

1 The effect of stimulus duration on preferences for gain adjustments in speech
2 Whitmer, William M.^{a, b}
3 Caswell-Midwinter, Benjamin^{a, b, c}
4 Naylor, Graham^a
5 a) Hearing Sciences – Scottish Section, School of Medicine, University of
6 Nottingham, Glasgow, UK.
7 b) Institute of Health and Wellbeing, College of Medical, Veterinary and Life
8 Science, University of Glasgow, Glasgow, UK.
9 c) Otolaryngology – Head and Neck Surgery, Massachusetts Eye and Ear, Harvard
10 Medical School, Boston, Massachusetts, USA
11 Corresponding author: bill.whitmer@nottingham.ac.uk
12 Word count: 5768

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

13 Abstract

14 Objectives

15 In the personalisation of hearing aid fittings, gain is often clinically adjusted to
16 patient preferences using live speech. When using brief sentences as stimuli, the
17 minimum gain adjustments necessary to elicit preferences ('preference thresholds')
18 were previously found to be much greater than typical adjustments in current
19 practice. The current study examined the role of duration on preference thresholds.

20 Design

21 Participants heard 2, 4 and 6-s segments of a continuous monologue presented in
22 pairs. Participants judged whether the second stimulus of each pair, with a ± 0 -12 dB
23 gain adjustment in one of three frequency bands, was "better", "worse" or "no
24 different" from the first at their individual real-ear or prescribed gain.

25 Study Sample

26 Twenty-nine adults, all with hearing-aid experience.

27 Results

28 The minimum gain adjustments to elicit "better" or "worse" judgments decreased
29 with increasing duration for most adjustments. Inter-participant agreement and
30 intra-participant reliability increased with increasing duration. The effect of duration,
31 however, decreased with increasing duration, with no increase in agreement or
32 reliability for 6-s vs. 4-s segments.

33 Conclusions

34 Providing longer stimuli improves the likelihood of patients providing reliable
35 judgments of hearing-aid gain adjustments, but the effect is limited, and alternative
36 fitting methods may be more viable for effective hearing-aid personalisation.

37 Introduction

38 In the treatment of hearing loss, clinicians fit hearing aids to reach a balance
39 between audibility and comfort for each patient. The balancing act begins with
40 prescribed gains across frequencies based on each patient's pure-tone thresholds.
41 These prescribed gains, based on average data, are then personalised through
42 adjustments made by the clinician using patient feedback (Anderson et al., 2018;
43 Jenstad et al., 2003; Kuk, 1999; Thielemans et al., 2017). The patient's feedback is
44 often based solely on the effect the adjustments have on the perception of the
45 clinician's voice, the most readily available stimulus in any clinic.

46 We have previously shown what gain adjustments are discriminable for short
47 sentences presented in quiet. Median just-noticeable differences (JNDs) for gain
48 increments in broad low-, mid- and high-frequency bands were 4, 4 and 7 dB,
49 respectively (Caswell-Midwinter and Whitmer, 2019). Using the same speech corpus,
50 we have subsequently shown what gain adjustments are necessary to elicit
51 preferences (Caswell-Midwinter and Whitmer, 2020). Median preference thresholds
52 ranged from 4-12 dB for gain decrements and 5-9 dB for increments in the same
53 broad low-, mid-, and high-frequency bands. In Caswell-Midwinter and Whitmer
54 (2019), it was posited that the greater JNDs for speech in quiet re speech-shaped
55 noise were due to the spectro-temporal sparsity of the speech. That is, for a given
56 gain adjustment in any given band, the clean speech signal provided a smaller
57 number of glimpses of the adjustment than speech plus noise. In Caswell-Midwinter
58 and Whitmer (2020), it was further hypothesised that the large preference thresholds
59 were due in part to the short duration of the stimuli. Although patients typically
60 make quick comparisons on adjustments in the clinic, audiologists may talk for
61 longer, which might elicit more frequent and reliable preferences.

62 Previous psychophysical research has shown durational effects on level
63 discriminability, albeit mostly limited to short pure-tone stimuli. Increasing the
64 duration of a 0.5 or 8-kHz tone up to 2 s can improve level discrimination in normal-
65 hearing listeners (Florentine, 1986), and improves discrimination in fixed and roving
66 pedestal level conditions (Oxenham and Buus, 2000). For the discrimination of a
67 tone's level within a complex (i.e., profile analysis), performance improves up to a
68 duration of 100 ms (Green et al., 1984; Dai & Green, 1993). The ability to
69 discriminate a gain adjustment in particular band(s) of speech bears partial
70 resemblance to increment detection, the detection of a temporary increase or 'bump'
71 in level in an ongoing sound. Valente et al. (2011) showed that increasing the
72 duration of the standard tone decreased the threshold more than increasing the
73 duration of the increment of a tone. In all past studies of level discrimination and
74 increment detection with varying duration, though, performance improves with
75 frequency (e.g., Moore et al., 1997), whereas the discriminability of gain adjustments
76 decreases with the frequency band of the adjustment for speech (Caswell-Midwinter
77 and Whitmer, 2019). There is some evidence of a duration effect with broadband
78 stimuli: studying the detection of an 8-dB peak at 3.5 kHz in a broadband noise,
79 Farrar et al. (1987) found that thresholds decreased as duration increased up to 300

80 ms, the maximum duration tested. Isarangura et al. (2019) found that spectral
81 modulation detection thresholds in a broadband noise carrier also decreased with
82 increasing duration but were asymptotic by 200 ms. For speech stimuli, the evidence
83 of duration effects on level discrimination is scant; in a study of overall level
84 discrimination of speech, the threshold for words (mean duration 450 ms) was only
85 significantly worse (greater) than for sentences (mean duration 1533 ms) when
86 participants were aided (Whitmer and Akeroyd, 2011).

87 In sound-quality evaluations such as comparing the adjustments of hearing-aid
88 settings, a balance must be struck in sound-sample duration. The sample must be
89 long enough in order to perceive the acoustic changes, but also short enough to be
90 able to compare the adjusted sound with the previous (reference) sound.
91 International Telecommunication Union (ITU) recommendations for subjective
92 sound-quality evaluations note that for paired comparisons, durations should not
93 exceed 15-20 s due to “short-term human memory limitations,” but can be “a few
94 seconds” (ITU, 2019, p. 6). These memory limitations – the ability to maintain
95 features of the first sound for comparison to the second (e.g., auditory sensory
96 memory trace; Sams et al., 1993) – are often measured as the effect of the inter-
97 stimulus interval (ISI) duration. In the clinic, the adjustment is often without any
98 gaps outwith natural pauses in ongoing speech. The memory limitation for comparing
99 ongoing stimuli, as experienced in the clinic, has been previously modelled as modest
100 exponential decay over many seconds, albeit for pure-tone stimuli (Durlach and
101 Braida, 1969; Massaro, 1970). Despite qualitative recommendations and a long
102 history of auditory memory research (cf. Cowan, 1984), the effect of duration on
103 preferences for speech stimuli, as presented in the clinic during hearing-aid
104 adjustments, is not known.

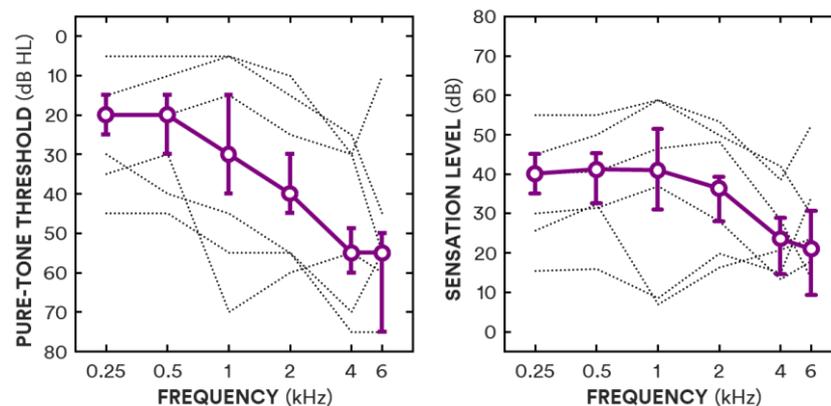
105 On the basis of the foregoing evidence, we hypothesize that extending the
106 duration of the stimulus will elicit more frequent and reliable preferences for gain
107 adjustments. The current study used most of the same methods, including most of
108 the same participants, as Caswell-Midwinter and Whitmer (2020) did when
109 measuring preferences for gain adjustments. The main difference is the primary
110 experimental contrast: stimulus duration. To avoid potential memory confounds, the
111 maximum stimulus duration was 6 s (cf. ITU-R 2003); the minimum was 2 s (vs.
112 0.855-2.3 s in the previous study). To better mimic elements of a clinical session,
113 there were five other methodological differences. First, the stimuli were consecutive
114 segments from a continuous story instead of repeated (within a trial) sentences.
115 Second, the gain adjustment was always the second interval on each trial, not
116 randomised. Third, the number of gain adjustments was reduced from six (± 4 , 8 &
117 12 dB) to four (± 6 & 12 dB). Fourth, there was no ISI. Finally, given the lack of
118 agreement or reliability in using descriptors (e.g., “tinny”) to describe the effect of a
119 gain adjustment in Caswell-Midwinter and Whitmer (2020), the current study only
120 measured preferences.

121 Methods

122 Participants

123 Twenty-nine adults (14 female) were recruited from a sample who had
124 previously participated in a gain discrimination experiment (Caswell-Midwinter and
125 Whitmer, 2019). The median age was 68 years (range 51-74 years). The median
126 better-ear four-frequency (0.5, 1, 2 & 4 kHz) pure-tone threshold average (BE4FA)
127 was 35 dB HL (range 12-56 dB HL). The median sensation level for amplified
128 stimuli, averaged across the same four frequencies, was 35 dB SL (range 15-51 dB
129 SL). None of the participants had a conductive loss (i.e., all participants' average air-
130 bone threshold differences were less than 20 dB; British Academy of Audiology,
131 2016).

132 For 19 participants habitually wearing hearing aids at the time of the study,
133 the real-ear insertion gain provided by their hearing aids in their better ear was
134 measured and used as their gain prescription. For ten participants who were not
135 currently wearing hearing aids, linear NAL-R gain prescriptions (Byrne and Dillon,
136 1986) for their better ear were used. Median hearing-aid experience was 10 years
137 (range 2-35 years). Twenty-six of the 29 participants took part 18 mos. earlier in the
138 aforementioned preference experiment with short sentences (Caswell-Midwinter and
139 Whitmer, 2020).



140

141 Figure 1. Left panel shows median pure-tone thresholds as a function of frequency (circles, solid line)
142 and interquartile ranges (error bars), with the individual curves for the three lowest and highest
143 average thresholds (dotted lines). Right panel shows median sensation level as a function of frequency
144 (circles, solid line) and interquartile ranges (error bars), with the individual curves for the three lowest
145 and highest average sensation levels (dotted lines).

146 All participants had also performed visual letter and digit monitoring tasks
147 during a previous study (min. 18 mos. prior to current study) as an estimate of their
148 cognitive abilities (specifically working memory; Gatehouse et al., 2006). The tasks
149 involved identifying sequences at two different ISIs (1 and 2 s); a full description is in
150 Caswell-Midwinter and Whitmer (2019b). The resulting d' measures were averaged
151 across letter and digit tasks and ISIs to a single cognitive score.

152 Stimuli

153 The stimuli were consecutive segments of a Sherlock Holmes story read by a
154 professional male actor with a Southern English accent (“The Naval Treaty”; Doyle,
155 2011). The original stimuli were collapsed from stereo to mono and resampled to 24
156 kHz from an original recording sample rate of 44.1 kHz. Any silent gaps greater than
157 250 ms were truncated to 250 ms. On each trial, two consecutive segments were
158 presented to the participants’ better ear, both with equal duration of either 2, 4 or 6
159 s. For each segment, 50-ms linear onset and offset gates were applied. To better
160 mimic adjustments in the clinic, the standard stimulus was always the first stimulus
161 in the pair, and there was no ISI beyond the offset and onset gating.

162 For the standard stimulus, real-ear or prescribed gain was applied across six
163 frequency bands: a 0.25 kHz low-pass band, four octave bands centred at 0.5, 1, 2
164 and 4 kHz, and a 6 kHz high-pass band. For the target stimulus, additional gain
165 (Δ Gain) of either -12, -6, 0, +6 and +12 dB was applied in one of three broad
166 frequency bands: a low-frequency band combining 0.25 (low-pass) and 0.5 kHz
167 (octave) bands (LF), a mid-frequency band combining 1 and 2 kHz octave bands
168 (MF), and a high-frequency band combining the 4 kHz and 6 kHz (high-pass) bands
169 (HF). Stimuli were generated by convolving each segment with a 140-tap finite
170 impulse response filter optimised for NAL-R equalisation at 24-kHz sample rate by
171 Kates and Arehart (2010). The overall long-term A-weighted presentation level was
172 60 dB SPL to approximate in-quiet conversation level (Olsen, 1998). Presentation
173 level was verified with an artificial ear and sound level meter (Bruel & Kjaer 4152
174 and 2260), prior to any prescription or gain adjustment. Audibility of the segments
175 was confirmed with each participant after the first trial.

176 We additionally analysed the effect of the natural variation in power across
177 the consecutive segments of each trial (i.e., when Δ Gain = 0). There were significant
178 mean level differences between the two segments in any given trial as a function of
179 both frequency band and segment duration [$F(2,56) = 13.06$ & 19.41 , respectively].
180 The differences, however, were small; variation in band-specific level increased from
181 0.2 dB for the LF band to 0.3 dB for MF and HF bands [$t(28) = 4.76$; $p \ll 0.001$],
182 and variation decreased from 0.3 to 0.2 to 0.1 dB when duration increased from 2 to
183 4 to 6 s, respectively [$t(28) = -2.58$ & -4.39 ; $p = 0.015$ & 0.0002 , respectively].

184 Procedure

185 Participants were seated in a sound-isolated booth (IAC Acoustics), and
186 listened to the stimuli through circumaural headphones (AKG K702) without hearing
187 aids. The change in stimulus within each trial from first to second segment was
188 synchronously indicated on a touch screen in front of the participant. Participants
189 were asked on each trial to listen to each presentation and decide “How did the
190 second sound compare to the first sound?” by selecting either the “better”, “worse” or
191 “no difference” button on the touch screen.

192 There were three segment durations (2, 4 and 6 s) and 13 gain adjustments
193 (± 6 and ± 12 dB adjustments in the LF, MF and HF bands plus a no-adjustment

194 control), resulting in 39 stimulus conditions. Each stimulus condition was repeated
 195 ten times, resulting in 390 trials (3×13×10). The order of presentation was
 196 randomised for each participant. The trial run was broken into equal blocks of 130
 197 trials with breaks between. Prior to testing, each participant completed 12 practice
 198 trials consisting of one trial each of 2-s and 6-s segments with ±12 dB gain
 199 adjustments in each of the three bands.

200 Ethical approval for the study was given by the West of Scotland research
 201 ethics committee (18/WS/0007) and NHS Scotland R&D (GN18EN094). All
 202 participants provided written informed consent prior to testing.

203 Results

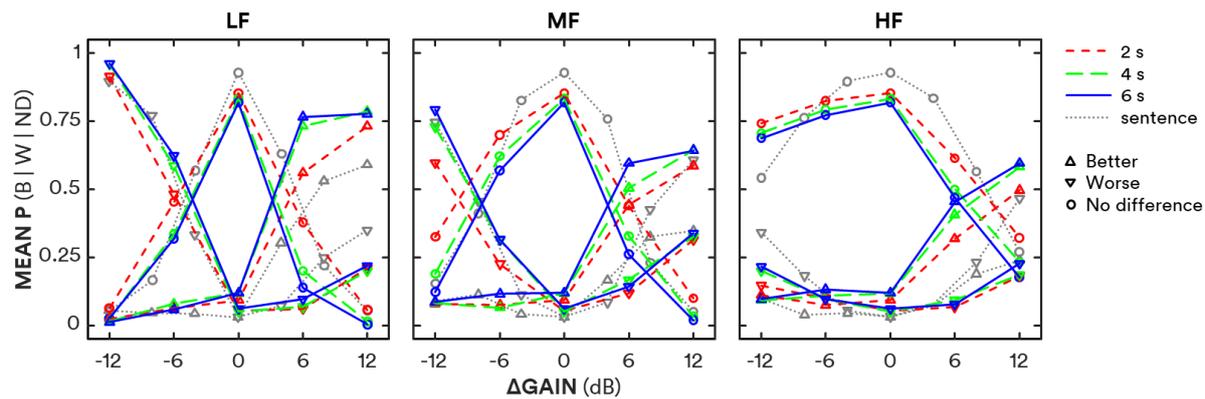
204 Preferences

205 Mean preference ratings – rates of “better,” “worse” and “no difference”
 206 judgments – were calculated for each participant for gain adjustments in each
 207 frequency band (see Figure 2). A repeated-measures analysis of variance was run on
 208 the entire dataset (5 gain adjustments × 3 frequency bands × 3 segment durations)
 209 using individual mean combined “better” and “worse” preference rates [P(B|W) = 1 –
 210 P(ND)] as the dependent variable (see Table 1). Amount of gain adjustment,
 211 frequency band and duration all showed significant main effects on better-and-worse
 212 preferences. Better and worse judgments increased with increasing duration, from 2
 213 to 4 s [$t_{(28)} = 8.44$; $p \ll 0.001$] and 4 to 6 s [$t_{(28)} = 2.80$; $p = 0.0092$].

214 Table 1. Results of the repeated-measures analysis of variance on mean preferences, showing degrees
 215 of freedom (*df*), *F*-statistics and *p* values and partial eta-squared effect sizes. Degrees of freedom (*df*)
 216 and probabilities (*p*) reflect Greenhouse-Geisser (1959) corrections for non-sphericity.

Main effects	<i>df</i> (effect, error)	<i>F</i>	<i>p</i>	η^2
<i>Band</i>	1.46, 40.81	128.30	$\ll 0.001$	0.82
<i>Gain</i>	2.92, 81.72	376.12	$\ll 0.001$	0.93
<i>Duration</i>	1.86, 52.05	55.20	$\ll 0.001$	0.66
Interactions				
<i>Band · Gain</i>	5.10, 142.68	43.24	$\ll 0.001$	0.61
<i>Band · Duration</i>	3.76, 105.30	2.14	0.085	0.07
<i>Gain · Duration</i>	5.64, 157.88	4.87	0.0002	0.15
<i>Band · Gain · Duration</i>	8.05, 225.30	2.28	0.023	0.08

217 The greatest rates of “better” and “worse” responses were for LF adjustments.
 218 Compared to preferences elicited for short sentences in Caswell-Midwinter and
 219 Whitmer (2020; grey triangles and dotted lines in Figure 2), the consecutive
 220 segments elicited more “better” and less “worse” ratings for +12-dB adjustments in
 221 the MF band [$t_{(59)} = 3.11$ & -3.10 for better and worse, respectively; $p = 0.0028$ &
 222 0.0030] and HF band [$t_{(59)} = 5.32$ & -3.77 , respectively; both $p < 0.001$]. There also
 223 appear to be more “better” and less “worse” ratings in the LF band for +12 dB
 224 adjustments (comparing grey with coloured triangles in the left panel of Figure 2) in
 225 the current study compared to the previous, but these differences were not
 226 statistically significant [$t_{(59)} = 1.99$ & -1.60 ; both $p > 0.05$].



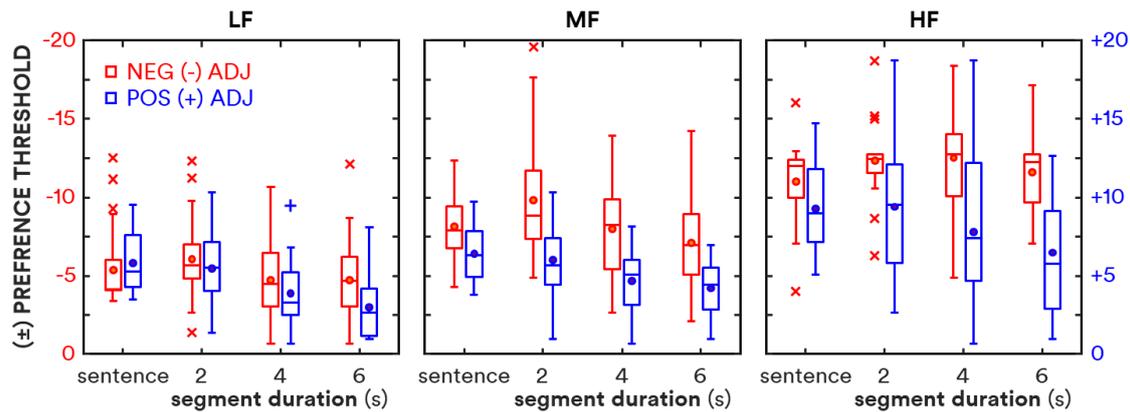
227

228 Figure 2. Mean preferences as a function of gain adjustment for low-frequency (LF; ≤ 0.5 kHz), mid-
229 frequency (MF; 1-2 kHz) and high-frequency (HF; ≥ 4 kHz) bands (left, middle and right panels,
230 respectively) with 2-s, 4-s and 6-s durations (red short-dashed, green long-dashed and blue solid lines,
231 respectively). Better, worse and no difference preferences are shown as upward triangles, downward
232 triangles and circles, respectively. Grey dotted lines and symbols show results using short sentences
233 from Caswell-Midwinter and Whitmer (2020).

234 Participants were less prone to choose “no difference” when there was no gain
235 adjustment in the current study compared to the previous study. The proportion of
236 no difference responses at $\Delta\text{Gain} = 0$ was 0.84 across segment durations compared to
237 0.94 previously for short sentences [$t(56) = 3.31$; $p = 0.0017$].

238 Preference thresholds

239 The minimum gain adjustment required to elicit either a “better” or “worse”
240 preference – the preference threshold – was estimated by fitting each individual’s
241 mean better and worse preferences ($= 1 - \text{no difference preferences}$) with a logistic
242 function. Separate functions were fit to negative and positive gain adjustments (i.e.,
243 decrements and increments) in each frequency band. The threshold was defined as
244 $P(B|W) = 0.55$ [$P(ND) = 0.45$] which corresponds to $d' = 1$ for an unbiased
245 differencing observer in a same-different discrimination task (Macmillan and
246 Creelman, 2005). Shapiro-Wilk tests of normality were violated in three of the 18
247 conditions: 4-s and 6-s LF (+) increment and 2-s MF (-) decrement thresholds ($W =$
248 $0.91, 0.87 \ \& \ 0.88$, respectively; $p = 0.018, 0.0034 \ \& \ 0.0064$); nevertheless, we use
249 Tukey boxplots (Tukey, 1977) in Figure 3 to show the range of preference thresholds
250 for each condition. The Holm-Bonferroni method (Holm, 1979) was used to adjust
251 the rejection probabilities for multiple comparisons where necessary.



252

253 Figure 3. Preference just-noticeable differences (JNDs) as a function of stimulus duration: sentences
 254 (average duration 1.6 s; Caswell-Midwinter & Whitmer, 2020), 2 s, 4 s and 6 s consecutive segments of
 255 a story. Preference thresholds for negative and positive gain adjustments are shown in red and blue,
 256 respectively. Dots show means; lines show medians; boxes show interquartile ranges (IQR); whiskers
 257 show extent of data < 1.5·IQR; crosses and pluses show outliers for negative and positive adjustments,
 258 respectively.

259 A repeated-measures analysis of variance showed main effects of frequency
 260 band, direction (\pm) of gain adjustment and segment duration (see Table 2).
 261 Preference thresholds decreased with segment duration, increased with frequency
 262 band and were greater for decrements than increments. There was a significant
 263 interaction as frequency band \times gain direction; decrement thresholds increased more
 264 than increments with increasing (centre frequency) band. There were also a
 265 significant albeit modest ($\eta^2 = 0.11$) interaction between gain direction and duration;
 266 preference thresholds decreased generally more for increments than decrements.
 267 There was additionally a significant but modest three-way interaction in the MF
 268 band: preference thresholds decreased with increasing segment duration more for
 269 decrements than increments.

270 Table 2. Results of the repeated-measures analysis of variance on preference JNDs, showing degrees of
 271 freedom (df), F -statistics and p values and partial eta-squared effect sizes. Degrees of freedom (df) and
 272 probabilities (p) reflect Greenhouse-Geisser (1959) corrections for non-sphericity.

Main effects	df (effect, error)	F	p	η^2
<i>Band</i>	1.65, 46.34	139.05	$\ll 0.001$	0.83
<i>Direction</i>	1, 28	70.80	$\ll 0.001$	0.72
<i>Duration</i>	1.91, 53.38	48.43	$\ll 0.001$	0.63
Interactions				
<i>Band · Direction</i>	1.69, 47.33	11.54	$\ll 0.001$	0.29
<i>Band · Duration</i>	2.94, 82.27	1.24	0.30	0.04
<i>Direction · Duration</i>	1.66, 46.49	3.35	0.042	0.11
<i>Band · Direction · Duration</i>	3.52, 98.69	3.76	0.0066	0.12

273 Mean thresholds with 95% repeated-measures confidence intervals (Loftus and
 274 Masson, 1994) are shown in Table 3. Thresholds significantly decreased with
 275 increasing duration for gain increments in the LF, MF and HF frequency bands, and
 276 for gain decrements in the for LF and MF bands, respectively; the thresholds for
 277 decrements in the HF band (12.1 dB) did not significantly change across durations.

278 The overall rate of change, derived from a linearisation of mean thresholds not
 279 including HF decrements, decreased as a function of duration from -0.7 to -0.3 dB/s.
 280 That is, preference thresholds decreased more for duration changing from 2 to 4 s
 281 than from 4 to 6 s.

282 Table 3. Mean preference thresholds and 95% confidence intervals in brackets for all conditions
 283 including mean data from Caswell-Midwinter and Whitmer (2020).

	LF		MF		HF	
	-	+	-	+	-	+
sentence	5.3 [4.6-6.1]	5.8 [5.1-6.6]	8.1 [7.4-8.9]	6.4 [5.6-7.1]	11.0 [10.2-11.8]	9.3 [8.5-10.0]
2 s	6.0 [5.3-6.8]	5.5 [4.7-6.2]	9.8 [9.0-10.6]	6.0 [5.2-6.8]	12.3 [11.6-13.1]	9.4 [8.6-10.1]
4 s	4.7 [4.0-5.5]	3.9 [3.1-4.6]	8.0 [7.2-8.7]	4.7 [3.9-5.4]	12.5 [11.7-13.3]	7.8 [7.0-8.5]
6 s	4.7 [3.9-5.5]	3.0 [2.2-3.7]	7.1 [6.3-7.9]	4.2 [3.4-5.0]	11.6 [10.8-12.3]	6.5 [5.7-7.2]

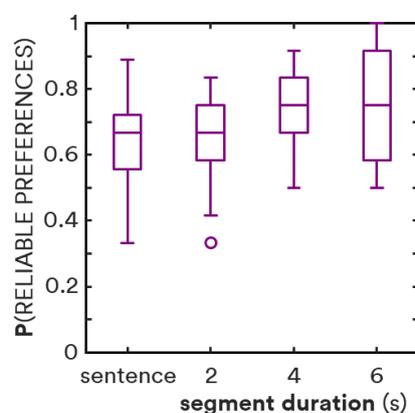
284

285 The preference thresholds here for 2-s consecutive segments of a continuous
 286 story were similar to the thresholds for short sentences in Caswell-Midwinter and
 287 Whitmer (2020) with the exception of MF and HF decrements, for which the current
 288 thresholds were significantly greater ($t = 2.75$ and 2.49 ; $p = 0.011$ and 0.030 ,
 289 respectively). Thresholds for 2-s stimuli were positively correlated across frequency
 290 bands with thresholds in the previous study for both increments and decrements (ρ
 291 $= 0.55$ and 0.72 , respectively; both $p \ll 0.001$). Preference thresholds were not
 292 correlated with age, better-ear four frequency pure-tone average, or hearing-aid
 293 experience after applying Holm-Bonferroni (1979) corrections for multiple
 294 comparisons (all $p > 0.05$). HF increment preference thresholds were positively
 295 correlated with HF pure-tone thresholds ($\rho = 0.44$; $p = 0.032$), and negatively
 296 correlated with HF sensation level ($\rho = -0.48$; $p = 0.019$). Preference thresholds were
 297 not correlated with cognitive score, but the individual decrease in threshold with
 298 duration, characterised as the dB/s slope, was negatively correlated with cognitive
 299 score ($r = -0.50$; $p = 0.0073$). That is, duration had a greater effect on those with
 300 greater letter/digit-monitoring ability.

301 Preference agreement and reliability

302 Fleiss' κ (Fleiss, 1971) was used to measure inter-participant agreement,
 303 comparing participants' most frequent judgment of each adjustment condition. To
 304 simplify the analysis, judgments were collapsed across adjustments for each direction
 305 and frequency band; the $\Delta\text{Gain} = 0$ condition was not included in the analysis.
 306 Fleiss' κ was 0.39 [0.36-0.42 95% confidence intervals (CI)], 0.50 (0.47-0.53) and 0.50
 307 (0.47-0.53) for segments of 2-s, 4-s and 6-s duration, respectively, representing fair (2
 308 s) and moderate (4 & 6 s) agreement (ibid.). That is, agreement significantly
 309 increased from 2-4 s, but not from 4-6 s.

310 For each participant, a given gain adjustment was considered reliable if it
311 elicited seven or more “better,” “worse” or “no difference” judgments, a reliability
312 threshold based on binomial probability theory (Kuk and Lau, 1995). The $\Delta\text{Gain} =$
313 0 condition was not included. Because the proportions of reliable preferences in the
314 current study were not normally distributed based on Shapiro-Wilk tests ($W = 0.92,$
315 0.90 & 0.92 for 2-s, 4-s & 6-s stimuli), non-parametric tests were used. Figure 4
316 shows individual proportions of adjustments with reliable preferences. Reliability
317 increased significantly from a median value of 67% for short sentences and 2-s
318 segments to 75% for 4-s and 6-s segments [$\chi^2 = 11.10; p = 0.011$]. There was no
319 significant difference in reliability between sentences and 2-s segments ($z = 0.65; p =$
320 0.51) nor 4-s and 6-s segments ($z = 0.72; p = 0.47$). The percentage of participants
321 with $\geq 90\%$ reliable preferences, however, did increase from 14% at 4 s to 28% at 6 s.
322 Individual reliabilities for short sentences and 2-s stimuli were not correlated, but
323 reliabilities for 4-s and 6-s stimuli were ($r = 0.61; p = 0.0004$).



324 Figure 4. Proportion of reliable preferences as a function of stimulus duration. Horizontal lines show
325 medians; boxes show interquartile ranges (IQR); whiskers show extent of data $\leq 1.5 \cdot \text{IQR}$; circles
326 show outliers. Sentence data is from Caswell-Midwinter and Whitmer (2020).
327

328 Discussion

329 By having participants compare and judge consecutive segments of a single-
330 narrator story, we have shown that longer durations promote more frequent and
331 reliable preference judgments for gain adjustments in broad frequency bands. That is,
332 the gain adjustments required to elicit preferences decreased with increasing stimulus
333 duration. Preferences were more frequent, ergo preference thresholds were smaller, for
334 increments compared to decrements, in agreement with Caswell-Midwinter and
335 Whitmer (2020) as well as previous psychophysical literature (Ellermeier 1996; Moore
336 et al. 1989). Preferences were less frequent with increasing centre frequency of the
337 adjustment band, as previously shown for short sentences (Caswell-Midwinter and
338 Whitmer, 2020).

339 Despite differences in the method, the median preference thresholds in the
340 current study for 2-s segments were similar to the thresholds for 1.6-s average
341 duration sentences in our previous study (Caswell-Midwinter and Whitmer, 2020),
342 and correlated with the previous thresholds. As with the previous study, the

343 strongest preferences were for increased LF gain and against decreased LF gain, as
344 found in self-fitting studies (Keidser and Convery, 2018; Nelson et al., 2018; Vaisberg
345 et al., 2021). The spectral peaks of the stimuli being in the LF band may have
346 influenced discriminability of LF adjustments (Jesteadt et al. 2017), increasing
347 preferences and reliability. There were preference differences between the two studies,
348 with increases in “better” vs. “worse” judgments for MF and HF increments in the
349 current study. The differences in the long-term spectra between the current
350 monologue and previous sentences – 0.9, 0.2 and -5.6 dB in the LF, MF and HF
351 bands, respectively – can explain the increase in “better” preferences for the HF
352 band, but not the MF band.

353 Participants were less likely to respond “no difference” in the current study
354 where consecutive segments were presented without gain adjustments compared to
355 the previous study (Caswell-Midwinter and Whitmer, 2020) where the same sentence
356 was presented. This difference can be attributed to the comparison of two different
357 speech segments; the naturally occurring differences in the spectrotemporal patterns
358 across the two intervals (without any gain difference) decreases the likelihood of a
359 “no difference” response. The effect of this decrease in no-difference responses on
360 threshold estimation was minimal, decreasing threshold estimates by 0.4 dB on
361 average when comparing with individual no-difference responses for the 26
362 participants from the previous study. Nevertheless, the change demonstrates a
363 limitation of using sequential stimuli for comparison.

364 The use of an ongoing story (cf. hearing the same utterance twice) provided a
365 greater degree of participant engagement with the material, as might occur in the
366 clinic, where the patient is selecting parameters for real-world use. Conversely, any
367 greater engagement with the stimulus content, however, may have been detrimental
368 to performing the task. Beyond the decrease in no-difference responses, the effect of
369 comparing different stimuli (two consecutive segments) versus comparing identical
370 stimuli was otherwise small. Using non-repeating intervals may have introduced
371 decision noise, inflating thresholds (cf. Whitmer and Akeroyd, 2011). The natural
372 variations in spectrum between the consecutive segments on any given trial were
373 modest on average, and excluding trials with the greatest inter-stimulus variation in
374 any frequency band only affected particular thresholds, and increased – not decreased
375 – those thresholds modestly (0.2-0.3 dB). That is, there is scant evidence that the
376 natural variation in the consecutive intervals affected the pattern of results.

377 The delivery of stimuli for appraisal by the patient in the clinic may, however,
378 be different to paired or sequential comparisons. Instead of a pre- and post-
379 adjustment comparison, the appraisal may take the form of a single interval. Single
380 interval ratings of hearing-aid sound quality have shown moderate test-retest
381 reliability (Narendran and Humes, 2003) and good inter-rater reliability (Gabrielsson
382 et al. 1990), but these studies were with stimuli durations of 50-60 s. Using such long
383 stimuli within clinical fine-tuning may not be feasible.

384 It is not clear from the current results if talking even longer (i.e., for durations
385 > 6 s) would provide even greater discriminability and more reliable preferences.
386 While the thresholds across most conditions decreased significantly from 4-s to 6-s,
387 the trend was asymptotic. The overall rate of change decreased from -0.8 dB/s at 4 s
388 to -0.4 dB/s at 6 s, resembling the modest exponential decay of memory-performance
389 models (e.g., Durlach and Braida, 1969). In line with memory-performance models,
390 there was a negative correlation between participants' monitoring-task cognitive
391 scores and the rate of decrease in their preference thresholds with increasing
392 duration. That is, the better their cognitive scores, the stronger the effect of stimulus
393 duration on preference thresholds. This suggests that the effect of duration in the
394 judgment of gain adjustments is limited by each individual's cognitive abilities. The
395 mean preferences were very similar for 4-s and 6-s stimuli (Figure 2), and there was
396 no increase in inter-participant agreement nor intra-participant reliability (Figure 4).
397 It is therefore unlikely for thresholds to decrease, or reliability to increase, much
398 further beyond the results here for 6-s stimuli (cf. Sams et al., 1993).

399 The improvement in thresholds and reliability with stimulus duration is also
400 small relative to the thresholds and reliabilities themselves. Talking or presenting
401 stimuli for 6 s to a hearing-aid wearer in the clinic will help elicit preferences for
402 adjustments, but those adjustments still need to be large: 3-6 dB for increments, 5-12
403 dB for decrements. These thresholds are still well above common troubleshooting
404 adjustments, especially for adjustments in the higher frequencies. In the
405 personalisation of hearing aids in the clinic, it is therefore important to not only say
406 more than a few words (e.g., "how's that sound?") immediately following an
407 adjustment, but to ensure the adjustment is large enough to elicit reliable feedback.
408 Given these constraints, alternative methods of fitting, such as self-adjustments
409 (Boothroyd and Mackersie, 2017; Nelson et al., 2018), may be more viable for
410 effective hearing-aid personalisation.

411 **Acknowledgments**

412 The authors would like to thank David McShefferty for his assistance in conducting
413 the study. This work was supported by funding from the Medical Research Council
414 [grant numbers MR/S003576/1 and 1601056]; and the Chief Scientist Office of the
415 Scottish Government.

416 **Disclosure Statement**

417 No potential conflict of interest was reported by the authors.

418 **ORCID**

419 William M. Whitmer: <https://orcid.org/0000-0001-8618-6851>

420 Benjamin Caswell-Midwinter: <https://orcid.org/0000-0002-3386-3860>

421 Graham Naylor: <https://orcid.org/0000-0003-1544-1944>

422 Funding

423 This work was supported by funding from the Medical Research Council [grant
424 numbers MR/S003576/1 and 1601056]; and the Chief Scientist Office of the Scottish
425 Government.

426 References

- 427 Anderson MC, Arehart KH, Souza PE (2018) Survey of current practice in the fitting
428 and fine-tuning of common signal-processing features in hearing aids for
429 adults. *J Am Acad Audiol* 29: 118-124. DOI: 10.3766/jaaa.16107.
- 430 Caswell-Midwinter B, Whitmer WM (2019) Discrimination of gain increments in
431 speech. *Trends Hear* 23. DOI: 10.1177/2331216519886684.
- 432 Caswell-Midwinter B, Whitmer WM (2019b) Discrimination of gain increments in
433 speech-shaped noises. *Trends Hear* 23. DOI: 10.1177/2331216518820220.
- 434 Caswell-Midwinter B, Whitmer WM (2020) The perceptual limitations of
435 troubleshooting hearing aids based on patients' descriptions. *Int J Audiol*.
436 DOI: 10.1080/14992027.2020.1839679
- 437 Doyle AC (2011) *The Memoirs of Sherlock Holmes* (D. Jacobi, narr.) [Audiobook].
438 London: AudioGO Ltd.
- 439 Dai H, Green DM (1993) Discrimination of spectral shape as a function of stimulus
440 duration. *J Acoust Soc Am* 93(2): 957-965. DOI: 10.1121/1.405456.
- 441 Durlach NI, Braida LD (1969) Intensity perception. I. Preliminary theory of intensity
442 resolution. *J Acoust Soc Am* 46(2): 373-383. DOI: 10.1121/1.1911699.
- 443 Ellermeier W (1996) Detectability of increments and decrements in spectral profiles.
444 *J Acoust Soc Am* 99(5): 3119-3125. DOI: 10.1121/1.414797.
- 445 Farrar CL, Reed CM, Ito Y, Durlach NI, Delhorne LA, Zurek PM, Braida LM (1987)
446 Spectral-shape discrimination I: Results from normal-hearing listeners for
447 stationary broadband noises. *J Acoust Soc Am* 81(4): 1085-1092. DOI:
448 10.1121/1.394628.
- 449 Florentine M (1986) Level discrimination of tones as a function of duration. *J Acoust*
450 *Soc Am* 79(3): 792-798. DOI: 10.1121/1.393469.
- 451 Gatehouse S, Naylor G, Elberling C (2006) Linear and nonlinear hearing aid fittings
452 – 2. Patterns of candidature. *Int J Audiol* 45(3): 153-171. DOI:
453 10.1080/14992020500429484.
- 454 Green DM, Mason CR, Kidd G (1984) Profile analysis: Critical bands and duration.
455 *J Acoust Soc Am* 75(4): 1163-1167. DOI: 10.1121/1.390765.
- 456 Greenhouse SW, Geisser S (1959) On methods in the analysis of profile data.
457 *Psychometrika* 24: 95-112. DOI: 10.1007/BF02289823.
- 458 Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat*
459 6: 65-70.
- 460 International Telecommunication Union, Radiocommunication Sector (2019) General
461 methods for the subjective assessment of sound quality. Recommendation
462 ITU-R BS.1284-2
- 463 International Telecommunication Union, Telecommunication Standardization Sector
464 (2003) Subjective test methodology for evaluating speech communication

- 465 systems that include noise suppression algorithm. Recommendation ITU-T
466 P.835.
- 467 Isarangura S, Eddins AC, Ozmeral EJ, Eddins DA (2019) The effects of duration and
468 level on spectral modulation perception. *J Speech Lang Hear Res*, 62: 3876-
469 3886.
- 470 Jenstad LM, Van Tasell DJ, and Ewert C (2003) Hearing Aid Troubleshooting Based
471 on Patients' Descriptions. *J Am Acad Audiol*, 14 (7): 347–360.
- 472 Jesteadt W, Walker SM, Oluwaseye AO, Ohlrich B, Brunette KE, Wróblewski M,
473 Schmid KK (2017) Relative contributions of specific frequency bands to the
474 loudness of broadband sounds. *J Acoust Soc Am*, 142(3): 1597-1610. DOI:
475 10.1121/1.5003778.
- 476 Keidser G, Convery E (2018). Outcomes with a self-fitting hearing aid. *Trends Hear*,
477 22: 1–12. DOI: 10.1177/2331216518768958.
- 478 Kuk FK, Lau C (1995) The application of binomial probability theory to paired
479 comparison responses. *Am J Audiol*, 4 (1): 37–42. DOI:10.1044/1059-
480 0889.0401.37
- 481 Kuk FK, Ludvigsen C (1999) Variables affecting the use of prescriptive formulae to
482 fit modern nonlinear hearing aids. *J Am Acad Audiol*, 10: 453-465.
- 483 Loftus GR, Masson MEJ (1994) Using confidence intervals in within-subject designs.
484 *Psychon Bull Rev*, 1: 476-490. DOI: 10.3758/BF03210951
- 485 Macmillan NA, Creelman CD (2005) *Detection theory a user's guide*. Mahwah, NJ:
486 Lawrence Erlbaum Associates.
- 487 Mackersie CL, Boothroyd A, Lithgow A. (2019) A "Goldilocks" approach to hearing
488 aid self-fitting: ear-canal output and speech intelligibility index. *Ear Hearing*
489 40(1): 107-115. DOI: 10.1097/AUD.0000000000000617.
- 490 Moore BCJ, Oldfield SR, Dooley GJ (1989) Detection and discrimination of spectral
491 peaks and notches at 1 and 8 kHz. *J Acoust Soc Am* 85: 820-836. DOI:
492 10.1121/1.397554.
- 493 Moore BCJ, Peters RW, Kohlrausch A, van de Par S (1997) Detection of increments
494 and decrements in sinusoids as a function of frequency, increment, and
495 decrement duration and increment duration. *J Acoust Soc Am* 102(5): 2954-
496 2965. DOI: 10.1121/1.420350.
- 497 Narendran MM, Humes LE (2003) Reliability and validity of judgments of sound
498 quality in elderly hearing aid wearers. *Ear Hear* 24(1): 4-11. DOI:
499 10.1097/01.AUD.0000051745.69182.14.
- 500 Nelson PB, Perry TT, Gregan M, Van Tasell D (2018) Self-adjusted amplification
501 parameters produce large between-subject variability and preserve speech
502 intelligibility. *Trends Hear*, 22. DOI: 10.1177/2331216518798264.
- 503 Oxenham AJ, Buus S (2000) Level discrimination of sinusoids as a function of
504 duration and level for fixed-level, roving-level, and across-frequency
505 conditions. *J Acoust Soc Am* 107(3): 1605-1614. DOI: 10.1121/1.428445.
- 506 Pollack I (1972) Memory for auditory waveform. *J Acoust Soc Am* 52(4): 1209-1215.
507 DOI: 10.1121/1.1913234.

- 508 Sams M, Riitta H, Rif J, Knuutila J (1993) The human auditory sensory memory
509 trace persists about 10 sec: Neuromagnetic evidence. *J Cogn Neurosci* 5(3):
510 363-370. DOI: 10.1162/jocn.1993.5.3.363.
- 511 Thielemans TD, Pans M, Chenault, and Anteunis L (2017) Hearing aid fine-tuning
512 based on Dutch descriptions. *Int J Audiol*, 56 (7): 507–515.
513 DOI:10.1080/14992027.2017.1288302.
- 514 Vaisberg JM, Beaulac S, Glista D, Macpherson EA, Scollie SD (2021) Perceived
515 sound quality dimensions influencing frequency-gain shaping preferences for
516 hearing aid-amplified speech and music. *Trends Hear*, 25. DOI:
517 10.1177/2331216521989900.
- 518 Valente DL, Patra H, Jesteadt W (2011) Relative effects of increment and pedestal
519 duration on the detection of intensity increments. *J Acoust Soc Am* 129(4):
520 2095-2103. DOI: 10.1121/1.3557043.
- 521 Whitmer WM, & Akeroyd MA (2011) Level discrimination of speech sounds by
522 hearing-impaired individuals with and without hearing amplification. *Ear*
523 *Hear* 32, 391–398. DOI:10.1097/AUD.0b013e318202b620