and
395 the term for frequency band). There was a main effect of narrator on peak latencies but no interaction with
396 wave (narrator $p = 0.001$, wave-narrator interactions $p > 0.087$). Therefore, as with broadband peaky
397 speech, frequency-specific peaky responses were more robust with the male narrator and the frequency-
398 specific responses peaked earlier for a narrator with a lower fundamental frequency.

399

400 *Frequency-specific responses can be measured simultaneously in each ear (dichotically)*

401   We have so far demonstrated the effectiveness of peaky speech for diotic stimuli, but there is often
402 a need to evaluate auditory function in each ear, and the most efficient tests assess both ears
403 simultaneously. Applying this principle to generate multiband peaky speech, we investigated whether ear-
404 specific responses could be evoked across the 5 standard audiological octave frequency bands (500–8000
405 Hz) using dichotic multiband speech. We created 10 independent pulse trains, 2 for each ear in each of the
406 5 frequency bands (see Multiband peaky speech and Band filters in Methods).

407   We recorded responses to 64 minutes (60 epochs of 64 s each) each of male- and female-narrated
408 dichotic multiband peaky speech in 11 subjects. The frequency-specific (i.e., common component-
409 subtracted) group average waveforms for each ear and frequency band are shown in Figure 9A. The ten
410 waveforms were small, especially for female-narrated speech, but a wave V was identifiable for both
411 narrators. Also, waves I and III of the ABR were visible in the group average waveforms of the 4 kHz band
412 (≥45% and 18% of subjects for the male and female narrator respectively) and 8 kHz band (≥90% and 72%

413   of subjects for the male and female narrator respectively). MLR waves were not clearly identifiable for

414   responses to female-narrated speech. Therefore, correlations between responses were performed for ABR

415   lags between 0–15 ms. As shown in Figure 9B, the median (interquartile range) left-right ear correlation

416   coefficients (averaged across narrators) ranged from 0.28 (0.02–0.52) for the 0.5 kHz band to 0.73 (0.62–

417   0.84) for the 8 kHz band. Male-female correlation coefficients (averaged across ear) ranged from 0.44

418   (0.00–0.58) for the 0.5 kHz band to 0.76 (0.51–0.80) for the 8 kHz band. Although the female-narrated

419   responses were smaller than the male-narrated responses, these male-female coefficients did not

420   significantly differ from correlations of same EEG split into even-odd trials and averaged across ear ($W_{(10)} \geq$

421   12.0, $p \geq 0.067$), likely reflecting the variability in such small responses.

422   Figure 9C shows the mean ± SEM peak latencies of wave V for each ear and frequency band for

423   the male- and female-narrated dichotic multiband peaky speech. The ICC3 for wave V was 0.97 (95%

424   confidence interval 0.97–0.98), indicating reliable peak latency choices. The nonlinear change in wave V

425   latency with frequency was modeled by a power law using log-log mixed effects regression with fixed

426   effects of narrator, ear, logged band center frequency normalized to 1 kHz, and the interactions between

427   narrator and frequency. Random effects included an intercept and frequency term for each subject. The

428   experiment 2 average estimates for a of 5.06 and 5.58 for the male- and female-narrated responses were

429   used for each subject. Details of the model are described in Supplementary File 1C. For wave V, the

430   estimated mean parameters were $b = 4.13$ ms (calculated as $10^{intercept}$) and $d = -0.36$ for the male narrator,

431   and $b = 4.25$ ms and $d = -0.41$ for the female narrator, which corresponded to previously reported ranges

432   for tone pips and derived-bands at 65 dB ppeSPL (Neely et al., 1988; Rasetshwane et al., 2013; Strelcyk et

433   al., 2009). Latency decreased with increasing frequency (slope, d, $p < 0.001$) but did not differ between

434   ears ($p = 0.265$). The b parameter differed by narrator but the slope did not change with narrator

435   (interaction with intercept $p = 0.004$, interaction with slope $p = 0.085$). Taken together, these results confirm

436   that, while small in amplitude, frequency-specific responses can be elicited in both ears across 5 different

437   frequency bands and show characteristic latency changes across the different frequency bands.
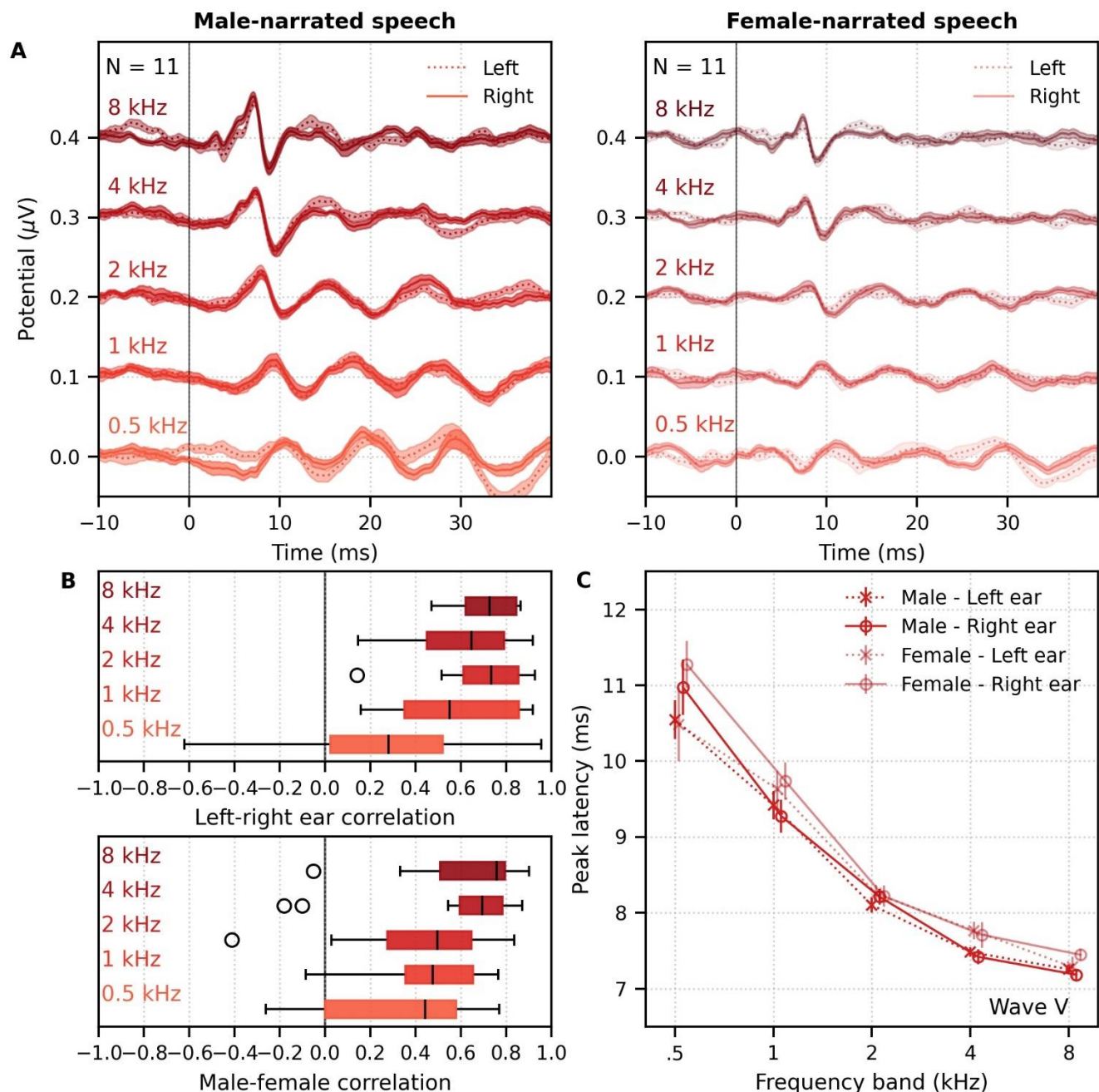
17

Figure 9. *Comparison of responses to ~60 minutes each of male- and female-narrated dichotic multiband peaky speech with standard audiological frequency bands. (A) Average frequency-specific waveforms across subjects (areas show ± 1 SEM; common component removed) are shown for each band for the left ear (dotted lines) and right ear (solid lines). Responses were high-pass filtered at 30 Hz. (B) Left-right ear correlation coefficients (top, averaged across gender) and male-female correlation coefficients (bottom, averaged across ear) during ABR time lags (0–15 ms) for each frequency band. Black lines denote the median. (C) Mean ± SEM wave V latencies for male- (dark red) and female-narrated (light red) speech for the left (dotted line, cross symbol) and right ear (solid line, circle symbol) decreased with increasing frequency band. Lines are given a slight horizontal offset to make the error bars easier to see. Details of the mixed effects model for (C) are provided in Supplementary File 1C.*

438

439

440 <u>Responses are obtained quickly for male-narrated broadband peaky speech but not multiband speech</u>

441   Having demonstrated that peaky broadband and multiband speech provides canonical waveforms
442 with characteristic changes in latency with frequency, we next evaluated the acquisition time required for
443 waveforms to reach a decent SNR. We chose 0 dB SNR based on visual assessment of when waveforms
444 were easily inspected and based on what we have done previously (Maddox and Lee, 2018; Polonenko

18

445 and Maddox, 2019). SNR was calculated by comparing the variance in the MLR time interval 0–30 ms (for
446 responses high-pass filtered at 30 Hz) or ABR time interval 0–15 ms (for responses high-pass filtered at
447 150 Hz) to the variance in the pre-stimulus noise interval −480 to −20 ms (see Response SNR calculation
448 in Methods for details).

449 Figure 10 shows the cumulative proportion of subjects who had responses with ≥ 0 dB SNR to
450 unaltered and broadband peaky speech as a function of recording time. Acquisition times for 22 subjects
451 were similar for responses to both unaltered and broadband peaky male-narrated speech, with 0 dB SNR
452 achieved by 7–8 minutes in 50% of subjects and about 20 minutes for all subjects. For responses high-
453 pass filtered at 150 Hz to highlight the ABR (0–15 ms interval), the time reduced to 2 and 5 minutes for
454 50% and 100% of subjects respectively for broadband peaky speech but increased to 8 and >20 minutes
455 for 50% and 100% of subjects respectively for unaltered speech. The increased time for the unaltered
456 speech reflects the broad morphology of the response during ABR lags. These times for male-narrated
457 broadband peaky speech were confirmed in our second cohort of 11 subjects. However, acquisition times
458 were at least 2.1 times – but in some cases over 10 times – longer for female-narrated broadband peaky
459 speech. In contrast to male-narrated speech, not all subjects achieved this threshold for female-narrated
460 speech by the end of the 32-minute recording. Taken together, these acquisition times confirm that
461 responses with useful SNRs can be measured quickly for male-narrated broadband peaky speech but
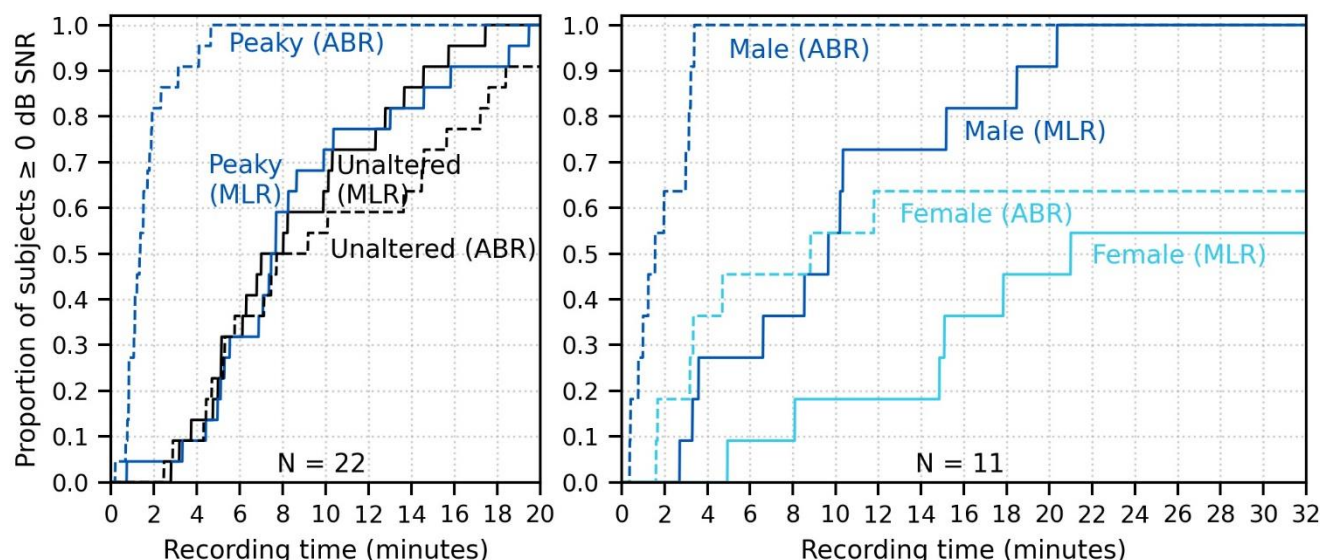462 longer recording sessions are necessary for narrators with higher fundamental frequencies.

Figure 10. *Cumulative proportion of subjects who have responses with ≥ 0 dB SNR as a function of recording time. Time required for unaltered (black) and broadband peaky speech (dark blue) of a male narrator are shown for 22 subjects in the left plot, and for male (dark blue) and female (light blue) broadband peaky speech is shown for 11 subjects in the right plot. Solid lines denote SNRs calculated using variance of the signal high-pass filtered at 30 Hz over the ABR/MLR interval 0–30 ms, and dashed lines denote SNR variances calculated on signals high-pass filtered at 150 Hz over the ABR interval 0–15 ms. Noise variance was calculated in the pre-stimulus interval −480 to −20 ms.*

463

464 The longer recording times necessary for a female narrator became more pronounced for the
465 multiband peaky speech. Figure 11A shows the cumulative density function for responses high-pass-
466 filtered at 150 Hz and the SNR estimated over the ABR interval. Many subjects (72%) had frequency-
467 specific responses (common component subtracted) with ≥ 0 dB SNR for all 4 frequency bands by the end
468 of the 32-minute recording for the male-narrated speech, but this was achieved in only 45% of subjects for
469 the female-narrated speech. Multiband peaky speech required significantly longer recording times than
470 broadband peaky speech, with 50% of subjects achieving 0 dB SNR by 15 minutes compared to 2 minutes
471 for the male-narrated responses across the ABR 0–15 ms interval and 17 minutes compared to 5 minutes
472 for the MLR 0–30 ms interval. Even more time was required for dichotic multiband speech, which was
473 comprised of a larger number of frequency bands (Figure 11B). All 10 audiological band responses

19

474 achieved ≥0 dB SNR in 45% of ears (10 / 22 ears from 11 subjects) by 64 minutes for male-narrated
475 speech and in 27% of ears (6 / 22 ears) for female-narrated speech. The smaller and broader responses in
476 the low frequency bands were slower to obtain and were the main constraint on testing time. These
477 recording times suggest that deriving multiple frequency-specific responses will require at least more than
478 30 minutes per condition for < 5 bands, and more than an hour session for one condition of peaky
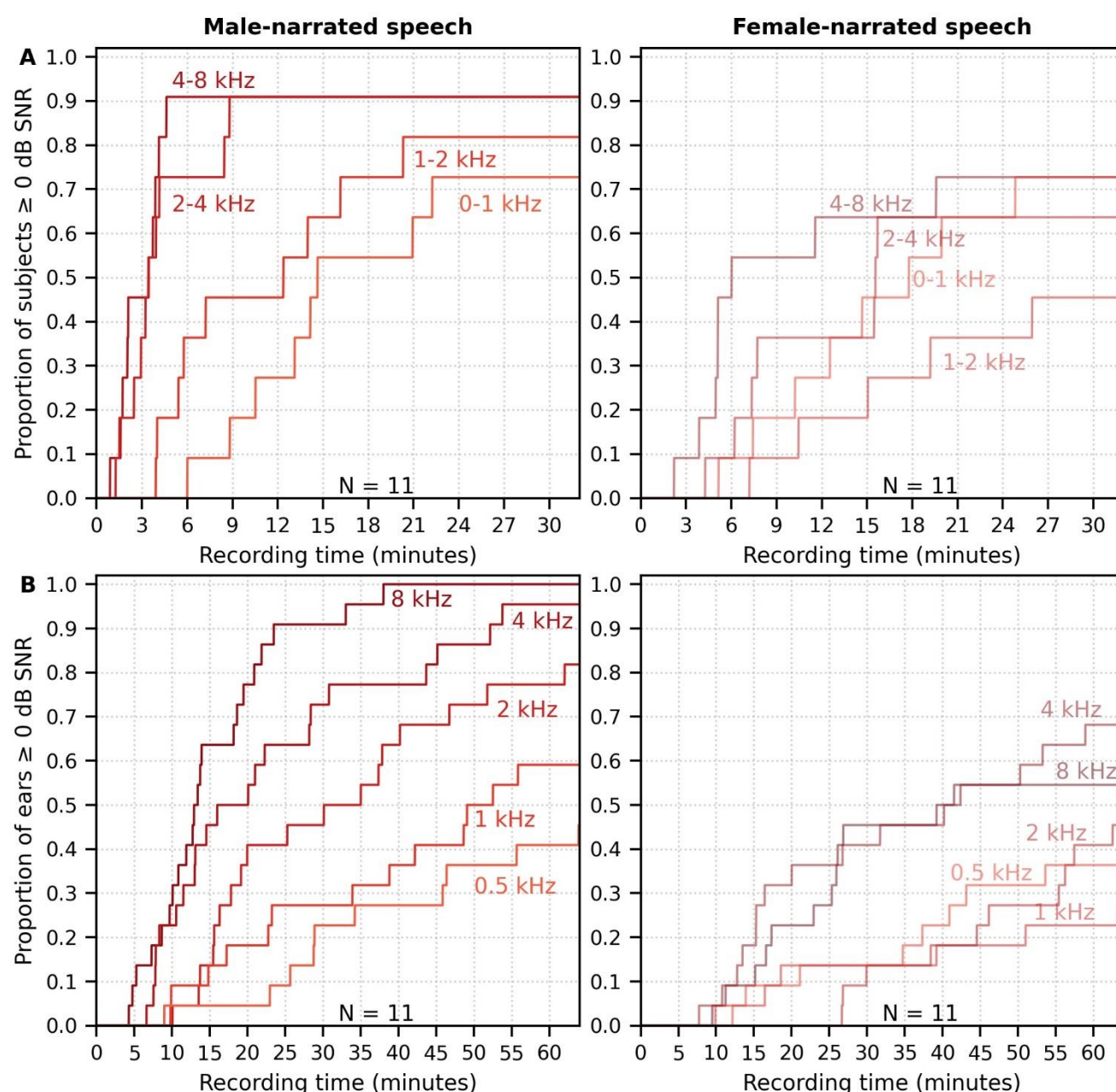479 multiband speech with 10 bands.

Figure 11. *Cumulative proportion of subjects who have frequency-specific responses (common component subtracted) with ≥ 0 dB SNR as a function of recording time. Acquisition time was faster for male (left) than female (right) narrated multiband peaky speech with (A) 4 frequency bands presented diotically, and with (B) 5 frequency bands presented dichotically (total of 10 responses, 5 bands in each ear). SNR was calculated by comparing variance of signals high-pass filtered at 150 Hz across the ABR interval of 0–15 ms to variance of noise in the pre-stimulus interval −480 to −20 ms.*

480

481

DISCUSSION

The major goal of this work was to develop a method to investigate early stages of naturalistic speech processing. We re-synthesized continuous speech taken from audio books so that the phases of all harmonics aligned at each glottal pulse during voiced segments, thereby making speech as impulse-like (peaky) as possible to drive the auditory brainstem. Then we used the glottal pulse trains as the regressor in deconvolution to derive the responses. Indeed, comparing waveforms to broadband peaky and unaltered speech validated the superior ability of peaky speech to evoke additional waves of the canonical ABR and MLR, reflecting neural activity from multiple subcortical structures. Robust ABR and MLR responses were recorded in less than 5 and 20 minutes respectively for all subjects, with half of the subjects exhibiting a strong ABR within 2 minute and MLR within 8 minutes. Longer recording times were required for the smaller responses generated by a narrator with a higher fundamental frequency. We also demonstrated the flexibility of this stimulus paradigm by simultaneously recording up to 10 frequency-specific responses to multiband peaky speech that was presented either diotically or dichotically, although these responses required much longer recording times. Taken together, our results show that peaky speech effectively yields responses from distinct subcortical structures and from different frequency bands, paving the way for new investigations of speech processing and new tools for clinical application.

<u>Peaky speech responses reflect activity from distinct subcortical components</u>

Canonical responses can be derived from speech with impulse-like characteristics

For the purpose of investigating responses from different subcortical structures, we accomplished our goal of creating a stimulus paradigm that overcame some of the limitations of current methods using natural speech. Methods that do not use re-synthesized impulse-like speech generate responses characterized by a broad peak between 6–9 ms (Forte et al., 2017; Maddox and Lee, 2018), with contributions predominantly from the inferior colliculus (Saiz-Alia and Reichenbach, 2020). In contrast, for the majority of our subjects, peaky speech evoked responses with canonical morphology comprised of waves I, III, V, P_0, N_a, P_a (Figure 1), reflecting neural activity from distinct stages of the auditory system from the auditory nerve to thalamus and primary auditory cortex (e.g., Picton et al., 1974). Although clicks also evoke responses with multiple component waves, current studies of synaptopathy show quite varied results with clicks and poor correlation with speech in noise (Bramhall et al., 2019; Prendergast et al., 2017). Obtaining click-like responses to stimuli with all of speech's spectrotemporal richness may provide a better connection to the specific neural underpinnings of speech encoding, similar to how the complex cross-correlation to a fundamental waveform can change based on the context of attention (Forte et al., 2017; Saiz-Alía et al., 2019).

The same ABR waves evoked here were also evoked by a method using embedded chirps intermixed within alternating octave bands of speech, particularly if presented monaurally over headphones instead of in free field (Backer et al., 2019; Miller et al., 2017). Chirps are transients that compensate for the cochlear traveling delay wave by introducing different phases across frequency, leading to a more synchronized response across the cochlea and a larger brainstem response than for clicks (Dau et al., 2000; Elberling and Don, 2008; Shore and Nuttall, 1985). The responses to embedded chirps elicited waves with larger mean amplitude than those to our broadband peaky speech (~0.4 versus ~0.2 µV, respectively), although a similar proportion of subjects had identifiable waves, the SNR was good, and several other factors may contribute to amplitude differences. For example, higher click rates (e.g., Burkard et al., 1990; Burkard and Hecox, 1983; Chiappa et al., 1979; Don et al., 1977; Jiang et al., 2009) and higher fundamental frequencies (Maddox and Lee, 2018; Saiz-Alía et al., 2019; Saiz-Alia and Reichenbach, 2020) reduce the brainstem response amplitude, and dynamic changes in rate may create interactions across neural populations that lead to smaller amplitudes. Our stimuli kept the dynamic changes in pitch across all frequencies (instead of alternate octave bands of chirps and speech) and created impulses at every glottal pulse, with an average pitch of ~115 Hz and ~198 Hz for the male and female narrators respectively. These presentation rates were much

529 higher and more variable than the flat 42 Hz rate at which the embedded chirps were presented (pitch
530 flattened to 82 Hz and chirps presented every other glottal pulse). We could evaluate whether chirps would
531 improve response amplitude to our dynamic peaky speech by simply all-pass filtering the re-synthesized
532 voiced segments by convolving with a chirp prior to mixing the re-synthesized parts with the unvoiced
533 segments. While maintaining the amplitude spectrum of speech, the harmonics would then have the different
534 phases associated with chirps at each glottal pulse instead of all phases set to 0. Regardless, our peaky
535 speech generated robust canonical responses with good SNR while maintaining a natural-sounding, if very
536 slightly "buzzy," quality to the speech. Overall, continuous speech re-synthesized to contain impulse-like
537 characteristics is an effective way to elicit responses that distinguish contributions from different subcortical
538 structures.

539 *Latencies of component waves are consistent with distinct subcortical structures*

540 The latencies of the component waves of the responses to peaky speech are consistent with
541 activity arising from known subcortical structures. The inter-wave latencies between I-III, III-V and I-V fall
542 within the expected range for brainstem responses elicited by transients at 50–60 dB sensation level (SL)
543 and 50–100 Hz rates (Burkard and Hecox, 1983; Chiappa et al., 1979; Don et al., 1977), suggesting the
544 transmission times between auditory nerve, cochlear nucleus and rostral brainstem remain similar for
545 speech stimuli. However, these speech-evoked waves peak at later absolute latencies than responses to
546 transient stimuli at 60 dB SL and 90–100 Hz, but at latencies more similar to those presented at 50 dB SL
547 or 50 dB nHL in the presence of some masking noise (Backer et al., 2019; Burkard and Hecox, 1983;
548 Chiappa et al., 1979; Don et al., 1977; Maddox and Lee, 2018; Miller et al., 2017). There are reasons why
549 the speech-evoked latencies may be later. First, our level of 65 dB sound pressure level (SPL) may be
550 more similar to click levels of 50 dB SL. Second, although spectra of both speech and transients are broad,
551 clicks, chirps and even our previous speech stimuli (which was high-pass filtered at 1 kHz; Maddox and
552 Lee, 2018) have relatively greater high-frequency energy than the unaltered and peaky broadband speech
553 used in the present work. Neurons with higher characteristic frequencies respond earlier due to their basal
554 cochlear location, and contribute relatively more to brainstem responses (e.g., Abdala and Folsom, 1995),
555 leading to quicker latencies for stimuli that have greater high frequency energy. Also consistent with having
556 greater lower frequency energy, our unaltered and peaky speech responses were later than the response
557 from the same speech segments that were high-pass filtered at 1 kHz (Maddox and Lee, 2018). In fact, the
558 ABR to broadband peaky speech bore a close resemblance to the summation of each frequency-specific
559 response and the common component to peaky multiband speech (Figure 7), with peak wave latencies
560 representing the relative contribution of each frequency band. Third, higher stimulation rates prolong
561 latencies due to neural adaptation, and the 115–198 Hz average fundamental frequencies of our speech
562 were much higher than the 41 Hz embedded chirps and 50–100 Hz click rates (e.g., Burkard et al., 1990;
563 Burkard and Hecox, 1983; Chiappa et al., 1979; Don et al., 1977; Jiang et al., 2009). The effect of
564 stimulation rate was also demonstrated by the later ABR wave I, III, and V peak latencies for the female
565 narrator with the higher average fundamental frequency of 198 Hz (Figure 6A&C). Therefore, the differing
566 characteristics of typical periodic transients (such as clicks and chirps) and continuous speech may give
567 rise to differences in brainstem responses, even though they share canonical waveforms arising from
568 similar contributing subcortical structures.

569 The latency of the peaky speech-evoked response also differed from the non-standard, broad
570 responses to unaltered speech. However, latencies from these waveforms are difficult to compare due to
571 the differing morphology and the different analyses that were used to derive the responses. Evidence for
572 the effect of analysis comes from the fact that the same EEG collected in response to peaky speech could
573 be regressed with pulse trains to give canonical ABRs (Figures 1, 2), or regressed with the half-wave
574 rectified peaky speech to give the different, broad waveform (Figure 4). Furthermore, non-peaky
575 continuous speech stimuli with similar ranges of fundamental frequencies (between 100–300 Hz) evoke
576 non-standard, broad brainstem responses that also differ in morphology and latency depending on whether

577 the EEG is analyzed by deconvolution with the half-wave rectified speech (Figure 2, Maddox and Lee,
578 2018) or complex cross-correlation with the fundamental frequency waveform (Forte et al., 2017).
579 Therefore, again, even though the inferior colliculus and lateral lemniscus may contribute to generating
580 these different responses (Møller and Jannetta, 1983; Saiz-Alia and Reichenbach, 2020; Starr and
581 Hamilton, 1976), the morphology and latency may differ (sometimes substantially) depending on the
582 analysis technique used.

583 *Responses reflect activity from different frequency regions*

584 In addition to evoking canonical brainstem responses, peaky speech can be exploited for other
585 traditional uses of ABR, such as investigating subcortical responses across different frequencies.
586 Frequency-specific responses were measurable to two different types of multiband peaky speech: 4
587 frequency bands presented diotically (Figures 6, 8), and 5 frequency bands presented dichotically (Figure
588 9). Peak wave latencies of these responses decreased with increasing band frequency in a similar way to
589 responses evoked by tone pips and derived-bands from clicks in noise (Gorga et al., 1988; Neely et al.,
590 1988; Rasetshwane et al., 2013; Strelcyk et al., 2009), thereby representing activity evoked from different
591 areas across the cochlea. In fact, our estimates of the power law parameters of a (the central conduction
592 time), d (the frequency dependence) and b (the latency corresponding to 1 kHz and 65 dB SPL) for wave V
593 fell within the corresponding ranges that were previously reported for tone pips and derived-bands at 65 dB
594 ppeSPL (Neely et al., 1988; Rasetshwane et al., 2013; Strelcyk et al., 2009). Interestingly, the frequency-
595 specific responses across frequency band were similar in amplitude, or possibly slightly smaller for the
596 lower frequency bands (Figures 6, 8, 9), even though the relative energy of each band decreased with
597 increasing frequency, resulting in a ~30 dB difference between the lowest and highest frequency bands
598 (Figure 14). A greater response elicited by higher frequency bands is consistent with the relatively greater
599 contribution of neurons with higher characteristic frequencies to ABRs (Abdala and Folsom, 1995), as well
600 as the need for higher levels to elicit low frequency responses to tone pips that are close to threshold
601 (Gorga et al., 2006, 1993; Hyde, 2008; Stapells and Oates, 1997). Also, canonical waveforms were derived
602 in the higher frequency bands of diotically presented speech, with waves I and III identifiable in most
603 subjects. Multiband peaky speech will not replace the current frequency specific ABR, but there are
604 situations where it may be advantageous to use speech over tone pips. Measuring waves I, III, and V of
605 high frequency responses in the context of all the dynamics of speech may have applications to studying
606 effects of cochlear synaptopathy on speech comprehension (Bharadwaj et al., 2014; Liberman et al.,
607 2016). Another exciting potential application is the evaluation of supra-threshold hearing across frequency
608 in toddlers and individuals who do not provide reliable behavioral responses, as they may be more
609 responsive to sitting for longer periods of time while listening to a narrated story than to a series of tone
610 pips. Giving a squirmy toddler an iPad and presenting a story for 30 minutes could allow responses to
611 multiband peaky speech that confirm audibility at normal speech levels. Such a screening or metric of
612 audibility is a useful piece of knowledge in pediatric clinical management and is also being investigated for
613 the frequency following response (Easwar et al., 2020, 2015), which provides different information than the
614 canonical responses. An extension of this assessment would be to evaluate neural speech processing in
615 the context of hearing loss, as well as rehabilitation strategies such as hearing aids and cochlear implants.
616 Auditory prostheses have algorithms specifically tuned for the spectrotemporal dynamics of speech that
617 behave very differently in response to standard diagnostic stimuli such as trains of clicks or tone pips.
618 Peaky speech responses could allow us to assess how the auditory system is encoding the amplified
619 speech and validate audibility of the hearing aid fittings before the infant or toddler is old enough to provide
620 reliable speech perception testing. Therefore, the ability of peaky speech to yield both canonical waveforms
621 and frequency-specific responses makes this paradigm a flexible method that assesses speech encoding
622 in new ways.

623

624 Practical considerations for using peaky speech and deconvolution

625 *Filtering*

626 Having established that peaky speech is a flexible stimulus for investigating different aspects of
627 speech processing, there are several practical considerations for using the peaky speech paradigm. First,
628 filtering should be performed carefully. As recommended in Maddox and Lee (2018), causal filters – which
629 have impulse responses with non-zero values at positive lags only – should be used to ensure cortical
630 activity at later peak latencies does not spuriously influence earlier peaks corresponding to subcortical
631 origins. Applying less aggressive, low-order filters (i.e., broadband with shallow roll-offs) will help reduce
632 the effects of causal filtering on delaying response latency. The choice of high-pass cutoff will also affect
633 the response amplitude and morphology. After evaluating several orders and cutoffs to the high-pass
634 filters, we determined that early waves of the peaky broadband ABRs were best visualized with a 150 Hz
635 cutoff, whereas a lower cutoff frequency of 30 Hz was necessary to view the ABR and MLR of the
636 broadband responses. When evaluating specific contributions to the earliest waves, we recommend at
637 least a first order 150 Hz high-pass filter or a more aggressive second order 200 Hz high-pass filter to deal
638 with artifacts arising from nonlinearities that are not taken into account by the pulse train regressor or any
639 potential influences by responses from later sources (Figure 3). For multiband responses, the 150 Hz high-
640 pass filter significantly reduced the response but also decreased the low-frequency noise in the pre-
641 stimulus interval. For the 4-band multiband peaky speech the 150 Hz and 30 Hz filters provided similar
642 acquisition times for 0 dB SNR, but better SNRs were obtained quicker with 150 Hz filtering for the 10-band
643 multiband peaky speech.

644 *Choice of narrator (stimulation rate)*

645 Second, the choice of narrator impacts the responses to both broadband and multiband peaky
646 speech. Although overall morphology was similar, the male-narrated responses were larger, contained
647 more clearly identifiable component waves in a greater proportion of subjects, and achieved a 0 dB SNR at
648 least 2.1 to over 10 times faster than those evoked by a female narrator. These differences likely stemmed
649 from the ~77 Hz difference in average pitch, as higher stimulation rates evoke smaller responses due to
650 adaptation and refractoriness (e.g., Burkard et al., 1990; Burkard and Hecox, 1983; Chiappa et al., 1979;
651 Don et al., 1977; Jiang et al., 2009). Indeed, a 50 Hz change in fundamental frequency yields a 24%
652 reduction in the modelled auditory brainstem response that was derived as the complex cross-correlation
653 with the fundamental frequency (Saiz-Alia and Reichenbach, 2020). The narrator differences exhibited in
654 the present study may be larger than those in other studies with continuous speech (Forte et al., 2017;
655 Maddox and Lee, 2018; Saiz-Alía et al., 2019) as a result of the different regressors. These response
656 differences do not preclude using narrators with higher fundamental frequencies in future studies, but the
657 time required for usable responses from each narrator must be considered when planning experiments,
658 and caution taken when interpreting comparisons between conditions with differing narrators. The
659 strongest results will come from comparing responses to the same narrator (or even the same speech
660 recordings) under different experimental conditions.

661 *SNR and recording time for multiple responses*

662 Third, the necessary recording time depends on the chosen SNR threshold, experimental demands,
663 and stimulus. We chose a threshold SNR of 0 dB based on when waveforms became clearly identifiable,
664 but of course a different threshold would change our recording time estimates (though, notably, not the
665 ratios between them). With this SNR threshold, acquisition times were quick enough for broadband peaky
666 responses to allow multiple conditions in a reasonable recording session. With male-narrated broadband
667 peaky speech, all subjects achieved 0 dB SNR ABRs in < 5 minutes and MLRs in < 20 minutes, thereby
668 affording between 3 and 12 conditions in an hour recording session. These recording times are
669 comparable, if not faster, than the 8 minutes for the broad response to unaltered speech, 6–12 minutes for
670 the chirp-embedded speech (Backer et al., 2019), ~10 minutes for the broad complex-cross correlation

24

671 response to the fundamental waveform (Forte et al., 2017), and 33 minutes for the broad response to high-
672 passed continuous speech (Maddox and Lee, 2018). However, using a narrator with a higher fundamental
673 frequency could increase testing time by 2- to over 10-fold. In this experiment, at most 2 conditions per
674 hour could be tested with the female-narrated broadband peaky speech. Unlike broadband peaky speech,
675 the testing times required for all frequency-specific responses to reach 0 dB SNR were significantly longer,
676 making only 1 condition feasible within a recording session. At least 30 minutes was necessary for the
677 multiband peaky speech, but based on extrapolated testing times, about 1 hour is required for 90% of
678 subjects and 2 hours for 75% of subjects to achieve this threshold for all 4 bands of diotic speech and all
679 10 bands of dichotic speech respectively. The longer testing times are important to consider when planning
680 studies using multiband peaky speech with several frequency bands.

681 *Number of frequency bands*

682 Fourth, as mentioned above, the number of frequency bands incorporated into multiband peaky
683 speech decreases SNR and increases testing time. Although it is possible to simultaneously record up to
684 10 frequency-specific responses, the significant time required to obtain decent SNRs reduces the feasibility
685 of testing multiple conditions or having recording sessions lasting less than 1–2 hours. However, pursuing
686 shorter testing times with multiband peaky speech is possible. Depending on the experimental question,
687 different multiband options could be considered. For male-narrated speech, the 2–4 and 4–8 kHz
688 responses had good SNRs and exhibited waves I, III, and V within 9 minutes for 90% of subjects.
689 Therefore, if researchers were more interested in comparing responses in these higher frequency bands,
690 they could stop recording once these bands reach threshold but before the lower frequency bands reach
691 criterion (i.e., within 9 minutes). Alternatively, the lower frequencies could be combined into a single
692 broader band in order to reduce the total number of bands, or the intensity could be increased to evoke
693 responses with larger amplitudes. Therefore, different band and parameter considerations could reduce
694 testing time and improve the feasibility, and thus utility, of multiband peaky speech.

695 *Flexible analysis windows for deriving responses from auditory nerve to cortex*

696 Fifth, and finally, a major advantage of deconvolution analysis is that the analysis window for the
697 response can be extended arbitrarily in either direction to include a broader range of latencies (Maddox
698 and Lee, 2018). Extending the pre-stimulus window leftward provides a better estimate of the SNR, and
699 extending the window rightward allows parts of the response that come after the ABR and MLR to be
700 analyzed as well, which are driven by the cortex. These later responses can be evaluated in response to
701 broadband peaky speech, but as shown in Figures 8 and 9, only ABR and early MLR waves are present in
702 the frequency-specific responses. The same broadband peaky speech data from Figure 5 are high-pass
703 filtered at 1 Hz and displayed with an extended time window in Figure 12, which shows component waves
704 of the ABR, MLR and late latency responses (LLR). Thus, this method allows us to simultaneously
705 investigate speech processing ranging from the earliest level of the auditory nerve all the way through the
706 cortex without requiring extra recording time. Usually the LLR is larger than the ABR/MLR, but our subjects
707 were encouraged to relax and rest, yielding a passive LLR response. Awake and attentive subjects may
708 improve the LLR; however, other studies that present continuous speech to attentive subjects also report
709 smaller and different LLR (Backer et al., 2019; Maddox and Lee, 2018), possibly from cortical adaptation to
710 a continuous stimulus. Here we used a simple 2-channel montage that is optimized for recording ABRs, but
711 a full multi-channel montage could also be used to more fully explore the interactions between subcortical
712 and cortical processing of naturalistic speech. The potential for new knowledge about how the brain
713 processes naturalistic and engaging stimuli cannot be undersold.
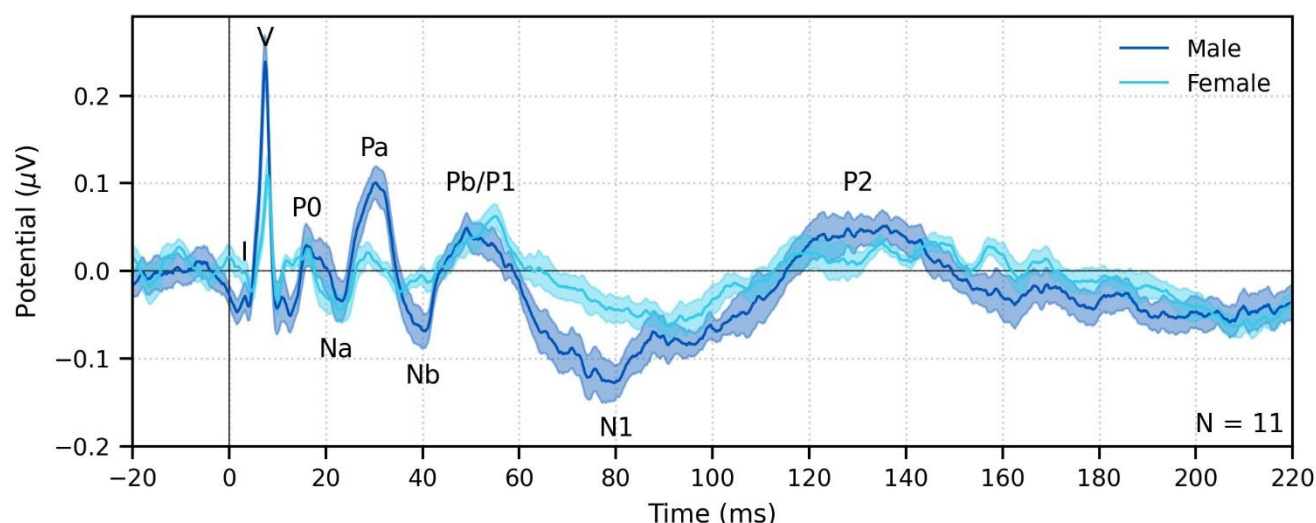
Figure 12. *The range of lags can be extended to allow early, middle and late latency responses to be analyzed from the same recording to broadband peaky speech. Average waveforms across subjects (areas show ± 1 SEM) are shown for responses measured to 32 minutes of broadband peaky speech narrated by a male (dark blue) and female (light blue). Responses were high-pass filtered at 1 Hz using a first order Butterworth filter, but different filter parameters can be used to focus on each stage of processing. Canonical waves of the ABR, MLR and LLR are labeled for the male-narrated speech. Due to adaptation, amplitudes of the late potentials are smaller than typically seen with other stimuli that are shorter in duration with longer inter-stimulus intervals than our continuous speech. Waves I and III become more clearly visible by applying a 150 Hz high-pass cutoff.*

714

715

Peaky speech is a tool that opens up new lines of query

716

717   The peaky speech paradigm is a viable method for recording broadband and frequency-specific
718 responses from distinct subcortical structures using an engaging, continuous speech stimulus. The
719 customizability and flexibility of peaky speech facilitates new lines of query, both in neuroscientific and
720 clinical domains. Speech often occurs within a mixture of sounds, such as other speech sources,
721 background noise, or music. Furthermore, visual cues from a talker's face are often available to aid speech
722 understanding, particularly in environments with low SNR (e.g., Bernstein and Grant, 2009; Grant et al.,
723 2007). Peaky speech facilitates investigation into the complex subcortical encoding and processing that
724 underpins successful listening in these scenarios using naturalistic, engaging tasks. Indeed, previous
725 methods have been quite successful in elucidating cortical processing of speech under these conditions
726 (O'Sullivan et al., 2019; Teoh and Lalor, 2019). Whereas these cortical studies could use regressors that
727 are not acoustically-based – such as semantics or surprisal – the fast responses of subcortical structures
728 necessitate a regressor that can allow timing of components to separate subcortical from cortical origins.
729 The similarity of the peaky speech response to the click response is an advantage because it is the only
730 way to understand what is occurring in separate subcortical regions during the dynamic spectrotemporal
731 context of speech. As with other work in this area, how informative the response is about speech
732 processing will depend on how it is deployed experimentally. Experiments that measure the response to
733 the same stimuli in different cognitive states (unattended/attended, understood/not understood) will
734 illuminate the relationship between those states and subcortical encoding. Such an approach was recently
735 used to show how the brainstem complex cross-correlation to a fundamental waveform can change based
736 on top-down attention (Forte et al., 2017). Finally, the ability to customize peaky speech for measuring
737 frequency-specific responses provides potential applications to clinical research in the context of facilitating
738 assessment of supra-threshold hearing function and changes to how speech may be encoded following
739 intervention strategies and technologies while using a speech stimulus that algorithms in hearing aids and
740 cochlear implants are designed to process.

741

METHODS

Participants

Data were collected over 3 experiments that were conducted under a protocol approved by the University of Rochester Research Subjects Review Board (#1227). All subjects gave informed consent before the experiment began and were compensated for their time. In each of experiments 1 and 2, there were equipment problems during testing for one subject, rendering data unusable in the analyses. Therefore, there were a total of 22, 11, and 11 subjects included in experiments 1, 2, and 3 respectively. Four subjects completed both experiments 1 and 2, and 2 subjects completed both experiments 2 and 3. The 38 unique subjects (25 females, 66%) were aged 18–32 years with a mean ± SD age of 23.0 ± 3.6 years. Audiometric screening confirmed subjects had normal hearing in both ears, defined as thresholds ≤ 20 dB HL from 250 to 8000 Hz. All subjects identified English as their primary language.

Stimulus presentation and EEG measurement

In each experiment, subjects listened to 128 minutes of continuous speech stimuli while reclined in a darkened sound booth. They were not required to attend to the speech and were encouraged to relax and to sleep. Speech was presented at an average level of 65 dB SPL over ER-2 insert earphones (Etymotic Research, Elk Grove, IL) plugged into an RME Babyface Pro digital soundcard (RME, Haimhausen, Germany) via an HB7 headphone amplifier (Tucker Davis Technologies, Alachua, FL). Stimulus presentation was controlled by a custom python (Python Programming Language, RRID:SCR_008394) script using publicly available software (Expyfun, RRID:SCR_019285; available at https://github.com/LABSN/expyfun; Larson et al., 2014). We interleaved conditions in order to prevent slow impedance drifts or transient periods of higher EEG noise from unevenly affecting one condition over the others. Physical measures to reduce stimulus artifact included: 1) hanging earphones from the ceiling so that they were as far away from the EEG cap as possible; and 2) sending an inverted signal to a dummy earphone (blocked tube) attached in the same physical orientation to the stimulus presentation earphones in order to cancel electromagnetic fields away from transducers. The soundcard also produced a digital signal at the start of each epoch, which was converted to trigger pulses through a custom trigger box (modified from a design by the National Acoustic Laboratories, Sydney, NSW, Australia) and sent to the EEG system so that audio and EEG data could be synchronized with sub-millisecond precision.

EEG was recorded using BrainVision's PyCorder software (RRID:SCR_019286). Ag/AgCl electrodes were placed at the high forehead (FCz, active non-inverting), left and right earlobes (A1, A2, inverting references), and the frontal pole (Fpz, ground). These were plugged into an EP-Preamp system specifically for recording ABRs, connected to an ActiCHamp recording system, both manufactured by BrainVision. Data were sampled at 10,000 Hz and high-pass filtered at 0.1 Hz. Offline, raw data were high-pass filtered at 1 Hz using a first-order causal Butterworth filter to remove slow drift in the signal, and then notch filtered with 5 Hz wide second-order infinite impulse response (IIR) notch filters to remove 60 Hz and its first 3 odd harmonics (180, 300, 420 Hz). To optimize parameters for viewing the ABR and MLR components of peaky speech responses, we evaluated several orders and high-pass cutoffs to the filters. Early waves of the broadband peaky ABRs were best visualized with a 150 Hz cutoff, whereas a lower cutoff frequency of 30 Hz was necessary to view the ABR and MLR of the broadband responses. Conservative filtering with a first order filter was sufficient with these cutoff frequencies.

Speech stimuli and conditions

Speech stimuli were taken from two audiobooks. The first was *The Alchemyst* (Scott, 2007), read by a male narrator and used in all 3 experiments. The second was *A Wrinkle in Time* (L'Engle, 2012), read

787 by a female narrator and used in experiments 2 and 3. These stimuli were used in Maddox and Lee (2018),
788 but in that study a gentle high-pass filter was applied which was not done for this study. Briefly, the
789 audiobooks were resampled to 44,100 Hz and then silent pauses were truncated to 0.5 s. Speech was
790 segmented into 64 s epochs with 1 s raised cosine fade-in and fade-out. Because conditions were
791 interleaved, the last 4 s of a segment were repeated in the next segment so that subjects could pick up
792 where they left off if they were listening.

793   In experiment 1 subjects listened to 3 conditions of male speech (42.7 min each): unaltered speech,
794 re-synthesized broadband peaky speech, and re-synthesized multiband peaky speech (see below for a
795 description of re-synthesized speech). In experiment 2 subjects listened to 4 conditions of re-synthesized
796 peaky speech (32 minutes each): male and female narrators of both broadband and multiband peaky
797 speech. For these first 2 experiments, speech was presented diotically (same speech to both ears). In
798 experiment 3 subjects listened to both male and female dichotic (slightly different stereo speech of the
799 same narrator in each ear) multiband peaky speech designed for audiological applications (64 min of each
800 narrator). The same 64 s of speech was presented simultaneously to each ear, but the stimuli were dichotic
801 due to how the re-synthesized multiband speech was created (see below).

802

803 Stimulus design

804   The brainstem responds best to impulse-like stimuli, so we re-synthesized the speech segments
805 from the audiobooks (termed "unaltered") to create 3 types of "peaky" speech, with the objectives of 1)
806 evoking additional waves of the ABR reflecting other neural generators, and 2) measuring responses to
807 different frequency regions of the speech. The process is described in detail below, but is best read in
808 tandem with the code that is publicly available (https://github.com/maddoxlab). Figure 13 compares the
809 unaltered speech and re-synthesized broadband and multiband peaky speech. Comparing the pressure
810 waveforms shows that the peaky speech is as click-like as possible, but comparing the spectrograms (how
811 sound varies in amplitude at every frequency and time point) shows that the overall spectrotemporal
812 content that defines speech is basically unchanged by the re-synthesis. Audio files 1–6 provide examples
813 of each stimulus type for both narrators, which demonstrate the barely perceptible difference between
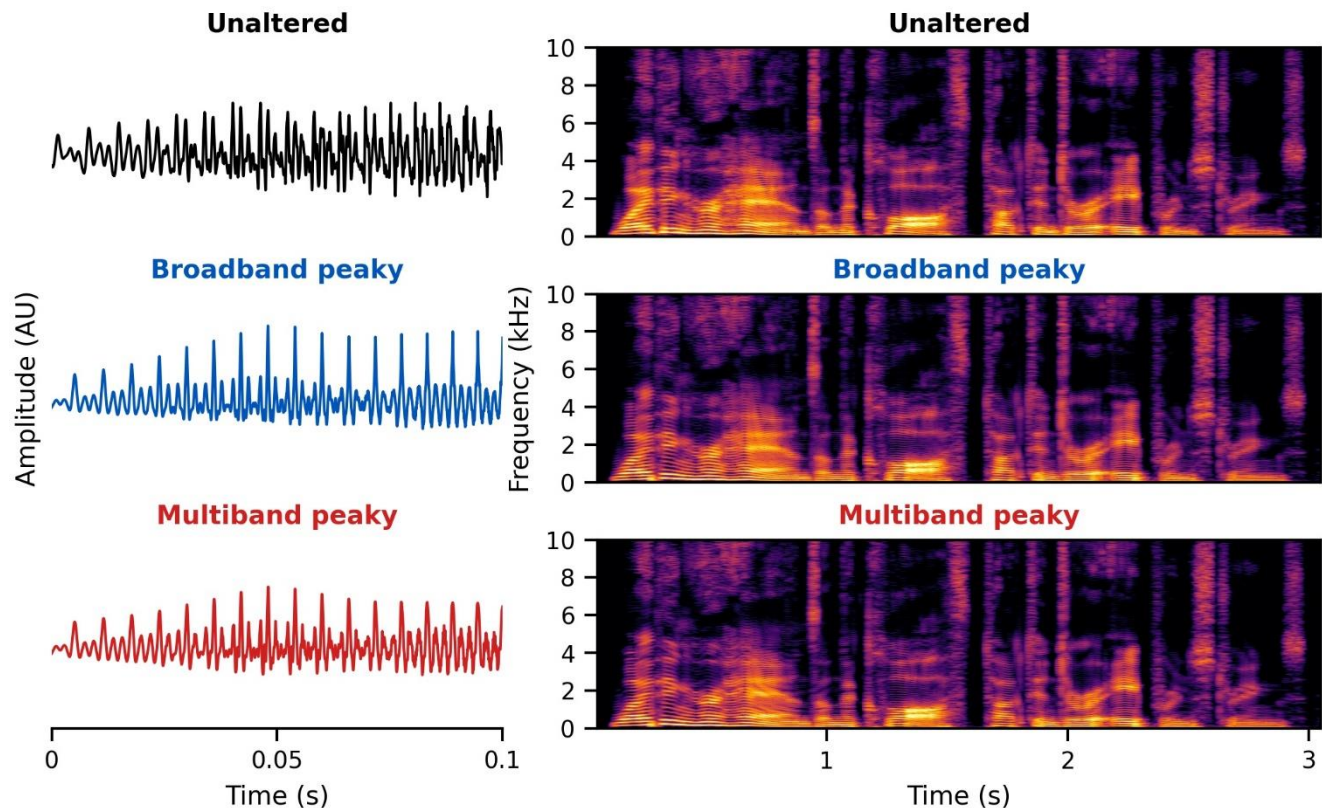814 unaltered and peaky speech.

815

Figure 13. *Unaltered speech waveform (top left) and spectrogram (top right) compared to re-synthesized broadband peaky speech (middle left and right) and multiband peaky speech (bottom left and right). Comparing waveforms shows that the peaky speech is as "click-like" as possible, while comparing the spectrograms shows that the overall spectrotemporal content that defines speech is basically unchanged by the re-synthesis. A naïve listener is unlikely to notice that any modification has been performed, and subjective listening confirms the similarity. Yellow/lighter colors represent larger amplitudes than purple/darker colors in the spectrogram. Audio files 1–6 provide examples of each stimulus type for both narrators.*

816 *Broadband peaky speech*

817   Voiced speech comprises rapid openings and closings of the vocal folds which are then filtered by
818 the mouth and vocal tract to create different vowel and consonant sounds. The first processing step in
819 creating peaky speech was to use speech processing software (PRAAT, RRID:SCR_016564; Boersma
820 and Weenink, 2018) to extract the times of these glottal pulses. Sections of speech where glottal pulses
821 were within 17 ms of each other were considered voiced (vowels and voiced consonants like /z/). 17 ms is
822 the longest inter-pulse interval one would expect in natural speech because it is the inverse of 60 Hz, the
823 lowest pitch at which someone with a deep voice would likely speak. A longer gap in pulse times was
824 considered a break between voiced sections. These segments were identified in a "mixer" function of time,
825 with indices of 1 indicating unvoiced and 0 indicating voiced segments (and would later be responsible for
826 time-dependent blending of re-synthesized and natural speech, hence its name). Transitions of the binary
827 mixer function were smoothed using a raised cosine envelope spanning the time between the first and
828 second pulses, as well as the last two pulses of each voiced segment. During voiced segments, the glottal
829 pulses set the fundamental frequency of speech (i.e., pitch), which were allowed to vary from a minimum to
830 maximum of 60–350 Hz for the male narrator and 90–500 Hz for the female narrator. For the male and
831 female narrators, these pulses gave a mean ± SD fundamental frequency (i.e., pulse rate) in voiced
832 segments of 115.1 ± 6.7 Hz and 198.1 ± 20 Hz respectively, and a mean ± SD pulses per second over the
833 entire 64 s, inclusive of unvoiced periods and silences, of 69.1 ± 5.7 Hz and 110.8 ± 11.4 respectively.
834 These pulse times were smoothed using 10 iterations of replacing pulse time p_i with the mean of pulse

835   times p_{i-1} to p_{i+1} if the $\log_2$ absolute difference in the time between p_i and p_{i-1} and p_{i+1} was less than
836   $\log_2(1.6)$.

837   The fundamental frequency of voiced speech is dynamic, but the signal always consists of a set of
838   integer-related frequencies (harmonics) with different amplitudes and phases. To create the waveform
839   component at the fundamental frequency, $f_0(t)$, we first created a phase function, $\varphi(t)$, which increased
840   smoothly by 2π between glottal pulses within the voiced sections as a result of cubic interpolation. We then
841   computed the spectrogram of the unaltered speech waveform – which is a way of analyzing sound that
842   shows its amplitude at every time and frequency (Figure 13, top-right) – which we called $A[t, f_0(t)]$. We
843   then created the fundamental component of the peaky speech waveform as:

844   $$h_0(t) = A[t, f_0(t)] \cos[\varphi(t)].$$

845   This waveform has an amplitude that changes according to the spectrogram but always peaks at the time
846   of the glottal pulses.

847   Next the harmonics of the speech were synthesized. The kth harmonic of speech is at a frequency
848   of $(k+1)f_0$ so we synthesized each harmonic waveform as:

849   $$h_k(t) = A[t, (k+1)f_0(t)] \cos[(k+1)\varphi(t)].$$

850   Each of these harmonic waveforms has multiple peaks per period of the fundamental, but every harmonic
851   also has a peak at exactly the time of the glottal pulse. Because of these coincident peaks, when the
852   harmonics are summed to create the re-synthesized voiced speech, there is always a large peak at the
853   time of the glottal pulse. In other words, the phases of all the harmonics align at each glottal pulse, making
854   the pressure waveform of the speech appear "peaky" (left-middle panel of Figure 13).

855   The resultant re-synthesized speech contained only the voiced segments of speech and was
856   missing unvoiced sounds like /s/ and /k/. Thus the last step was to mix the re-synthesized voiced segments
857   with the original unvoiced parts. This was done by cross-fading back and forth between the unaltered
858   speech and re-synthesized speech during the unvoiced and voiced segments respectively, using the binary
859   mixer function created when determining where the voiced segments occurred. We also filtered the peaky
860   speech to an upper limit of 8 kHz, and used the unaltered speech above 8 kHz, to improve the quality of
861   voiced consonants such as /z/. Filter properties for the broadband peaky speech are further described
862   below in the "Band filters" subsection.

863

864   *Multiband peaky speech*

865   The same principles to generate broadband peaky speech were applied to create stimuli designed
866   to investigate the brainstem's response to different frequency bands that comprise speech. This makes use
867   of the fact that over time, speech signals with slightly different f_0 are independent, or have (nearly) zero
868   cross-correlation, at the lags for the ABR. To make each frequency band of interest independent, we
869   shifted the fundamental frequency and created a fundamental waveform and its harmonics as:

870   $$h_k(t) = A[t, (k+1)f_0(t)] \cos[(k+1)\varphi(t)],$$

871   where

872   $$\varphi(t) = 2\pi \int_0^t (f_0(\tau) + f_\Delta)d\tau,$$

873   and where f_Δ is the small shift in fundamental frequency.

874   In these studies, we increased fundamentals for each frequency band by the square root of each
875   successive prime number and subtracting one, resulting in a few tenths of a hertz difference between
876   bands. The first, lowest frequency band contained the un-shifted f_0. Responses to this lowest, un-shifted

877 frequency band showed some differences from the common component for latencies > 30 ms that were not
878 present in the other, higher frequency bands (Figure 6, 0–1 kHz band), suggesting some low-frequency
879 privilege/bias in this response. Therefore, we suggest that following studies create independent frequency
880 bands by synthesizing a new fundamental for each band. The static shifts described above could be used,
881 but we suggest an alternative method that introduces random dynamic frequency shifts of up to ±1 Hz over
882 the duration of the stimulus. From this random frequency shift we can compute a dynamic random phase
883 shift, to which we also add a random starting phase, θ_Δ, which is drawn from a uniform distribution between
884 0 and 2π. The phase function from the above set of formulae would be replaced with this random dynamic
885 phase function:

886
$$\varphi(t) = 2\pi \int_0^t [f_0(\tau) + f_\Delta(\tau)]d\tau + \theta_\Delta$$

887 Validation data from one subject is provided in Figure 6-figure supplement 1. Responses from all four
888 bands show more consistent resemblance to the common component, indicating that this method is
889 effective at reducing stimulus-related bias. However, low-frequency dependent differences remained,
890 suggesting there is also unique neural-based low-frequency activity to the speech-evoked responses.

891   This re-synthesized speech was then band-pass filtered to the frequency band of interest (e.g. from
892 0–1 kHz or 2–4 kHz). This process was repeated for each independent frequency band, then the bands
893 were mixed together and then these re-synthesized voiced parts were mixed with the original unaltered
894 voiceless speech. This peaky speech comprised octave bands with center frequencies of: 707, 1414, 2929,
895 5656 Hz for experiments 1 and 2, and of 500, 1000, 2000, 4000, 8000 Hz for experiment 3. Note that for
896 the lowest band, the actual center frequency was slightly lower because the filters were set to pass all
897 frequencies below the upper cutoff. Filter properties for these two types of multiband speech are shown in
898 the middle and right panels of Figure 16 and further described below in the "Band filters" subsection. For
899 the dichotic multiband peaky speech, we created 10 fundamental waveforms – 2 in each of the five filter
900 bands for the two different ears, making the output audio file stereo (or dichotic). We also filtered this
901 dichotic multiband peaky speech to an upper limit of 11.36 kHz to allow for the highest band to have a
902 center frequency of 8 kHz and octave width. The relative mean-squared magnitude in decibels for
903 components of the multiband peaky speech (4 filter bands) and dichotic (audiological) multiband peaky
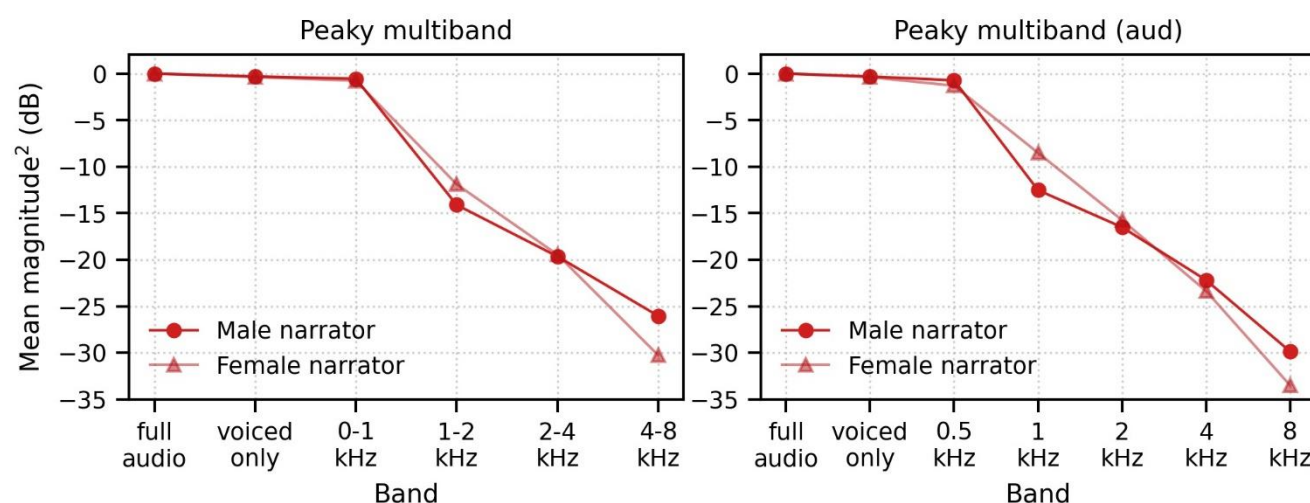904 speech (5 filter bands) are shown in Figure 14.

Figure 14. *Relative mean-squared magnitude in decibels of multiband peaky speech with 4 filter bands (left) and 5 filter bands (right) for male-(blue) and female-(orange) narrated speech. The full audio comprises unvoiced and re-synthesized voiced sections, which was presented to the subjects during the experiments. The other bands reflect the relative magnitude of the voiced sections (voiced only), and each filtered frequency band.*

905

906   For peaky speech, the re-synthesized speech waveform was presented during the experiment but
907   the pulse trains were used as the input stimulus for calculating the response (i.e., the regressor, see
908   Response derivation section below). These pulse trains all began and ended together in conjunction with
909   the onset and offset of voiced sections of the speech. To verify which frequency ranges of the multiband
910   pulse trains were independent across frequency bands, and would thus yield truly band-specific responses,
911   we conducted a spectral coherence analyses on the pulse trains. All 60 unique 64 s sections of each male-
912   and female-narrated multiband peaky speech used in the three experiments were sliced into 1 s segments
913   for a total of 3,840 slices. Phase coherence across frequency was then computed across these slices for
914   each combination of pulse trains according to the formula:

915
$$C_{xy} = \frac{\left| E[\mathcal{F}\{x_i\}^* \, \mathcal{F}\{y_i\}] \right|}{\sqrt{E[\mathcal{F}\{x_i\}^* \, \mathcal{F}\{x_i\}] \, E[\mathcal{F}\{y_i\}^* \, \mathcal{F}\{y_i\}]}}$$

916   where C_{xy} denotes coherence between bands x and y, $E[\ ]$ the average across slices, $\mathcal{F}$ the fast Fourier
917   transform, * complex conjugation, x_i the pulse train for slice i in band x, and y_i the pulse train for slice i in
918   band y.

919   Spectral coherence for each narrator is shown in Figure 15. For the 4-band multiband peaky
920   speech used in experiments 1 and 2 there were 6 pulse train comparisons. For the audiological multiband
921   peaky speech used in experiment 3, there were 5 bands for each of 2 ears, resulting in 10 pulse trains and
922   45 comparisons. All 45 comparisons are shown in Figure 15A for the stimuli created with static frequency
923   shifts, and the 6 comparisons for the stimuli created with random dynamic frequency shifts (used in the
924   pilot experiment) are shown in Figure 15B. Pulse trains were coherent (> 0.1) up to a maximum of 71 and
925   126 Hz for male- and female-narrated speech respectively, which roughly correspond to the mean ± SD
926   pulse rates (calculated as total pulses / 64 s) of 69.1 ± 5.7 Hz and 110.8 ± 11.4 respectively. This means
927   that above ~130 Hz the stimuli were no longer coherent and evoked frequency-specific responses.
928   Importantly, responses would be to correlated stimuli, i.e., not frequency-specific, at frequencies below this
929   cutoff and would result in a low-frequency response component that is present in (or common to) all band
930   responses. There were two separated collections of curves for the stimuli with static frequency shifts, which
931   stemmed from the shifted frequency bands having different coherence with each other than the lowest
932   unshifted band. This stimulus-related bias in the lowest frequency band was reduced by using dynamic
933   random frequency shifts (used in the pilot experiment), as indicated by the similar spectral coherence
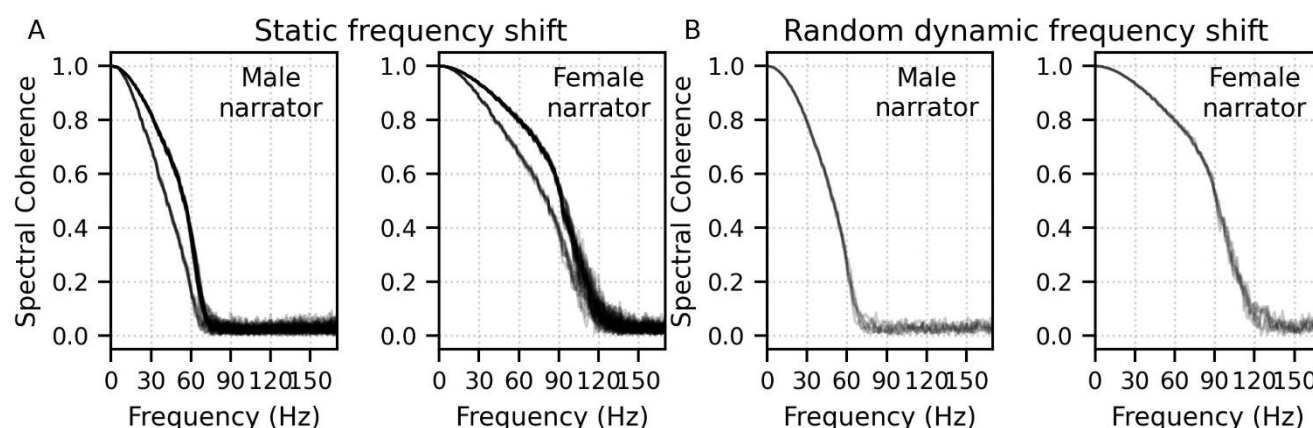934   curves shown in Figure 15B.

935

Figure 15. *Spectral coherence of pulse trains for multiband peaky speech narrated by a male and female. Spectral coherence was computed across 1 s slices from 60 unique 64 s multiband peaky speech segments (3,840 total slices) for each combination of bands. Each light gray line represents the coherence for one band comparison. (A) There were 45 comparisons across the 10-band (audiological) speech used in experiment 3 (5 frequency bands x 2 ears). The lowest band was unshifted and the other 9 bands had static frequency shifts. (B) There were 6 comparisons across 4 pulse trains of the bands in the pilot experiment, which all had dynamic random frequency shifts. Pulse trains (i.e., the input stimuli, or regressors, for the deconvolution) were frequency-dependent (coherent) below 72 Hz for the male multiband speech and 126 Hz for the female multiband speech.*

936   To identify the effect of the low-frequency stimulus coherence in the responses, we computed the
937 common component across pulse trains by creating an averaged response to 6 additional "fake" pulses
938 trains that were created during stimulus design but were not used during creation of the multiband peaky
939 speech wav files. The common component was assessed for both "fake" pulse trains taken from shifts
940 lower than the original fundamental frequency and those taken from shifts higher than the highest "true" re-
941 synthesized fundamental frequency. For the dynamic frequency shift method (pilot experiment) an
942 additional 6 pulse trains were created. Figure 15-figure supplement 1 shows that the common component
943 was similar for "fake" pulse trains created using static or dynamic random frequency shifts. To assess
944 frequency-specific responses to multiband speech, we subtracted this common component from the band
945 responses. Alternatively, one could simply high-pass the stimuli at 150 Hz using a first-order causal
946 Butterworth filter (being mindful of edge artifacts). However, this high-pass filtering reduces response
947 amplitude and may affect response detection (see Results for more details).

948   We also verified the independence of the stimulus bands by treating the regressor pulse train as the
949 input to a system whose output was the rectified stimulus audio and performed deconvolution (see
950 Deconvolution and Response derivation section below). Further details are provided in Figure 15-figure
951 supplement 2. The responses showed that the non-zero responses only occurred when the correct pulse
952 train was paired with the correct audio.

953

954   *Band filters*

955   Because the fundamental frequencies for each frequency band were designed to be independent
956 over time, the band filters for the speech were designed to cross over in frequency at half power. To make
957 the filter, the amplitude was set by taking the square root of the specified power at each frequency. Octave
958 band filters were constructed in the frequency domain by applying trapezoids – with center bandwidth and
959 roll-off widths of 0.5 octaves. For the first (lowest frequency) band, all frequencies below the high-pass
960 cutoff were set to 1, and likewise for all frequencies above the low-pass cutoff for the last (highest
961 frequency) band were set to 1 (Figure 16 top row). The impulse response of the filters was assessed by
962 shifting the inverse FFT (IFFT) of the bands so that time zero was in the center, and then applied a Nuttall
963 window, thereby truncating the impulse response to length of 5 ms (Fig 16 middle row). The actual

964 frequency response of the filter bands was assessed by taking the FFT of the impulse response and
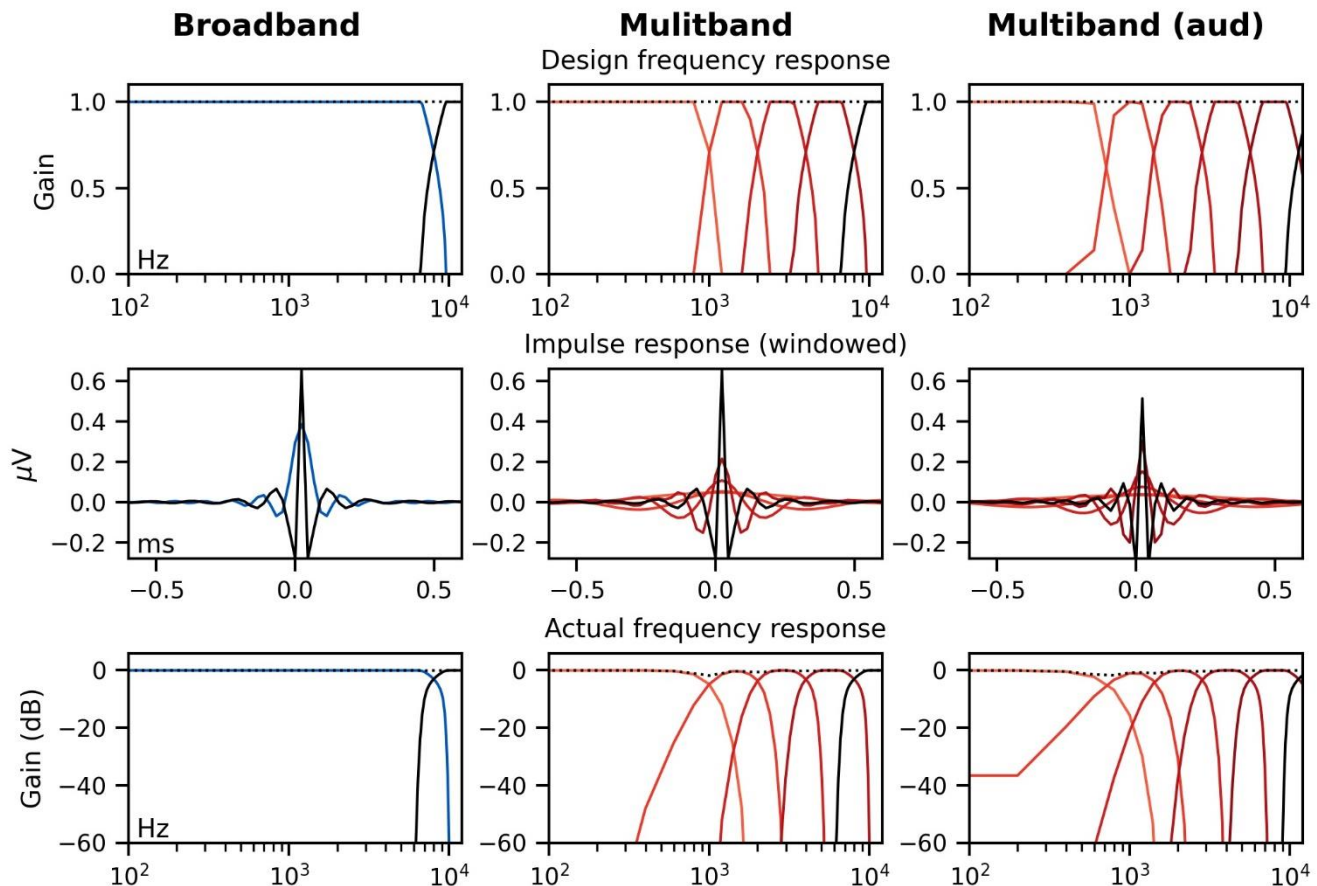965 plotting the magnitude (Figure 16 bottom row).

966

Figure 16. Octave band filters used to create re-synthesized broadband peaky speech (left, blue), diotic multiband peaky speech with 4 bands (middle, red), and dichotic multiband peaky speech using 5 bands with audiological center frequencies (right, red). The last band (2nd, 5th, 6th respectively, black line) was used to filter the high-frequencies of unaltered speech during mixing to improve the quality of voiced consonants. The designed frequency response using trapezoids (top) were converted into the time-domain using IFFT, shifted and Nuttall windowed to create impulse responses (middle), which were then used to assess the actual frequency response by converting into the frequency domain using FFT (bottom).

967

968   As mentioned above, broadband peaky speech was filtered to an upper limit of 8 kHz for diotic
969 peaky speech and 11.36 kHz for dichotic peaky speech. This band filter was constructed from the second
970 last octave band filter from the multiband filters (i.e., the 4–8 kHz band from the top-middle of Figure 16,
971 dark red line) by setting the amplitude of all frequencies less than the high-pass cutoff frequency to 1
972 (Figure 16 top-left panel, blue line). As mentioned above, unaltered (unvoiced) speech above 8 kHz (diotic)
973 or 11.36 kHz (dichotic) was mixed with the broadband and multiband peaky speech, which was
974 accomplished by applying the last (highest) octave band filter (8+ or 11.36+ kHz band, black line) to the
975 unaltered speech and mixing this band with the re-synthesized speech using the other bands.

976

977 *Alternating polarity*

978   To limit stimulus artifact, we also alternated polarity between segments of speech. To identify
979 regions to flip polarity, the envelope of speech was extracted using a first-order causal Butterworth low-

34

980 pass filter with a cutoff frequency of 6 Hz applied to the absolute value of the waveform. Then flip indices
981 were identified where the envelope became less than 1 percent of the median envelope value, and then a
982 function that changed back and forth between 1 and −1 at each flip index was created. This function of
983 spikes was smoothed using another first-order causal Butterworth low-pass filter with a cutoff frequency of
984 10,000 Hz, which was then multiplied with the re-synthesized speech before saving to a wav file.

985

986 Response derivation

987 *Deconvolution*

988   The peaky-speech ABR was derived by using deconvolution, as in previous work (Maddox and Lee,
989 2018), though the computation was performed in the frequency domain for efficiency. The speech was
990 considered the input to a linear system whose output was the recorded EEG signal, with the ABR
991 computed as the system's impulse response. As in Maddox and Lee (2018), for the unaltered speech, we
992 used the half-wave rectified audio as the input waveform. Half-wave rectification was accomplished by
993 separately calculating the response to all positive and all negative values of the input waveform for each
994 epoch and then combining the responses together during averaging. For our new re-synthesized peaky
995 speech, the input waveform was the sequence of impulses that occurred at the glottal pulse times and
996 corresponded to the peaks in the waveform. Figure 17 shows a section of stimulus and the corresponding
997 input signal of glottal pulses used in the deconvolution.
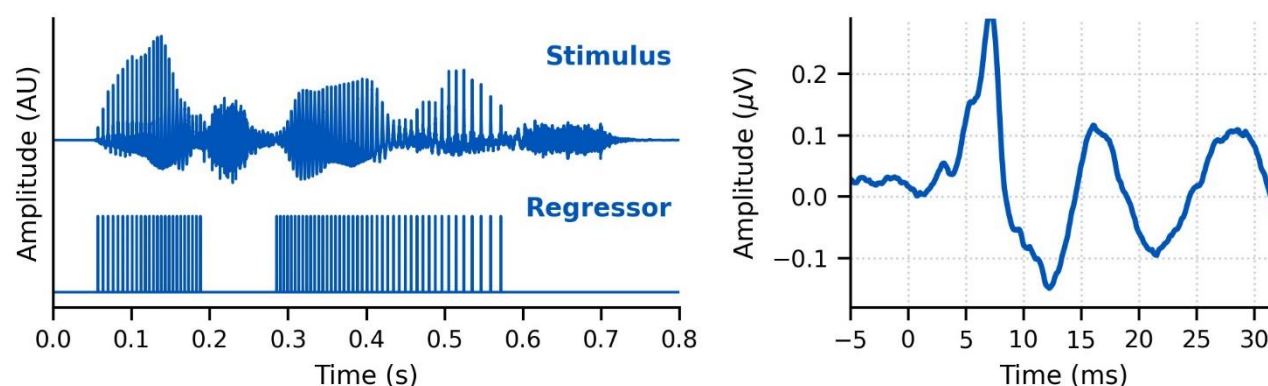
998

Figure 17. Left: A segment of broadband peaky speech stimulus (top) and the corresponding glottal pulse train (bottom) used in calculating the broadband peaky speech response. Right: An example broadband peaky speech response from a single subject. The response shows ABR waves I, III, and V at ~3, 5, 7 ms respectively. It also shows later peaks corresponding to thalamic and cortical activity at ~17 and 27 ms respectively.

999   The half-wave rectified waveforms and glottal pulse sequences were corrected for the small clock
1000 differences between the sound card and EEG system and then down-sampled to the EEG sampling
1001 frequency prior to deconvolution. To avoid temporal splatter due to standard downsampling, the pulse
1002 sequences were resampled by placing unit impulses at sample indices closest to each pulse time.
1003 Regularization was not necessary because the amplitude spectra of these regressors were sufficiently
1004 broadband. For efficiency, the time-domain response waveform, w, for a given 64 s epoch was calculated
1005 using frequency-domain division for the deconvolution, with the numerator the cross-spectral density
1006 (corresponding to the cross-correlation in the time domain) of the stimulus regressor and EEG response,
1007 and the denominator the power spectral density of the stimulus regressor (corresponding to its
1008 autocorrelation in the time domain). For a single epoch, that would be:

35

1009
$$w = \mathcal{F}^{-1}\left\{\frac{\mathcal{F}\{x\}^* \ \mathcal{F}\{y\}}{\mathcal{F}\{x\}^* \ \mathcal{F}\{x\}}\right\}$$

1010 where $\mathcal{F}$ denotes the fast Fourier transform, $\mathcal{F}^{-1}$ the inverse fast Fourier Transform, * complex conjugation,
1011 x the input stimulus regressor (half-wave rectified waveform or glottal pulse sequence), and y the EEG
1012 data for each epoch. We used methods incorporated into the MNE-python package (RRID:SCR_005972;
1013 Gramfort et al., 2013). In practice, we made adaptations to this formula to improve the SNR with Bayesian-
1014 like averaging (see below). For multiband peaky speech the same EEG was deconvolved with the pulse
1015 train of each band separately, and then with an additional 6 "fake" pulse trains to derive the common
1016 component across bands due to the pulse train coherence at low frequencies (shown in Figure 15). The
1017 averaged response across these 6 fake pulse trains, or common component, was then subtracted from the
1018 multiband responses to identify the frequency-specific band responses.

1019

1020 *Response averaging*

1021 The quality of the ABR waveforms as a function of each type of stimulus was of interest, so we
1022 calculated the averaged response after each 64 s epoch. We followed a Bayesian-like process (Elberling
1023 and Wahlgreen, 1985) to account for variations in noise level across the recording time (such as slow drifts
1024 or movement artifacts) and to avoid rejecting data based on thresholds. Each epoch was weighted by its
1025 inverse variance, $1/\sigma_i^2$, to the sum of the inverse variances of all epochs. Thus, epoch weights, b_i, were
1026 calculated as follows:

1027
$$b_i = \frac{1/\sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2}$$

1028 where i is the epoch number and n is the number of epochs collected. For efficiency, weighted averaging
1029 was completed during deconvolution. Because auto-correlation of the input stimulus (denominator of the
1030 frequency domain division) was similar across epochs it was averaged with equal weighting. Therefore, the
1031 numerator of the frequency domain division was summed across weighted epochs and the denominator
1032 averaged across epochs, according to the following formula:

1033
$$w = \mathcal{F}^{-1}\left\{\frac{\sum_{i=1}^{n} b_i \mathcal{F}\{x_i\}^* \ \mathcal{F}\{y_i\}}{\sum_{i=1}^{n} \frac{1}{n}\mathcal{F}\{x_i\}^* \ \mathcal{F}\{x_i\}}\right\}$$

1034

1035 where w is the average response waveform, i is again the epoch number, n is the number of epochs
1036 collected.

1037 Due to the circular nature of the discrete frequency domain deconvolution, the resulting response
1038 has an effective time interval of [0, 32] s at the beginning and [-32, 0) s at the end, so that concatenating
1039 the two – with the end first – yields the response from [-32, 32] s. Consequently, to avoid edge artifacts, all
1040 filtering was performed after the response was shifted to the middle of the 64 s time window. To remove
1041 high-frequency noise and some low-frequency noise, the average waveform was band-pass filtered
1042 between 30–2000 Hz using a first-order causal Butterworth filter. An example of this weighted average
1043 response to broadband peaky speech is shown in the right panel of Figure 17. This bandwidth of 30 to
1044 2000 Hz is sufficient to identify additional waves in the brainstem and middle latency responses (ABR and
1045 MLR respectively). To further identify earlier waves of the auditory brainstem responses (i.e., waves I and
1046 III), responses were high-pass filtered at 150 Hz using a first-order causal Butterworth filter. This filter was
1047 determined to provide the best morphology without compromising the response by comparing responses

1048 filtered with common high-pass cutoffs of 1, 30, 50, 100 and 150 Hz each combined with first, second and
1049 fourth order causal Butterworth filters.

1050

1051 *Response normalization*

1052   An advantage of this method over our previous one (Maddox and Lee, 2018) is that because the
1053 regressor comprises unit impulses, the deconvolved response is given in meaningful units which are the
1054 same as the EEG recording, namely microvolts. With a continuous regressor, like the half-wave rectified
1055 speech waveform, this is not the case. Therefore, to compare responses to half-wave rectified speech
1056 versus glottal pulses, we calculated a normalization factor, g, based on data from all subjects:

1057
$$g = \frac{1/n \; \sum_{i=1}^{n} \sigma_{u,i}}{1/n \; \sum_{i=1}^{n} \sigma_{p,i}}$$

1058 where n is the number of subjects, $\sigma_{u,i}$ is the SD of subject i's response to unaltered speech between 0–
1059 20 ms, and $\sigma_{p,i}$ is the same for the broadband peaky speech. Each subject's responses to unaltered
1060 speech were multiplied by this normalization factor to bring these responses within a comparable amplitude
1061 range as those to broadband peaky speech. Consequently, amplitudes were not compared between
1062 responses to unaltered and peaky speech. This was not our prime interest, rather we were interested in
1063 latency and presence of canonical component waves. In this study the normalization factor was 0.26, which
1064 cannot be applied to other studies because this number also depends on the scale when storing the digital
1065 audio. In our study, this unitless scale was based on a root-mean-square amplitude of 0.01. The same
1066 normalization factor was used when the half-wave rectified speech as used as the regressor with EEG
1067 collected in response to unaltered speech, broadband peaky speech and multiband peaky speech (Figure
1068 2, Figure 4, Figure 4-figure supplement 1).

1069

1070 *Response SNR calculation*

1071   We were also interested in the recording time required to obtain robust responses to re-synthesized
1072 peaky speech. Therefore, we calculated the time it took for the ABR and MLR to reach a 0-dB SNR. The
1073 SNR of each waveform in dB, SNR_w, was estimated as:

1074
$$SNR_w \; = 10 \log_{10} \left[\frac{\sigma_{S+N}^2 - \sigma_N^2}{\sigma_N^2} \right],$$

1075 where σ_{S+N}^2 represents the variance (i.e., mean-subtracted energy) of the waveform between 0 and 15 ms
1076 or 30 ms for the ABR and MLR respectively (contains both component signals as well as noise, $S + N$), and
1077 σ_N^2 represents the variance of the noise, N, estimated by averaging the variances of 15 ms (ABR) to 30 ms
1078 (MLR) segments of the pre-stimulus baseline between −480 and −20 ms. Then the SNR for 1 min of
1079 recording, SNR_{60}, was computed from the SNR_w as:

1080
$$SNR_{60} = \; SNR_w + 10 \log_{10}[60/t_w],$$

1081 where t_w is the duration of the recording in seconds, as specified in the "Speech stimuli and conditions"
1082 subsection. For example, in experiment 3, the average waveform resulted from 64 min of recording, or a t_w
1083 of 3,840 s. The time to reach 0 dB SNR for each subject, $t_{0dB \; SNR}$, was estimated from this SNR_{60} by:

1084
$$t_{0dB \; SNR} = 60 \; \times 10^{-SNR_{60}/10}.$$

1085 Cumulative density functions were used to show the proportion of subjects that reached an SNR ≥ 0 dB
1086 and to determine the necessary acquisition times that can be expected for each stimulus on a group level.

1087

37

Statistical Analyses

Data were checked for normality using the Shapiro-Wilk test. Waveform morphology of responses to different narrators was compared using Pearson correlations of the responses between 0 and 15 ms for the ABR waveforms or 0 and 40 ms for both ABR and MLR waveforms. The Wilcoxon signed rank test was used to determine whether narrator differences (waveform correlations) were significantly different than the correlations of the same EEG split into even and odd epochs with equal numbers of epochs from each narrator. The intraclass correlation coefficient type 3 (absolute agreement) was used to verify good agreement in peak latencies chosen by an experienced audiologist and neuroscientist (MJP) at two different time points, 3 months apart. Independent t-tests with $\mu = 0$ were conducted on the peak latency differences of ABR/MLR waves for unaltered and broadband peaky speech. For multiband peaky speech, the component wave peak latency changes across frequency band were assessed with power law regression (Harte et al., 2009; Neely et al., 1988; Rasetshwane et al., 2013; Strelcyk et al., 2009), conducted in the log-log domain with linear mixed effects regression using the lme4 (RRID:SCR_015654) and lmerTest (RRID:SCR_015656) packages in R (RRID:SCR_001905) and RStudio (RRID:SCR_000432) (Bates et al., 2015; Kuznetsova et al., 2017; R Core Team, 2020). The parameter a of the power law regression was estimated by adding an assumed synaptic delay of 0.8 (Eggermont, 1979; Strelcyk et al., 2009) to the I-V inter-wave delay from the subjects' responses to broadband peaky speech. For subjects who did not have an identifiable wave I in the broadband peaky response, the group mean I-V delay was used – this occurred for 1 of 22 subjects in experiment 1, and 2 of 11 subjects for responses to the female narrator in experiment 2. Because only multiband peaky speech was presented in experiment 3, the mean I-V intervals from experiment 2 were used for each subject for experiment 3. Random effects of subject and each frequency band term were included to account for individual variability that is not generalizable to the fixed effects. A power analysis was completed using the simR package (RRID:SCR_019287; Green and MacLeod, 2016), which uses a likelihood ratio test on 1000 Monte Carlo permutations of the response variables based on the fitted model.

ACKNOWLEDGMENTS

FUNDING

DATA AVAILABILITY

Python code is available on the lab GitHub account (https://github.com/maddoxlab/peaky-speech). All EEG recordings are available in the EEG-BIDS format (Pernet et al., 2019) on Dryad (https://doi.org/10.5061/dryad.12jm63xwd). Stimulus files and python code necessary to derive the peaky speech responses are deposited to the same Dryad repository.

REFERENCES

Abdala C, Folsom RC. 1995. Frequency contribution to the click-evoked auditory brain-stem response in human adults and infants. *J Acoust Soc Am* **97**:2394–2404. doi:10.1121/1.411961

Backer KC, Kessler AS, Lawyer LA, Corina DP, Miller LM. 2019. A novel EEG paradigm to simultaneously and rapidly assess the functioning of auditory and visual pathways. *J Neurophysiol* **122**:1312–1329. doi:10.1152/jn.00868.2018

Bajo VM, King AJ. 2012. Cortical modulation of auditory processing in the midbrain. *Front Neural Circuits* **6**:114. doi:10.3389/fncir.2012.00114

Bajo VM, Nodal FR, Moore DR, King AJ. 2010. The descending corticocollicular pathway mediates learning-induced auditory plasticity. *Nat Neurosci* **13**:253–260. doi:10.1038/nn.2466

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting Linear Mixed-Effects Models Using lme4. *J Stat Softw* **67**:1–48.

Bernstein JGW, Grant KW. 2009. Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. *J Acoust Soc Am* **125**:3358–3372. doi:10.1121/1.3110132

Bharadwaj HM, Verhulst S, Shaheen L, Liberman MC, Shinn-Cunningham BG. 2014. Cochlear neuropathy and the coding of supra-threshold sound. *Front Syst Neurosci* **8**. doi:10.3389/fnsys.2014.00026

Boersma P, Weenink D. 2018. Praat: doing phonetics by computer.

Bramhall N, Beach EF, Epp B, Le Prell CG, Lopez-Poveda EA, Plack CJ, Schaette R, Verhulst S, Canlon B. 2019. The search for noise-induced cochlear synaptopathy in humans: Mission impossible? *Hear Res* **377**:88–103. doi:10.1016/j.heares.2019.02.016

Burkard R, Hecox K. 1983. The effect of broadband noise on the human brainstem auditory evoked response. I. Rate and intensity effects. *J Acoust Soc Am* **74**:1204–1213. doi:10.1121/1.390024

Burkard R, Shi Y, Hecox KE. 1990. A comparison of maximum length and Legendre sequences for the derivation of brain-stem auditory-evoked responses at rapid rates of stimulation. *J Acoust Soc Am* **87**:1656–1664. doi:10.1121/1.399413

Carney LH, Li T, McDonough JM. 2015. Speech Coding in the Brain: Representation of Vowel Formants by Midbrain Neurons Tuned to Sound Fluctuations,,. *eNeuro* **2**. doi:10.1523/ENEURO.0004-15.2015

Chiappa KH, Gladstone KJ, Young RR. 1979. Brain Stem Auditory Evoked Responses: Studies of Waveform Variations in 50 Normal Human Subjects. *Arch Neurol* **36**:81–87. doi:10.1001/archneur.1979.00500380051005

Dau T, Wegner O, Mellert V, Kollmeier B. 2000. Auditory brainstem responses with optimized chirp signals compensating basilar-membrane dispersion. *J Acoust Soc Am* **107**:1530–1540. doi:10.1121/1.428438

Don M, Allen AR, Starr A. 1977. Effect of Click Rate on the Latency of Auditory Brain Stem Responses in Humans. *Ann Otol Rhinol Laryngol* **86**:186–195. doi:10.1177/000348947708600209

Easwar V, Purcell DW, Aiken SJ, Parsa V, Scollie SD. 2015. Evaluation of Speech-Evoked Envelope Following Responses as an Objective Aided Outcome Measure: Effect of Stimulus Level, Bandwidth, and Amplification in Adults With Hearing Loss. *Ear Hear* **36**:635–652. doi:10.1097/AUD.0000000000000199

Easwar V, Scollie S, Aiken S, Purcell D. 2020. Test-Retest Variability in the Characteristics of Envelope Following Responses Evoked by Speech Stimuli. *Ear Hear* **41**:150–164. doi:10.1097/AUD.0000000000000739

Eggermont JJ. 1979. Narrow-band AP latencies in normal and recruiting human ears. *J Acoust Soc Am* **65**:463–470. doi:10.1121/1.382345

Elberling C, Don M. 2008. Auditory brainstem responses to a chirp stimulus designed from derived-band latencies in normal-hearing subjects. *J Acoust Soc Am* **124**:3022–3037. doi:10.1121/1.2990709

Elberling C, Wahlgreen O. 1985. Estimation of Auditory Brainstem Response, Abr, by Means of Bayesian Inference. *Scand Audiol* **14**:89–96. doi:10.3109/01050398509045928

Forte AE, Etard O, Reichenbach T. 2017. The human auditory brainstem response to running speech reveals a subcortical mechanism for selective attention. *eLife* **6**:e27203. doi:10.7554/eLife.27203

Geisler CD, Frishkopf LS, Rosenblith WA. 1958. Extracranial Responses to Acoustic Clicks in Man. *Science* **128**:1210–1211. doi:10.1126/science.128.3333.1210

Goldstein R, Rodman LB. 1967. Early components of averaged evoked responses to rapidly repeated auditory stimuli. *J Speech Hear Res* **10**:697–705. doi:10.1044/jshr.1004.697

Gorga MP, Johnson TA, Kaminski JR, Beauchaine KL, Garner CA, Neely ST. 2006. Using a Combination of Click- and Tone Burst–Evoked Auditory Brain Stem Response Measurements to Estimate Pure-Tone Thresholds. *Ear Hear* **27**:60–74. doi:10.1097/01.aud.0000194511.14740.9c

Gorga MP, Kaminski JR, Beauchaine KA, Jesteadt W. 1988. Auditory brainstem responses to tone bursts in normally hearing subjects. *J Speech Hear Res* **31**:87–97. doi:10.1044/jshr.3101.87

Gorga MP, Kaminski JR, Beauchaine KL, Bergman BM. 1993. A Comparison of Auditory Brain Stem Response Thresholds and latencies Elicited by Air- and Bone-Conducted Stimuli. *Ear Hear* **14**:85–94.

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Goj R, Jas M, Brooks T, Parkkonen L, Hämäläinen M. 2013. MEG and EEG data analysis with MNE-Python. *Front Neurosci* **7**. doi:10.3389/fnins.2013.00267

Grant KW, Tufts JB, Greenberg S. 2007. Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing-impaired individuals. *J Acoust Soc Am* **121**:1164–1176. doi:10.1121/1.2405859

Green P, MacLeod CJ. 2016. SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol Evol* **7**:493–498. doi:10.1111/2041-210X.12504

Grothe B, Pecka M. 2014. The natural history of sound localization in mammals--a story of neuronal inhibition. *Front Neural Circuits* **8**:116. doi:10.3389/fncir.2014.00116

Harte JM, Pigasse G, Dau T. 2009. Comparison of cochlear delay estimates using otoacoustic emissions and auditory brainstem responses. *J Acoust Soc Am* **126**:1291–1301. doi:10.1121/1.3168508

Hashimoto I. 1982. Auditory evoked potentials from the human midbrain: slow brain stem responses. *Electroencephalogr Clin Neurophysiol* **53**:652–657. doi:10.1016/0013-4694(82)90141-9

Hyde M. 2008. Ontario Infant Hearing Program Audiologic Assessment Protocol Version 3.1.

Jiang ZD, Wu YY, Wilkinson AR. 2009. Age-related changes in BAER at different click rates from neonates to adults. *Acta Paediatr Oslo Nor 1992* **98**:1284–1287. doi:10.1111/j.1651-2227.2009.01312.x

Kileny P, Paccioretti D, Wilson AF. 1987. Effects of cortical lesions on middle-latency auditory evoked responses (MLR). *Electroencephalogr Clin Neurophysiol* **66**:108–120. doi:10.1016/0013-4694(87)90180-5

Kuznetsova A, Brockhoff PB, Christensen RHB. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *J Stat Softw* **82**:1–26. doi:10.18637/jss.v082.i13

Larson E, McCloy D, Maddox R, Pospisil D. 2014. expyfun: Python experimental paradigm functions, version 2.0.0. Zenodo. doi:10.5281/zenodo.11640

1215   L'Engle M. 2012. A Wrinkle in Time: 50th Anniversary Commemorative Edition. Macmillan.

1216   Liberman MC, Epstein MJ, Cleveland SS, Wang H, Maison SF. 2016. Toward a Differential Diagnosis of
1217   Hidden Hearing Loss in Humans. *PLOS ONE* **11**:e0162726. doi:10.1371/journal.pone.0162726

1218   Maddox RK, Lee AKC. 2018. Auditory Brainstem Responses to Continuous Natural Speech in Human
1219   Listeners. *eNeuro* **5**. doi:10.1523/ENEURO.0441-17.2018

1220   Mesgarani N, David SV, Fritz JB, Shamma SA. 2009. Influence of Context and Behavior on Stimulus
1221   Reconstruction From Neural Activity in Primary Auditory Cortex. *J Neurophysiol* **102**:3329–3339.
1222   doi:10.1152/jn.91128.2008

1223   Miller L, IV BM, Bishop C. 2017. Frequency-multiplexed speech-sound stimuli for hierarchical neural
1224   characterization of speech processing. US20170196519A1.

1225   Møller AR, Jannetta PJ. 1983. Interpretation of brainstem auditory evoked potentials: results from
1226   intracranial recordings in humans. *Scand Audiol* **12**:125–133.

1227   Moore JK. 1987. The human auditory brain stem as a generator of auditory evoked potentials. *Hear Res*
1228   **29**:33–43. doi:10.1016/0378-5955(87)90203-6

1229   Neely ST, Norton SJ, Gorga MP, Jesteadt W. 1988. Latency of auditory brain-stem responses and
1230   otoacoustic emissions using tone-burst stimuli. *J Acoust Soc Am* **83**:652–656.
1231   doi:10.1121/1.396542

1232   O'Sullivan AE, Lim CY, Lalor EC. 2019. Look at me when I'm talking to you: Selective attention at a
1233   multisensory cocktail party can be decoded using stimulus reconstruction and alpha power
1234   modulations. *Eur J Neurosci* **50**:3282–3295. doi:10.1111/ejn.14425

1235   Pernet CR, Appelhoff S, Gorgolewski KJ, Flandin G, Phillips C, Delorme A, Oostenveld R. 2019. EEG-
1236   BIDS, an extension to the brain imaging data structure for electroencephalography. *Sci Data* **6**:103.
1237   doi:10.1038/s41597-019-0104-8

1238   Picton TW, Hillyard SA, Krausz HI, Galambos R. 1974. Human auditory evoked potentials. I. Evaluation of
1239   components. *Electroencephalogr Clin Neurophysiol* **36**:179–190. doi:10.1016/0013-4694(74)90155-
1240   2

1241   Polonenko MJ, Maddox RK. 2019. The Parallel Auditory Brainstem Response. *Trends Hear*
1242   **23**:2331216519871395. doi:10.1177/2331216519871395

1243   Prendergast G, Guest H, Munro KJ, Kluk K, Léger A, Hall DA, Heinz MG, Plack CJ. 2017. Effects of noise
1244   exposure on young adults with normal audiograms I: Electrophysiology. *Hear Res* **344**:68–81.
1245   doi:10.1016/j.heares.2016.10.028

1246   R Core Team. 2020. R: A language and environment for statistical computing. Vienna, Austria: R
1247   Foundation for Statistical Computing.

1248   Rasetshwane DM, Argenyi M, Neely ST, Kopun JG, Gorga MP. 2013. Latency of tone-burst-evoked
1249   auditory brain stem responses and otoacoustic emissions: Level, frequency, and rise-time effects. *J
1250   Acoust Soc Am* **133**:2803–2817. doi:10.1121/1.4798666

1251   Rudnicki M, Schoppe O, Isik M, Völk F, Hemmert W. 2015. Modeling auditory coding: from sound to
1252   spikes. *Cell Tissue Res* **361**:159–175. doi:10.1007/s00441-015-2202-z

1253   Saiz-Alía M, Forte AE, Reichenbach T. 2019. Individual differences in the attentional modulation of the
1254   human auditory brainstem response to speech inform on speech-in-noise deficits. *Sci Rep* **9**:1–10.
1255   doi:10.1038/s41598-019-50773-1

1256 Saiz-Alia M, Reichenbach T. 2020. Computational modeling of the auditory brainstem response to
1257   continuous speech. *J Neural Eng*. doi:10.1088/1741-2552/ab970d

1258 Scott M. 2007. The Alchemyst: the secrets of the immortal Nicholas Flamel, Book 1. New York: Listening
1259   Library.

1260 Shore SE, Nuttall AL. 1985. High-synchrony cochlear compound action potentials evoked by rising
1261   frequency-swept tone bursts. *J Acoust Soc Am* **78**:1286–1295. doi:10.1121/1.392898

1262 Stapells DR, Oates P. 1997. Estimation of the Pure-Tone Audiogram by the Auditory Brainstem Response:
1263   A Review. *Audiol Neurotol* **2**:257–280. doi:10.1159/000259252

1264 Starr A, Hamilton AE. 1976. Correlation between confirmed sites of neurological lesions and abnormalities
1265   of far-field auditory brainstem responses. *Electroencephalogr Clin Neurophysiol* **41**:595–608.
1266   doi:10.1016/0013-4694(76)90005-5

1267 Strelcyk O, Christoforidis D, Dau T. 2009. Relation between derived-band auditory brainstem response
1268   latencies and behavioral frequency selectivity. *J Acoust Soc Am* **126**:1878–1888.
1269   doi:10.1121/1.3203310

1270 Teoh ES, Lalor EC. 2019. EEG decoding of the target speaker in a cocktail party scenario: considerations
1271   regarding dynamic switching of talker location. *J Neural Eng* **16**:036017. doi:10.1088/1741-
1272   2552/ab0cf1

1273 Verhulst S, Altoè A, Vasilkov V. 2018. Computational modeling of the human auditory periphery: Auditory-
1274   nerve responses, evoked potentials and hearing loss. *Hear Res*, Computational models of the
1275   auditory system **360**:55–75. doi:10.1016/j.heares.2017.12.018

1276 Winer JA. 2005. Decoding the auditory corticofugal systems. *Hear Res* **207**:1–9.
1277   doi:10.1016/j.heares.2005.06.007

1278 Zilany MSA, Bruce IC, Carney LH. 2014. Updated parameters and expanded simulation options for a
1279   model of the auditory periphery. *J Acoust Soc Am* **135**:283–286. doi:10.1121/1.4837815

1280

1281 **FIGURE CAPTIONS**

1282 **Figure 1.** Single subject and group average (bottom right) weighted-average auditory brainstem responses
1283 (ABR) to ~43 minutes of broadband peaky speech. Area for the group average shows ± 1 SEM.
1284 Responses were high-pass filtered at 150 Hz using a first order Butterworth filter. Waves I, III, and V of the
1285 canonical ABR are evident in most of the single subject responses (N = 22, 16, 22 respectively), and are
1286 marked by the average peak latencies on the average response.

1287 **Figure 2.** Comparison of auditory brainstem (ABR) and middle latency responses (MLR) to ~43 minutes
1288 each of unaltered speech and broadband peaky speech. (A) The average waveform to broadband peaky
1289 speech (blue) shows additional, and sharper, waves of the canonical ABR and MLR than the broader average
1290 waveform to unaltered speech (black). Responses were high-pass filtered at 30 Hz with a first order
1291 Butterworth filter. Areas show ± 1 SEM. (B) Comparison of peak latencies for ABR wave V (circles) and MLR
1292 waves N_a (downward triangles) and P_a (upward triangles) that were common between responses to
1293 broadband peaky and unaltered speech. Blue symbols depict individual subjects and black symbols depict
1294 the mean.

1295 **Figure 3.** Comparison of grand average (N = 22) measured and modeled responses to ~43 minutes of
1296 broadband peaky speech. Amplitudes of the linear (dashed line) and auditory nerve (AN; dotted line)
1297 modelled responses were in arbitrary units, and thus scaled to match the amplitude of the measured
1298 response (solid line) over the 0–20 ms lags. The pre-stimulus component was present in all three responses

1299 using a first order 30 Hz high-pass Butterworth filter (top row), but was minimized by aggressive high-pass
1300 filtering with a second order 200 Hz high-pass Butterworth filter (bottom row).

1301 **Figure 4.** Comparison of responses derived by using the same type of regressor in the deconvolution.
1302 Average waveforms (areas show ± 1 SEM) are shown for ~43 minutes each of unaltered speech (black) and
1303 broadband peaky speech (blue). EEG was regressed with the (A) half-wave rectified audio and (B) pulse
1304 train. Responses were high-pass filtered at 30 Hz using a first order Butterworth filter. **Figure 5.** Comparison
1305 of responses to 32 minutes each of male (dark blue) and female (light blue) narrated re-synthesized
1306 broadband peaky speech. (A) Average waveforms across subjects (areas show ± 1 SEM) are shown for
1307 auditory brainstem response (ABR) time lags with high-pass filtering at 150 Hz (top), and both ABR and
1308 middle latency response (MLR) time lags with a lower high-pass filtering cutoff of 30 Hz (bottom). (B)
1309 Histograms of the correlation coefficients between responses evoked by male- and female-narrated
1310 broadband peaky speech during ABR (top) and ABR/MLR (bottom) time lags. Solid lines denote the median
1311 and dotted lines the inter-quartile range. (C) Comparison of ABR (top) and MLR (bottom) wave peak latencies
1312 for individual subjects (gray) and the group mean (black). ABR and MLR responses were similar to both
1313 types of input but are smaller for female-narrated speech, which has a higher glottal pulse rate. Peak
1314 latencies for female-evoked speech were delayed during ABR time lags but faster for early MLR time lags.

1315 **Figure 6.** Comparison of responses to ~43 minutes of male-narrated multiband peaky speech. (A) Average
1316 waveforms across subjects (areas show ± 1 SEM) are shown for each band (colored solid lines) and for the
1317 common component (dot-dash gray line, same waveform replicated as a reference for each band), which
1318 was calculated using 6 false pulse trains. (B) The common component was subtracted from each band's
1319 response to give the frequency-specific waveforms (areas show ± 1 SEM), which are shown with high-pass
1320 filtering at 30 Hz (solid lines) and 150 Hz (dashed lines). (C) Mean ± SEM peak latencies for each wave
1321 decreased with increasing band frequency. Numbers of subjects with an identifiable wave are given for
1322 each wave and band. Details of the mixed effects models for (C) are provided in Supplementary File 1A.

1323 **Figure 7.** Comparison of responses to ~43 minutes of male-narrated peaky speech in the same subjects.
1324 Average waveforms across subjects (areas show ± 1 SEM) are shown for broadband peaky speech (blue)
1325 and for the summed frequency-specific responses to multiband peaky speech with the common component
1326 added (red), high-pass filtered at 150 Hz (left) and 30 Hz (right). Regressors in the deconvolution were
1327 pulse trains.

1328 **Figure 8.** Comparison of responses to 32 minutes each of male- and female-narrated re-synthesized
1329 multiband peaky speech. (A) Average frequency-specific waveforms across subjects (areas show ± 1 SEM;
1330 common component removed) are shown for each band in response to male- (dark red lines) and female-
1331 narrated (light red lines) speech. Responses were high-pass filtered at 30 Hz (left) and 150 Hz (right) to
1332 highlight the MLR and ABR respectively. (B) Correlation coefficients between responses evoked by male-
1333 and female-narrated multiband peaky speech during ABR/MLR (left) and ABR (right) time lags for each
1334 frequency band. Black lines denote the median. (C) Mean ± SEM peak latencies for male- (dark) and
1335 female- (light) narrated speech for each wave decreased with increasing frequency band. Numbers of
1336 subjects with an identifiable wave are given for each wave, band and narrator. Lines are given a slight
1337 horizontal offset to make the error bars easier to see. Details of the mixed effects models for (C) are
1338 provided in Supplementary File 1B.

1339 **Figure 9.** Comparison of responses to ~60 minutes each of male- and female-narrated dichotic multiband
1340 peaky speech with standard audiological frequency bands. (A) Average frequency-specific waveforms
1341 across subjects (areas show ± 1 SEM; common component removed) are shown for each band for the left
1342 ear (dotted lines) and right ear (solid lines). Responses were high-pass filtered at 30 Hz. (B) Left-right ear
1343 correlation coefficients (top, averaged across gender) and male-female correlation coefficients (bottom,
1344 averaged across ear) during ABR time lags (0–15 ms) for each frequency band. Black lines denote the
1345 median. (C) Mean ± SEM wave V latencies for male- (dark red) and female-narrated (light red) speech for

1346   the left (dotted line, cross symbol) and right ear (solid line, circle symbol) decreased with increasing
1347   frequency band. Lines are given a slight horizontal offset to make the error bars easier to see. Details of
1348   the mixed effects model for (C) are provided in Supplementary File 1C.

1349   **Figure 10.** Cumulative proportion of subjects who have responses with ≥ 0 dB SNR as a function of
1350   recording time. Time required for unaltered (black) and broadband peaky speech (dark blue) of a male
1351   narrator are shown for 22 subjects in the left plot, and for male (dark blue) and female (light blue)
1352   broadband peaky speech is shown for 11 subjects in the right plot. Solid lines denote SNRs calculated
1353   using variance of the signal high-pass filtered at 30 Hz over the ABR/MLR interval 0–30 ms, and dashed
1354   lines denote SNR variances calculated on signals high-pass filtered at 150 Hz over the ABR interval 0–15
1355   ms. Noise variance was calculated in the pre-stimulus interval −480 to −20 ms.

1356   **Figure 11.** Cumulative proportion of subjects who have frequency-specific responses (common component
1357   subtracted) with ≥ 0 dB SNR as a function of recording time. Acquisition time was faster for male (left) than
1358   female (right) narrated multiband peaky speech with (A) 4 frequency bands presented diotically, and with
1359   (B) 5 frequency bands presented dichotically (total of 10 responses, 5 bands in each ear). SNR was
1360   calculated by comparing variance of signals high-pass filtered at 150 Hz across the ABR interval of 0–15
1361   ms to variance of noise in the pre-stimulus interval −480 to −20 ms.

1362   **Figure 12.** The range of lags can be extended to allow early, middle and late latency responses to be
1363   analyzed from the same recording to broadband peaky speech. Average waveforms across subjects (areas
1364   show ± 1 SEM) are shown for responses measured to 32 minutes of broadband peaky speech narrated by
1365   a male (dark blue) and female (light blue). Responses were high-pass filtered at 1 Hz using a first order
1366   Butterworth filter, but different filter parameters can be used to focus on each stage of processing.
1367   Canonical waves of the ABR, MLR and LLR are labeled for the male-narrated speech. Due to adaptation,
1368   amplitudes of the late potentials are smaller than typically seen with other stimuli that are shorter in
1369   duration with longer inter-stimulus intervals than our continuous speech. Waves I and III become more
1370   clearly visible by applying a 150 Hz high-pass cutoff.

1371   **Figure 13.** Unaltered speech waveform (top left) and spectrogram (top right) compared to re-synthesized
1372   broadband peaky speech (middle left and right) and multiband peaky speech (bottom left and right).
1373   Comparing waveforms shows that the peaky speech is as "click-like" as possible, while comparing the
1374   spectrograms shows that the overall spectrotemporal content that defines speech is basically unchanged
1375   by the re-synthesis. A naïve listener is unlikely to notice that any modification has been performed, and
1376   subjective listening confirms the similarity. Yellow/lighter colors represent larger amplitudes than
1377   purple/darker colors in the spectrogram. See supplementary files for audio examples of each stimulus type
1378   for both narrators.

1379   **Figure 14.** Relative mean-squared magnitude in decibels of multiband peaky speech with 4 filter bands
1380   (left) and 5 filter bands (right) for male-(blue) and female-(orange) narrated speech. The full audio
1381   comprises unvoiced and re-synthesized voiced sections, which was presented to the subjects during the
1382   experiments. The other bands reflect the relative magnitude of the voiced sections (voiced only), and each
1383   filtered frequency band.

1384   **Figure 15.** Spectral coherence of pulse trains for multiband peaky speech narrated by a male (left) and
1385   female (right). Spectral coherence was computed across 1 s slices from 60 unique 64 s multiband peaky
1386   speech segments (3,840 total slices) for each combination of bands. Each light gray line represents the
1387   coherence for one band comparison. (A) There were 45 comparisons across the 10-band (audiological)
1388   speech used in experiment 3 (5 frequency bands x 2 ears). The lowest band was unshifted and the other 9
1389   bands had static frequency shifts. (B) There were 6 comparisons across 4 pulse trains of the bands in the
1390   pilot experiment, which all had dynamic random frequency shifts. Pulse trains (i.e., the input stimuli, or
1391   regressors, for the deconvolution) were frequency-dependent (coherent) below 72 Hz for the male
1392   multiband speech and 126 Hz for the female multiband speech.

Figure 16. Octave band filters used to create re-synthesized broadband peaky speech (left, blue), diotic multiband peaky speech with 4 bands (middle, red), and dichotic multiband peaky speech using 5 bands with audiological center frequencies (right, red). The last band (2nd, 5th, 6th respectively, black line) was used to filter the high-frequencies of unaltered speech during mixing to improve the quality of voiced consonants. The designed frequency response using trapezoids (top) were converted into the time-domain using IFFT, shifted and Nuttall windowed to create impulse responses (middle), which were then used to assess the actual frequency response by converting into the frequency domain using FFT (bottom).

Figure 17. Left: A segment of broadband peaky speech stimulus (top) and the corresponding glottal pulse train (bottom) used in calculating the broadband peaky speech response. Right: An example broadband peaky speech response from a single subject. The response shows ABR waves I, III, and V at ~3, 5, 7 ms respectively. It also shows later peaks corresponding to thalamic and cortical activity at ~17 and 27 ms respectively.

FIGURE SUPPLEMENTS

Figure 4-figure supplement 1. Comparison of responses derived by using the half-wave rectified audio as the regressor in the deconvolution with electroencephalography (EEG) recorded in response to ~43 minutes of speech. Average waveforms (areas show ± 1 SEM) are shown for EEG recorded to unaltered speech (black) relative to EEG recorded to multiband peaky speech (red). Responses were high-pass filtered at 30 Hz using a first order Butterworth filter.

Figure 6-figure supplement 1. Comparison of responses to 64 minutes each of male- (left) and female-narrated (right) multiband peaky speech created with the dynamic random frequency shift method. (A) Weighted-average waveforms for one subject are shown for each band (colored solid lines) and for the common component (dot-dashed gray line, same waveform replicated as a reference for each band), which was calculated using 6 false pulse trains. (B) The common component was subtracted from each band's response to give the frequency-specific waveforms, which are shown with high-pass filtering at 30 Hz (solid lines) and 150 Hz (dashed lines). Responses from all four bands show more consistent resemblance to the common component, indicating that this method is effective at reducing stimulus-related bias. However, differences still remain in the lowest frequency band for latencies >30 ms, suggesting that this new method reveals true underlying low-frequency neural activity that is unique.

Figure 15-figure supplement 1. Comparison of the common component derived from the average response to 6 fake pulse trains that were created using static frequency shifts (solid, darker lines; used in the paper) or dynamic random frequency shifts (dashed, lighter lines, pilot data and suggested in methods). Responses were to 32 minutes each of male- (left) and female- (right) narrated re-synthesized diotic multiband peaky speech. Areas show ±1 SEM. The EEG to diotic multiband peaky speech (4-bands) was regressed with 6 fake pulse trains created using the static shifts (used in the paper; the same common component displayed in Figure 4), as well as 6 fake pulse trains created using the dynamic random frequency shift method (used to create the common component in Supplementary Figure 2).

Figure 15-figure supplement 2. Multiband stimuli responses for male (left) and female (right) derived by deconvolving the absolute value of the dichotic (stereo) multiband peaky audio (from experiment 3) with the 10 associated pulse trains– 5 pulse trains were used for each band in each ear ("correct" pulse trains, top row). Each pulse train acted as a "wrong" pulse train for the associated band in the other ear (bottom row). Left ear responses are shown by dotted lines and right ear responses by solid lines. Higher frequency bands are shown by lighter red colors. The non-zero responses only occurred when the correct pulse train was paired with the correct audio. Both male- and female-narrated speech responses symmetrically surrounded zero ms and were largest for the first (lowest frequency) band when the correct, but not fake, pulse trains were used as the regressor in the deconvolution.

1439

AUDIO FILES

Audio 1. Unaltered speech sample from the male narrator (*The Alchemyst*; Scott, 2007).

Audio 2. Broadband peaky speech sample from the male narrator (*The Alchemyst*; Scott, 2007).

Audio 3. Multiband peaky speech sample from the male narrator (*The Alchemyst*; Scott, 2007).

Audio 4. Unaltered speech sample from the female narrator (*A Wrinkle in Time*; L'Engle, 2012).

Audio 5. Broadband peaky speech sample from the female narrator (*A Wrinkle in Time*; L'Engle, 2012).

Audio 6. Multiband peaky speech sample from the female narrator (*A Wrinkle in Time*; L'Engle, 2012).

SUPPLEMENTARY FILES

Supplementary File 1A. LMER model formula and summary output for multiband peaky speech in experiment 1.

Supplementary File 1B. LMER model formula and summary output for multiband peaky speech in experiment 2.

Supplementary File 1C. LMER model formula and summary output for multiband peaky speech in experiment 3.