medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

The effect of stimulus duration on preferences for gain adjustments in speech

Whitmer, William M.[a, b]

Caswell-Midwinter, Benjamin[a, b, c]

Naylor, Graham[a]

a) Hearing Sciences – Scottish Section, School of Medicine, University of Nottingham, Glasgow, UK.

b) Institute of Health and Wellbeing, College of Medical, Veterinary and Life Science, University of Glasgow, Glasgow, UK.

c) Otolaryngology – Head and Neck Surgery, Massachusetts Eye and Ear, Harvard Medical School, Boston, Massachusetts, USA

Corresponding author: bill.whitmer@nottingham.ac.uk

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NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Abstract

Objectives

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

Design

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

Study Sample

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

Results

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

Conclusions

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

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Introduction

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Methods

Participants

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

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

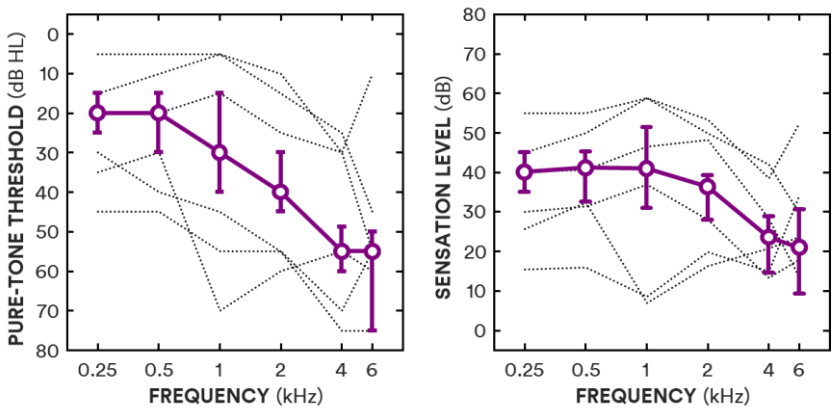

Figure 1. Left panel shows median pure-tone thresholds as a function of frequency (circles, solid line) and interquartile ranges (error bars), with the individual curves for the three lowest and highest average thresholds (dotted lines). Right panel shows median sensation level as a function of frequency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

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

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Stimuli

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

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

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

Procedure

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

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

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

control), resulting in 39 stimulus conditions. Each stimulus condition was repeated ten times, resulting in 390 trials (3×13×10). The order of presentation was randomised for each participant. The trial run was broken into equal blocks of 130 trials with breaks between. Prior to testing, each participant completed12 practice trials consisting of one trial each of 2-s and 6-s segments with ±12 dB gain adjustments in each of the three bands.

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

Results

Preferences

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

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

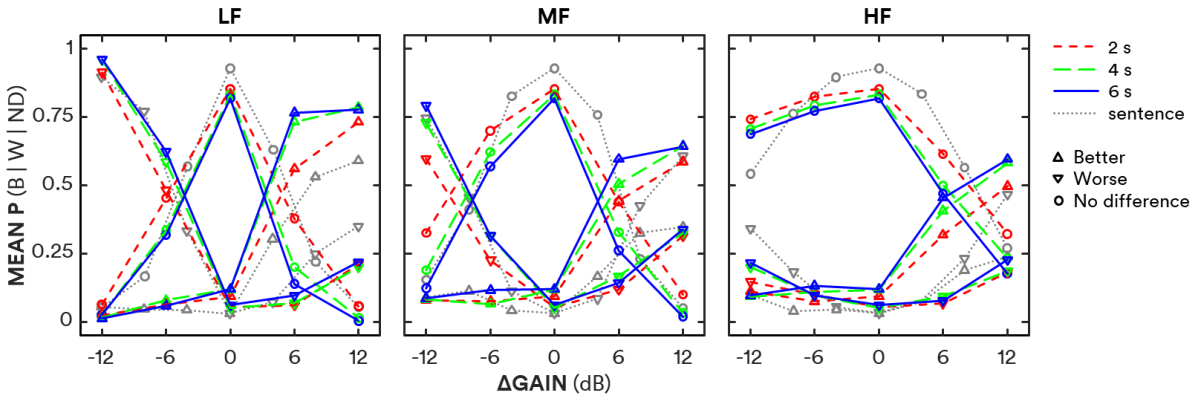

Figure 2. Mean preferences as a function of gain adjustment for low-frequency (LF; ≤ 0.5 kHz), mid-frequency (MF; 1-2 kHz) and high-frequency (HF; ≥ 4 kHz) bands (left, middle and right panels, respectively) with 2-s, 4-s and 6-s durations (red short-dashed, green long-dashed and blue solid lines, respectively).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

Participants were less prone to choose "no difference" when there was no gain adjustment in the current study compared to the previous study. The proportion of no difference responses at ΔGain = 0 was 0.84 across segment durations compared to 0.94 previously for short sentences [t(56) = 3.31; p = 0.0017].

Preference thresholds

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

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

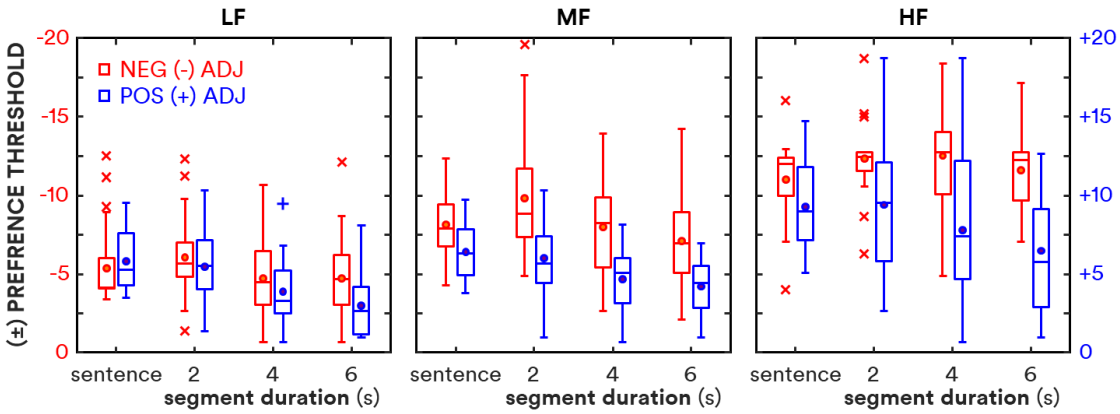

Figure 3. Preference just-noticeable differences (JNDs) as a function of stimulus duration: sentences (average duration 1.6 s; Caswell-Midwinter & Whitmer, 2020), 2 s, 4 s and 6 s consecutive segments of a story. Preference thresholds for negative and positive gain adjustments are shown in red and blue,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

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

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

Table 3. Mean preference thresholds and 95% confidence intervals in brackets for all conditions 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]

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

Preference agreement and reliability

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

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

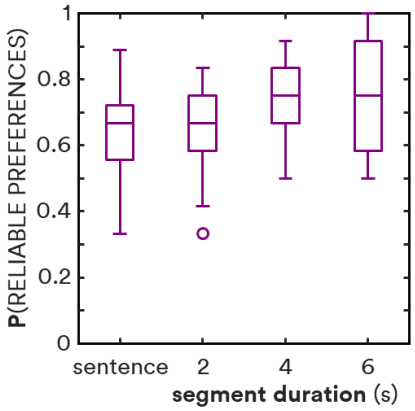

Figure 4. Proportion of reliable preferences as a function of stimulus duration. Horizontal lines show medians; boxes show interquartile ranges (IQR); whiskers show extent of data ⩽ 1.5 • IQR; circles show outliers. Sentence data is from Caswell-Midwinter and Whitmer (2020).

Discussion

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

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

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

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

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

Acknowledgments

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

Disclosure Statement

No potential conflict of interest was reported by the authors.

ORCID

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

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

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

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Funding

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

References

Anderson MC, Arehart KH, Souza PE (2018) Survey of current practice in the fitting and fine-tuning of common signal-processing features in hearing aids for adults. *J Am Acad Audiol* 29: 118-124. DOI: 10.3766/jaaa.16107.

Caswell-Midwinter B, Whitmer WM (2019) Discrimination of gain increments in speech. *Trends Hear* 23. DOI: 10.1177/2331216519886684.

Caswell-Midwinter B, Whitmer WM (2019b) Discrimination of gain increments in speech-shaped noises. *Trends Hear* 23. DOI: 10.1177/2331216518820220.

Caswell-Midwinter B, Whitmer WM (2020) The perceptual limitations of troubleshooting hearing aids based on patients' descriptions. *Int J Audiol.* DOI: 10.1080/14992027.2020.1839679

Doyle AC (2011) *The Memoirs of Sherlock Holmes* (D. Jacobi, narr.) [Audiobook]. London: AudioGO Ltd.

Dai H, Green DM (1993) Discrimination of spectral shape as a function of stimulus duration. *J Acoust Soc Am* 93(2): 957-965. DOI: 10.1121/1.405456.

Durlach NI, Braida LD (1969) Intensity perception. I. Preliminary theory of intensity resolution. *J Acoust Soc Am* 46(2): 373-383. DOI: 10.1121/1.1911699.

Ellermeier W (1996) Detectability of increments and decrements in spectral profiles. *J Acoust Soc Am* 99(5): 3119-3125. DOI: 10.1121/1.414797.

Farrar CL, Reed CM, Ito Y, Durlach NI, Delhorne LA, Zurek PM, Braida LM (1987) Spectral-shape discrimination I: Results from normal-hearing listeners for stationary broadband noises. *J Acoust Soc Am* 81(4): 1085-1092. DOI: 10.1121/1.394628.

Florentine M (1986) Level discrimination of tones as a function of duration. *J Acoust Soc Am* 79(3): 792-798. DOI: 10.1121/1.393469.

Gatehouse S, Naylor G, Elberling C (2006) Linear and nonlinear hearing aid fittings – 2. Patterns of candidature. *Int J Audiol* 45(3): 153-171. DOI: 10.1080/14992020500429484.

Green DM, Mason CR, Kidd G (1984) Profile analysis: Critical bands and duration. *J Acoust Soc Am* 75(4): 1163-1167. DOI: 10.1121/1.390765.

Greenhouse SW, Geisser S (1959) On methods in the analysis of profile data. *Psychometrika* 24: 95-112. DOI: 10.1007/BF02289823.

Holm S (1979) A simple sequentially rejective multiple test procedure. *Scand J Stat* 6: 65-70.

International Telecommunication Union, Radiocommunication Sector (2019) General methods for the subjective assessment of sound quality. Recommendation ITU-R BS.1284-2

International Telecommunication Union, Telecommunication Standardization Sector (2003) Subjective test methodology for evaluating speech communication

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

systems that include noise suppression algorithm. Recommendation ITU-T P.835.

Isarangura S, Eddins AC, Ozmeral EJ, Eddins DA (2019) The effects of duration and level on spectral modulation perception. *J Speech Lang Hear Res*, 62: 3876-3886.

Jenstad LM, Van Tasell DJ, and Ewert C (2003) Hearing Aid Troubleshooting Based on Patients' Descriptions. *J Am Acad Audiol,* 14 (7): 347–360.

Jesteadt W, Walker SM, Oluwaseye AO, Ohlrich B, Brunette KE, Wróblewski M, Schmid KK (2017) Relative contributions of specific frequency bands to the loudness of broadband sounds. *J Acoust Soc Am*, 142(3): 1597-1610. DOI: 10.1121/1.5003778.

Keidser G, Convery E (2018). Outcomes with a self-fitting hearing aid. *Trends Hear*, 22: 1–12. DOI: 10.1177/2331216518768958.

Kuk FK, Lau C (1995) The application of binomial probability theory to paired comparison responses. *Am J Audiol,* 4 (1): 37–42. DOI:10.1044/1059-0889.0401.37

Kuk FK, Ludvigsen C (1999) Variables affecting the use of prescriptive formulae to fit modern nonlinear hearing aids. *J Am Acad Audiol,* 10: 453-465.

Loftus GR, Masson MEJ (1994) Using confidence intervals in within-subject designs. *Psychon Bull Rev*, 1: 476-490. DOI: 10.3758/BF03210951

Macmillan NA, Creelman CD (2005) *Detection theory a user's guide*. Mahwah, NJ: Lawrence Erlbaum Associates.

Mackersie CL, Boothroyd A, Lithgow A. (2019) A "Goldilocks" approach to hearing aid self-fitting: ear-canal output and speech intelligibility index. *Ear Hearing* 40(1): 107-115. DOI: 10.1097/AUD.0000000000000617.

Moore BCJ, Oldfield SR, Dooley GJ (1989) Detection and discrimination of spectral peaks and notches at 1 and 8 kHz. *J Acoust Soc Am* 85: 820-836. DOI: 10.1121/1.397554.

Moore BCJ, Peters RW, Kohlrausch A, van de Par S (1997) Detection of increments and decrements in sinusoids as a function of frequency, increment, and decrement duration and increment duration. *J Acoust Soc Am* 102(5): 2954-2965. DOI: 10.1121/1.420350.

Narendran MM, Humes LE (2003) Reliability and validity of judgments of sound quality in elderly hearing aid wearers. *Ear Hear* 24(1): 4-11. DOI: 10.1097/01.AUD.0000051745.69182.14.

Nelson PB, Perry TT, Gregan M, Van Tasell D (2018) Self-adjusted amplification parameters produce large between-subject variability and preserve speech intelligibility. *Trends Hear*, 22. DOI: 10.1177/2331216518798264.

Oxenham AJ, Buus S (2000) Level discrimination of sinusoids as a function of duration and level for fixed-level, roving-level, and across-frequency conditions. *J Acoust Soc Am* 107(3): 1605-1614. DOI: 10.1121/1.428445.

Pollack I (1972) Memory for auditory waveform. *J Acoust Soc Am* 52(4): 1209-1215. DOI: 10.1121/1.1913234.

medRxiv preprint doi: https://doi.org/10.1101/2021.02.26.21252511; this version posted March 5, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license .

Sams M, Riitta H, Rif J, Knuutila J (1993) The human auditory sensory memory trace persists about 10 sec: Neuromagnetic evidence. *J Cogn Neurosci* 5(3): 363-370. DOI: 10.1162/jocn.1993.5.3.363.

Thielemans TD, Pans M. Chenault, and Anteunis L (2017) Hearing aid fine-tuning based on Dutch descriptions. *Int J Audiol,* 56 (7): 507–515. DOI:10.1080/14992027.2017.1288302.

Vaisberg JM, Beaulac S, Glista D, Macpherson EA, Scollie SD (2021) Perceived sound quality dimensions influencing frequency-gain shaping preferences for hearing aid-amplified speech and music. *Trends Hear*, 25. DOI: 10.1177/2331216521989900.

Valente DL, Patra H, Jesteadt W (2011) Relative effects of increment and pedestal duration on the detection of intensity increments. *J Acoust Soc Am* 129(4): 2095-2103. DOI: 10.1121/1.3557043.

Whitmer WM, & Akeroyd MA (2011) Level discrimination of speech sounds by hearing-impaired individuals with and without hearing amplification. *Ear Hear* 32, 391–398. DOI:10.1097/AUD.0b013e318202b620