2021

Intelligibility benefit for familiar voices does not depend on better discrimination of fundamental frequency or vocal tract length

Emma Holmes

Ingrid Johnsrude
Western University, ijohnsru@uwo.ca

Intelligibility benefit for familiar voices does not depend on better discrimination of fundamental frequency or vocal tract length

Emma Holmes

Department of Speech Hearing and Phonetic Sciences, UCL, London, WC1N 1PF, U.K.

emma.holmes@ucl.ac.uk

Ingrid S. Johnsrude

Brain and Mind Institute, University of Western Ontario, London, Ontario, N6A 5B7, Canada; School of Communication Sciences and Disorders, University of Western Ontario, London, Ontario, N6G 1H1, Canada

ijohnsru@uwo.ca

Abstract

Speech is more intelligible when it is spoken by familiar than unfamiliar people. Two cues to voice identity are glottal pulse rate (GPR) and vocal tract length (VTL): perhaps these features are more accurately represented for familiar voices in a listener's brain. If so, listeners should be able to discriminate smaller manipulations to perceptual correlates of these vocal parameters for familiar than unfamiliar voices. We recruited pairs of friends who had known each other for 0.5–22.5 years. We measured thresholds for discriminating pitch (correlate of GPR) and formant spacing (correlate of VTL; 'VTL-timbre') for voices that were familiar (friends) and unfamiliar (friends of other participants). When a competing talker was present, speech was substantially more intelligible when it was spoken in a familiar voice. Discrimination thresholds were not systematically smaller for familiar compared to unfamiliar talkers. Although, participants detected smaller deviations to VTL-timbre than pitch uniquely for familiar talkers, suggesting a different balance of characteristics contribute to discrimination of familiar and unfamiliar voices. Across participants, we found no relationship between the size of the intelligibility benefit for a familiar over an unfamiliar voice and the difference in discrimination thresholds for the same voices. Also, the intelligibility benefit was not affected by the acoustic manipulations we imposed on voices to assess discrimination thresholds. Overall, these results provide no evidence that two important cues to voice identity—pitch and VTL-timbre—are more accurately represented when voices are familiar, or are necessarily responsible for the large intelligibility benefit derived from familiar voices.

Keywords

Speech; voice; familiar; discrimination; vocal tract length; pitch

Introduction

We naturally become familiar with the voices of people we often interact with (such as friends and family). This allows us to recognize them by voice. In other words, familiar voices tell us about talker identity. Yet, words spoken by familiar people are also much more intelligible than the same words spoken by unfamiliar people when other sounds (e.g., competing speech or noise) are present—demonstrating that familiar voices also help us to retrieve linguistic content. A familiar-voice intelligibility benefit has been documented in a variety of contexts (Barker & Newman, 2004; Domingo, Holmes, & Johnsrude, 2020; Holmes, Domingo, & Johnsrude, 2018; Johnsrude et al., 2013; Kreitewolf, Mathias, & von Kriegstein, 2017; Levi, Winters, & Pisoni, 2011; Newman & Evers, 2007; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994; Souza, Gehani, Wright, & McCloy, 2013; Yonan & Sommers, 2000).

Traditional models of speech perception hold that recognition of words and phrases requires voice information to be stripped away from the acoustic signal to obtain discrete, abstract, linguistic units, which are the basic perceptual unit. Yet, the finding that talker information influences the perception of speech content (e.g., the words that are spoken), challenges this view (Lachs, McMichael, & Pisoni, 2003; see, for example, Nygaard et al., 1994; Pisoni, 1996; Remez, Fellowes, & Rubin, 1997). This finding can be explained under exemplar-based, or 'episodic' accounts of speech perception (Goldinger, 1996, 1998), which recognize that specific detail about particular instances of speech are stored in memory. Under these accounts, more memory traces would be stored for familiar than unfamiliar voices, and these memories should allow participants to better match acoustic properties of familiar voice. This finding can also be explained under the prototype account (Lavner, Rosenhouse, & Gath, 2001), which assumes that incoming speech is compared to a prototype (i.e., 'average' or common) voice, and different voices are represented as the distance from the prototype in acoustic space. Under this account, we may assume that familiar voices contribute more

65 strongly to the prototype representation and, therefore, this prototype should allow acoustic
66 properties to be better recovered for familiar than unfamiliar voices. However, specifically
67 which details of familiar voices are stored and subsequently utilized to benefit speech
68 intelligibility are unclear.

69 According to the source-filter model of speech production (Chiba & Kajiyama, 1941;
70 Fant, 1960) voice acoustics are the product of the vocal source (vocal-fold vibration) filtered
71 through the vocal tract. Vocal-fold vibration rate affects the perceived pitch of a vocalisation.
72 The length of the vocal tract (including the laryngeal cavity, the pharynx and the oral cavity)
73 determines its resonance characteristics, which manifest as the spacing of formants in speech.
74 This is perceived as a specific timbre (hereafter, referred to as VTL-timbre). These two
75 prominent characteristics—pitch and VTL-timbre—determine whether a voice is heard as
76 male or female, adult or child (e.g., Smith & Patterson, 2005), and are important cues to voice
77 identity (Holmes et al., 2018; LaRiviere, 1975; Lavner, Gath, & Rosenhouse, 2000; Lavner et
78 al., 2001; van Dommelen, 1987, 1990). Thus, it seems plausible that listeners could use similar
79 acoustic characteristics to achieve the familiar-voice intelligibility benefit.

80 One possible explanation for the familiar-voice benefit is that people are better at
81 predicting attributes (such as pitch and VTL-timbre) of a familiar voice than an unfamiliar
82 voice—and this may allow them to better understand speech spoken by familiar people when
83 it is masked by other sounds. For example, people might utilise precise predictions for a
84 familiar voice to help them focus their attention on that person's voice. This explanation aligns
85 with several theories of speech recognition: better predictions could be underpinned by more
86 stored 'episodes' for familiar voices (Goldinger, 1996, 1998) or because familiar voices are
87 closer to the prototype representation (Lavner et al., 2001). If either of these explanations
88 are correct, then we would expect that precise representations of familiar voices would allow
89 listeners to discriminate smaller deviations to voice acoustics for familiar than unfamiliar

voices. In other words, acuity should be better for familiar than unfamiliar voices, as demonstrated by smaller just-noticeable differences (JNDs) for discriminating voice attributes (such as pitch and VTL-timbre).

An alternative explanation for the familiar-voice benefit—which would *not* lead to better discrimination thresholds for familiar voices—is that familiarity affects the active cognitive processes engaged in speech perception (Heald & Nusbaum, 2014). For example, familiar voices may be processed more efficiently than unfamiliar voices. Indeed, Holmes and Johnsrude (2020) found that the magnitude of the familiar-voice benefit to intelligibility differed among maskers that were acoustically similar but differed in content: They found a benefit when the masker was a competing talker, but no familiar-voice benefit when the masker was unintelligible modulated noise. These results imply that voice familiarity helps listeners to resist interference from the content of a masker. Crucially, under this account, familiarity would *not* be expected to affect discrimination of speech presented in quiet.

Previous studies rule out several other explanations for the familiar-voice intelligibility benefit. For example, the familiar-voice benefit is not an artefact of listeners being more likely to guess at possible words when a voice is familiar (i.e., a shift in report criterion, or bias), since it is robustly observed using closed-set tests, in which participants report the same number of words on every trial (Domingo et al., 2020; Holmes et al., 2018; Johnsrude et al., 2013; Kreitewolf et al., 2017). Another benefit of closed-set tests is that transitional probabilities between words in sentences are strictly controlled and the materials are identical across familiarity conditions. This rules out any explanation based on listeners being more able to predict upcoming *words* in sentences spoken by familiar people. Familiar voices do not seem to be more attentionally salient than unfamiliar voices: If that were the case, then target speech would be harder to understand when masked by a familiar talker compared to an

unfamiliar talker, and no such pattern has been observed (Domingo et al., 2020; Johnsrude et al., 2013).

In this experiment, we compared perceptual discrimination thresholds for pitch and VTL-timbre for familiar and unfamiliar voices—to tease apart explanations based on better predictions of familiar-voice attributes compared with more cognitively efficient processing of familiar voices. We also tested intelligibility of the same voices in the presence of competing speech, using a closed-set speech corpus. We measured intelligibility of the voices in their original form and when their pitch and VTL-timbre had been manipulated to match the participant's discrimination threshold—to test whether manipulating these voice characteristics reduces the intelligibility benefit gained from familiar voices.

Methods

Participants

We recruited 10 pairs of participants, who had known each other for 0.5–22.5 years (median = 1.7 years, interquartile range = 3.1) and reported that they usually spoke to each other 3–78 hours in person each week (median = 17.0 hours, interquartile range = 29.5). Pairs of participants were friends, roommates, or siblings. Two participants did not complete the experiment and one participant was excluded due to a technical error during data collection. The remaining 17 participants (3 male) were aged 19–29 years (median = 20.8, interquartile range = 1.9) and were Canadian native English speakers with normal hearing (average pure-tone thresholds at octave frequencies between 0.5 and 4 kHz of 10 dB HL or better in each ear).

A power analysis (GPower 3.1; Faul, Erdfelder, Buchner, & Lang, 2009) showed that 17 participants is sufficient to detect within-subjects effects of size $d > 0.58$ with 0.8 power and correlations of size $r > 0.47$. The familiar voice-benefit to intelligibility found by Johnsrude

et al. (2013) was much larger than this (d = 1.44), and effects of this size should be detectable with power ~1.00 with 17 participants.

The experiment was cleared by Western University's Health Sciences Research Ethics Board. Informed consent was obtained from all participants.

Apparatus

The experiment was conducted in a single-walled sound-attenuating booth (Eckel Industries of Canada, Ltd.; Model CL-13 LP MR). Participants sat in a comfortable chair facing a 24-inch LCD visual display unit (either ViewSonic VG2433SMH or Dell G2410t).

Acoustic stimuli were recorded using a Sennheiser e845-S microphone connected to a Steinberg UR22 sound card (Steinberg Media Technologies).

Acoustic stimuli were presented through a Steinberg UR22 sound card (Steinberg Media Technologies) connected to Grado Labs SR225 headphones.

Stimuli

Each participant recorded 480 sentences from the Boston University Gerald (BUG) corpus (Kidd et al., 2008), which are of the form: "<Name> <verb> <number> <adjective> <noun>". An example is "Bob bought three green bags". To ensure that all sentences were spoken at similar rates, we played videos (Holmes, 2018) indicating the desired pace for each sentence while participants completed the recordings. The sentences had an average duration of 2.5 seconds (s.d. = 0.3). The levels of the digital recordings of the sentences were normalised to the same root-mean-square power.

We simulated manipulations to pitch and VTL-timbre using the 'Change Gender' function in Praat (version 5.4.04; www.fon.hum.uva.nl/praat). We shifted the median pitch of the sentence upwards, which changes the fundamental frequency (f_0) of the sentence and the

161 frequencies of the harmonics. We simulated a change to VTL by changing the formant spacing,
162 which affects the timbre. To ensure that distortions introduced by either manipulation were
163 not cues for discrimination, we created new 'unshifted' versions of the sentences by shifting
164 the formants upwards, then applying the inverse manipulation (to approximate the VTL of the
165 original sentence), then subsequently shifting median pitch up and then down again.

166 Throughout the experiment, each participant heard sentences spoken by their partner
167 (i.e. their familiar voice) and sentences spoken by two unfamiliar talkers, who were the
168 partners of other participants in the experiment who were the same sex as the participant's
169 partner. To counterbalance voice acoustics, we aimed to present sentences spoken by each
170 participant as a familiar voice to one participant (i.e., their partner) and as an unfamiliar voice
171 to two other participants in the experiment. The only exceptions were the partners of the
172 three participants who were not included in the analysis, who were presented as unfamiliar
173 voices but never as familiar. For the same reason, three voices were presented once as familiar
174 and only once as unfamiliar.

175 ***Procedure***

176 First, participants completed the discrimination task. On each trial, participants heard
177 three different sentences spoken by the same talker, presented sequentially. The three
178 sentences could be spoken by the familiar talker or by one of the two unfamiliar talkers. The
179 first sentence was presented in its 'unshifted' version. Either the second or third sentence
180 was the manipulated version (i.e., different pitch or VTL-timbre than the original recording)
181 and the remaining sentence was the 'unshifted' version. In a two-alternative forced-choice
182 task, participants had to indicate which of the two sentences (second or third) had been
183 manipulated. We used a weighted up-down adaptive procedure (Kaernbach, 1991) with a step
184 size of 0.001 and a ratio of 1:9 to estimate each participant's 90% JND for discriminating
185 manipulations to pitch and VTL-timbre. The starting value for each run was 0.0115% above

186 the original median pitch or VTL-timbre and the procedure stopped after 8 reversals. For
187 each talker, we adapted pitch and VTL-timbre separately, producing 6 separate runs (3 talkers
188 x 2 manipulations) that we interleaved.

189 Next, participants completed two tasks: a speech intelligibility task and an explicit
190 recognition task. Half completed the speech intelligibility task first and the other half
191 completed the explicit recognition task first. For both tasks, we presented three voice
192 manipulation conditions: (1) the original pitch and VTL-timbre were preserved ('unshifted'
193 condition), (2) pitch was manipulated to the participant's pitch discrimination threshold
194 ('pitch-manipulated' condition), and (3) formant spacing (an acoustic correlate of VTL-timbre)
195 was manipulated to the participant's formant spacing discrimination threshold ('VTL-
196 manipulated' condition).

197 In the speech intelligibility task, participants heard two sentences spoken
198 simultaneously by different talkers. They had to identify the 4 remaining words from the
199 sentence that began with either "Bob" or "Pat" (counterbalanced across participants), by
200 clicking buttons on a screen (Figure 1). We included two familiarity conditions: (1) the target
201 sentence was spoken by the participant's partner and the masker sentence was spoken by an
202 unfamiliar talker ("Familiar Target" condition), or (2) both sentences were spoken by
203 unfamiliar talkers ("Both Unfamiliar" condition). Both the target and masker sentences were
204 always manipulated in the same way (i.e., VTL-manipulated, pitch-manipulated, or unshifted).
205 Given that Johnsrude et al. (2013) found an interaction between familiarity and target-to-
206 masker ratio (TMR), we presented the two sentences at 4 different TMRs: -6, -3, 0, and +3
207 dB. To discourage participants from using absolute sound level as a cue for the target talker,
208 we roved the overall level of the combined sentences at four levels between ± 1.5 dB. All trial
209 types were randomly interleaved. Participants completed 768 trials, with a short break every

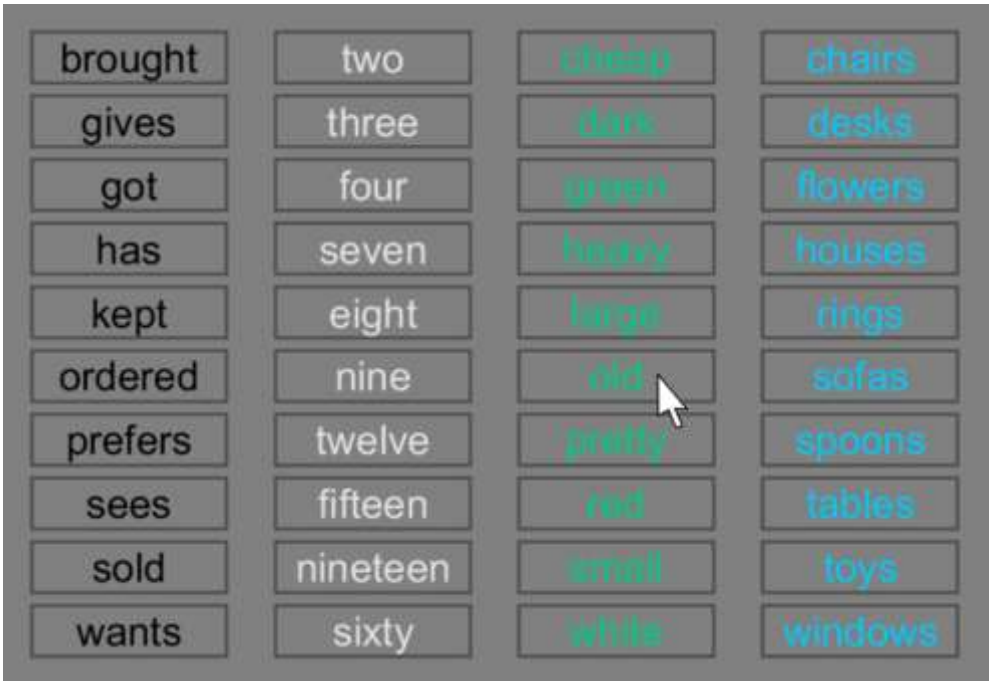


Figure 1. Response screen in the speech intelligibility task. On each trial, participants clicked one word from each column of buttons.

64 trials and a longer break after 384 trials, after which the target Name word (i.e. "Bob" or "Pat") was switched.

There were two different versions of the explicit recognition task. The first 6 participants completed a two-alternative forced-choice (2AFC) discrimination task. On each trial, they heard two sentences presented sequentially. One sentence was spoken by their partner and the other was spoken by one of the two unfamiliar talkers. Participants had to report which of the two sentences was spoken by their familiar talker (first or second sentence). Both sentences for the trial were manipulated in the same way (i.e., VTL-manipulated, pitch-manipulated, or unshifted). Participants completed 48 trials.

The remaining 11 participants completed a yes-no version of the explicit recognition task. On each trial of the yes-no task, listeners heard one sentence. The sentence could be

spoken by the participant's familiar voice or by one of the two unfamiliar voices, and was either VTL-manipulated, pitch-manipulated, or unshifted. Participants had to report whether each sentence was spoken by their familiar partner or not. Participants completed 63 trials: 21 in each voice manipulation condition. The reason we changed the task is because we thought the 2AFC task might inflate recognition—because the 2AFC task could be performed by identifying which of the two talkers was less familiar, rather than by recognizing the partner's voice. This strategy could be particularly useful when the familiar voice was manipulated. Whereas, by eliminating the direct comparison between two voices, the yes-no task assessed familiarity with the partner's voice independent from the other voices in the set.

Analyses

The JNDs were calculated as the median of the last five reversals. For each participant, we averaged JNDs across the two unfamiliar voices. We express the 90% JND threshold as a Weber fraction: The Weber fraction is the JND (i.e., the difference in VTL-timbre or median pitch at threshold) divided by the VTL-timbre or median pitch of the original sentence.

For the speech intelligibility task, we calculated the percentage of sentences in which participants reported all four words correctly.

For the 2AFC explicit recognition task, we calculated percent correct in each manipulation condition. For the yes-no explicit recognition task, we calculated sensitivity (d') with loglinear correction (Hautus, 1995). The loglinear correction means that chance performance is 0.3 for the yes-no task.

This study was not preregistered. Data are available at the following link: https://osf.io/b72d5/?view_only=2250063289014d289c6b4a1c4784d8d7

Results

Discrimination thresholds

Figure 2 illustrates the JNDs in each condition. We used a two-way within-subjects analysis of variance (ANOVA) to compare JNDs across Familiarity (familiar and unfamiliar) and Manipulation (pitch and VTL-timbre) conditions. Participants had significantly greater (i.e., worse) JNDs for detecting manipulations to pitch (mean = 0.065, s.d. = 0.012) than acoustic correlates of VTL (mean = 0.049, s.d. = 0.008) [$F(1, 16) = 16.86$, $p = 0.001$, $\omega_p^2 = 0.47$].

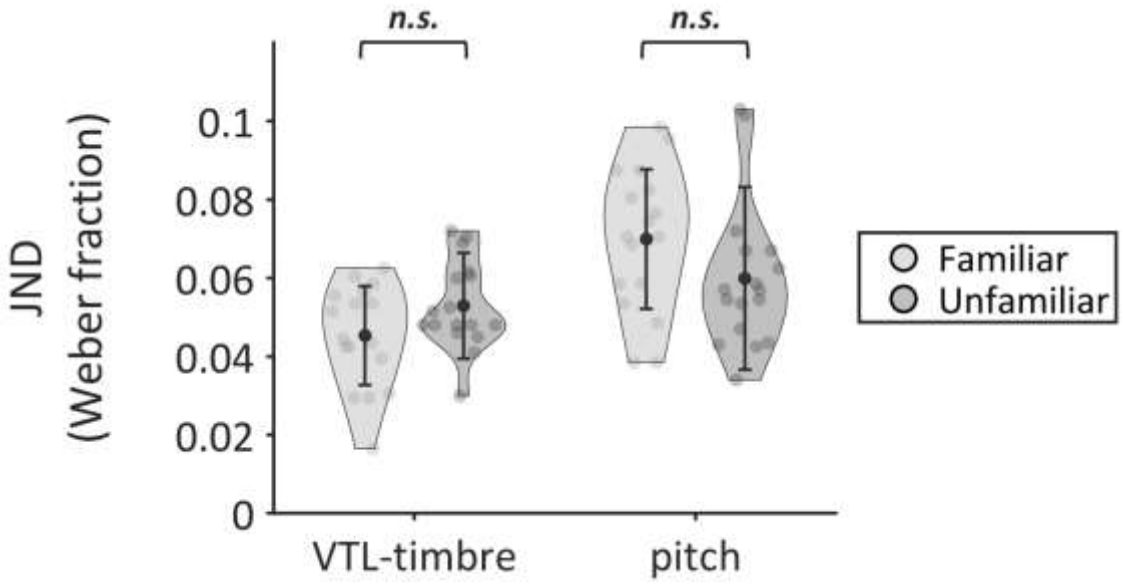


Figure 2. Pitch discrimination for familiar and unfamiliar voices (N=17). Just-noticeable difference (JND), expressed as a Weber fraction, for discriminating pitch and acoustic correlates of vocal tract length (VTL-timbre). Error bars show ± 1 standard error of the mean.

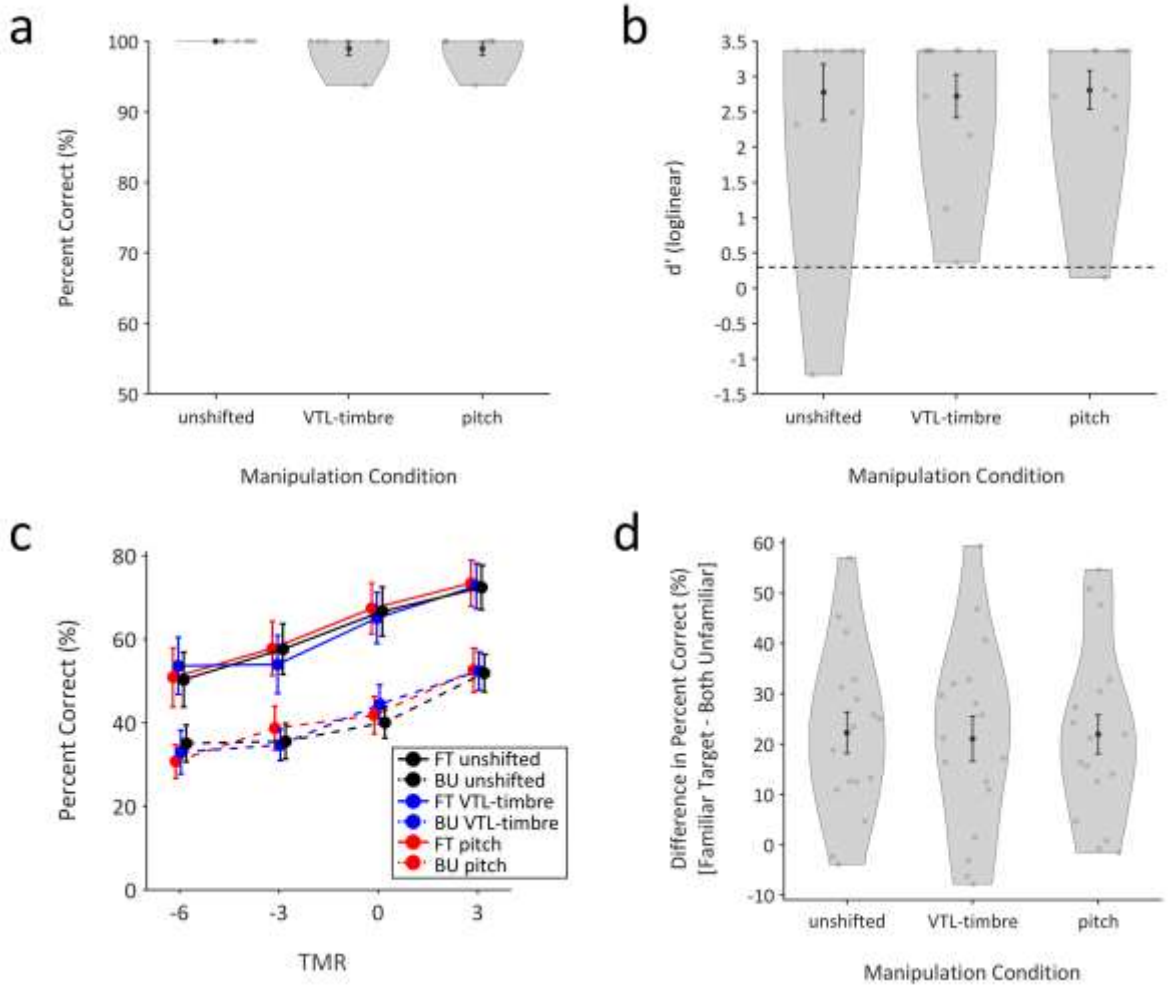
263 Overall, there was no evidence for a difference in JNDs between familiar (mean =
264 0.058, s.d. = 0.013) and unfamiliar (mean = 0.056, s.d. = 0.011) voices [$F(1, 16) = 0.07$, $p =$
265 0.80, $\omega_p^2 = -0.05$]. However, there was a significant interaction between the Familiarity and
266 Manipulation factors [$F(1, 16) = 8.12$, $p = 0.011$, $\omega_p^2 = 0.28$] (see Figure 2). We expected that
267 thresholds might be better for familiar than unfamiliar voices, but t-tests showed no significant
268 difference in JNDs between familiar and unfamiliar voices for pitch [$t(16) = 1.48$, $p = 0.16$, d_z
269 $= 0.36$] or VTL-timbre [$t(16) = 1.78$, $p = 0.10$, $d_z = 0.43$]. Instead, the interaction showed a
270 crossover pattern, with slightly (but not significantly) better VTL-timbre JNDs for familiar than
271 unfamiliar voices, and slightly (but not significantly) worse pitch JNDs for familiar than
272 unfamiliar voices. The interaction was explained by a significant difference between VTL-
273 timbre and pitch JNDs for familiar [$t(16) = 5.50$, $p < 0.001$, $d_z = 1.33$], but not unfamiliar [$t(16)$
274 $= 1.30$, $p = 0.21$, $d_z = 0.32$], voices. This pattern of results shows that participants can
275 discriminate smaller changes to VTL-timbre than pitch for familiar voices, but there is no
276 difference in the ability to discriminate changes to VTL-timbre and pitch for unfamiliar voices.

277 ***Explicit recognition***

278 Participants were able to identify their partner's voice with high accuracy in all voice
279 manipulation conditions.

280 Figure 3a illustrates percent correct on the 2AFC explicit recognition task (range =
281 87.5–100.0%). The data violated the assumption of normality (skewed distributions and $p <$
282 0.05 in Shapiro-Wilk test), so we compared percent correct across the three Manipulation
283 conditions using a Friedman test. The effect of Manipulation was not significant [$X^2(2) = 2.00$,
284 $p = 0.37$].

285 Figure 3b illustrates d′ for the subset of participants who completed the yes-no version
286 of the explicit recognition task. The d′ results also violated the assumption of normality



287

288 **Figure 3.** Explicit recognition and speech intelligibility for voices with their original
289 characteristics, voices manipulated in acoustic correlates of vocal tract length (VTL-timbre),
290 and voices manipulated in pitch. (a) Percentage of correct responses in the two-alternative
291 forced-choice (2AFC) version of the Explicit Recognition task (N=6). (b) Sensitivity (d′ with
292 loglinear correction) in the yes-no version of the Explicit Recognition task (N=11). The dashed
293 horizontal line shows chance d′ (0.3). (c) Percentage of trials in which participants reported
294 the words from the target sentence correctly in the Speech Intelligibility task (N=17), across
295 Familiar Target (FT; solid lines) and Both Unfamiliar (BU; dashed lines) conditions. (d) Familiar-
296 voice benefit (i.e. difference in percent correct between Familiar Target and Both Unfamiliar
297 conditions), collapsed across target-to-masker ratios, in the Speech Intelligibility task (N=17).
298 Error bars in all plots show ± 1 standard error of the mean.

299 (skewed distributions and p < 0.05 in Shapiro-Wilk test), so we compared d' across the three
300 Manipulation conditions using a Friedman test. The effect of Manipulation was not significant
301 [$X^2(2) = 0.95$, $p = 0.62$].

302 ***Speech intelligibility***

303 As can be seen in Figure 3C, intelligibility was better at more favourable TMRs.
304 Baseline performance in the Both Unfamiliar condition was similar across the four
305 Manipulation conditions. Therefore, for each manipulation condition we calculated the
306 speech-intelligibility benefit for the familiar voice by subtracting percent correct in the Both
307 Unfamiliar condition from percent correct in the Familiar Target condition. A large familiar-
308 voice intelligibility benefit, averaging 22%, was observed across all TMRs.

309 We compared the magnitude of the familiar-voice benefit across Manipulation (VTL-
310 manipulated, pitch-manipulated, and unshifted) and TMR (-6, -3, 0, +3) conditions using a two-
311 way within-subjects ANOVA. We found no significant main effect of Manipulation [$F(2, 32) =$
312 0.29, $p = 0.75$, $\omega_p^2 = -0.04$] or TMR [$F(3, 48) = 1.39$, $p = 0.26$, $\omega_p^2 = 0.02$]. The interaction
313 was not significant either [$F(6, 96) = 0.87$, $p = 0.52$, $\omega_p^2 = -0.01$].

314 Figure 3d illustrates the speech-intelligibility benefit across the four manipulation
315 conditions, collapsed across TMRs. One-sample t-tests for each Manipulation condition
316 showed that the familiar-voice benefit was significantly greater than zero in all four conditions
317 ($t \geq 4.61$, $p < 0.001$).

318 ***Correlations between measures***

319 To examine relationships between measures, we calculated Spearman's rank
320 correlation coefficients with Bonferroni correction for 9 tests.

321 First, we investigated whether participants who showed a greater familiar-voice benefit
322 to intelligibility (i.e., difference in percent correct between Familiar Target and Both Unfamiliar

in the unshifted condition) showed a greater difference in thresholds (i.e., difference in JNDs) between familiar and unfamiliar voices. We found no relationship between the magnitude of the familiar-voice benefit to intelligibility and the difference in thresholds—either for pitch thresholds (r = 0.11, p ~ 1.00) or VTL-timbre thresholds (r = -0.14, p ~ 1.00). Second, we assessed the relationship between the familiar-voice benefit and discrimination thresholds using the difference between pitch and VTL-timbre thresholds for familiar voices as the metric. We found no relationship between the familiar-voice intelligibility benefit and the difference in thresholds between pitch and VTL-timbre (r = -0.28, p ~ 1.00).

Table 1. Correlations between the extent of familiarity (measured either by the number of years the participant had known the familiar person or the number of hours they reported speaking to them each week) and familiar-unfamiliar differences in discrimination (pitch or VTL) or speech intelligibility.

	Familiar-unfamiliar difference in JNDs for pitch		**Familiar-unfamiliar difference in JNDs for VTL-timbre**		**Familiar-unfamiliar difference in speech intelligibility**	
	r	*p*	*r*	*p*	*r*	*p*
Years known	0.03	~ 1.00	0.30	~ 1.00	-0.48	0.48
Hours speak per week	-0.43	0.74	-0.43	0.77	-0.09	1.00

Finally, we sought to determine whether the extent of familiarity affected discrimination thresholds or speech intelligibility. We assessed the extent of familiarity using two metrics: the number of years the pair had known each other and the number of hours that the pair spoke to each other each week. As can be seen from Table 1, neither metric correlated significantly with the difference in thresholds between familiar and unfamiliar voices (neither for pitch or VTL-timbre) or the familiar-voice benefit to intelligibility.

Discussion

We replicated the finding that speech spoken by familiar people—here, a participant's friend—is more intelligible than speech spoken by unfamiliar people (Barker & Newman, 2004; Domingo et al., 2020; Domingo, Holmes, Macpherson, & Johnsrude, 2019; Holmes et al., 2018; Holmes & Johnsrude, 2020; Johnsrude et al., 2013; Kreitewolf et al., 2017; Levi et al., 2011; Newman & Evers, 2007; Nygaard & Pisoni, 1998; Nygaard et al., 1994; Souza et al., 2013; Yonan & Sommers, 2000). The magnitude of the familiar-voice benefit to intelligibility in the current experiment (10–25%) appears to be of a similar magnitude as was found by Johnsrude et al. (2013) (10–20%) and Holmes et al. (2018) (15–20%). Yet, when measuring discrimination thresholds, we found no evidence that Weber fractions for pitch or VTL-timbre were better for familiar than unfamiliar voices. We found some evidence that voice familiarity affects acuity, but this was a subtle effect: We found a crossover interaction, which reflected better Weber fractions (and thus acuity) for VTL-timbre than pitch for familiar but not unfamiliar voices. Across participants, discrimination thresholds did not co-vary with the intelligibility benefit or the extent of familiarity with the voice—and when we tested intelligibility with voices that were manipulated in pitch to the extent of the discrimination threshold, these manipulations had no significant effect on the speech intelligibility benefit within subjects. We conclude that the familiar-voice benefit to intelligibility is unlikely to be due to better thresholds for discriminating pitch, or VTL-timbre, for familiar than unfamiliar voices.

Discrimination thresholds are not reliably better for familiar voices

For both familiar and unfamiliar voices, thresholds were better for discriminating VTL-timbre than pitch—although for unfamiliar voices, this trend was non-significant. Broadly speaking, our results are consistent with previous results from Gaudrain et al. (2009), who found that unfamiliar voices needed to be manipulated to a greater extent in correlates of glottal pulse rate (corresponding to pitch) than in VTL for the unfamiliar talkers to be rated as different identities.

We expected to find better thresholds for familiar than unfamiliar voices, which was not supported by the results: We found no significant difference in discrimination thresholds between familiar and unfamiliar voices for either pitch or perceptual correlates of VTL. These findings imply that long-term memory representations for the pitch and for the timbral signature of a given formant spacing of a familiar voice are not more precise than the shorter-term representations used to perform the discrimination task with unfamiliar voices. This result cannot be because participants had become familiar with the unfamiliar voices throughout the experiment, because the discrimination task was run first, and so the unfamiliar voices were highly novel during this task. The lack of a significant difference is unlikely to be because the voices were not sufficiently familiar to provide perceptual benefits, given we found a large intelligibility benefit for the same familiar voices when a competing talker was present.

We found no evidence that variability in discrimination thresholds across participants related to the length of time participants had known each other, or to the familiar-voice intelligibility benefit—which is consistent with the idea that perceptual discrimination does not relate to the intelligibility benefit. In addition, we found no evidence that manipulations to pitch or VTL-timbre affected speech intelligibility (discussed in more detail in the next section),

386 which is also consistent with the hypothesis that better intelligibility for a familiar voice does
387 not depend on precise representations of acoustic characteristics.

388 Discrimination thresholds depended on both familiarity and the manipulated voice
389 characteristic, as demonstrated by an interaction. The difference in JND threshold between
390 VTL-timbre and pitch was larger for familiar than for unfamiliar voices. This interaction seems
391 to be explained by a non-significant trend towards better VTL-timbre discrimination
392 thresholds as well as worse pitch discrimination thresholds for familiar voices. This interaction
393 might reflect a small (but significant) shift in the perceptual weight assigned to the properties
394 of a voice, when a voice is familiar, compared to unfamiliar. Participants might have learnt to
395 rely more heavily on VTL than pitch for familiar voices because vocal tract length is a very
396 stable talker characteristic, whereas pitch varies within talkers. Consistent with this idea,
397 Holmes et al. (2018) found that large changes to VTL-timbre eliminate the ability to recognise
398 familiar voices, whereas perceptually equivalent changes to pitch reduce recognition by a
399 significantly smaller amount. Given that VTL is more stable within a talker than pitch is,
400 changes in VTL-timbre may be more salient for familiar voices than changes in pitch are.

401 While it is possible that the shift in balance towards better VTL-timbre and worse
402 pitch thresholds for familiar voices contributes to the intelligibility benefit, we found no
403 correlation between this difference and the magnitude of the benefit across participants. Also,
404 the effect is subtle: Thus, it is difficult to believe that it explains such a large intelligibility benefit
405 of 20-25% in word-report accuracy.

406 ***Perceptually detectable manipulations to pitch and VTL-timbre have no***
407 ***detectable effect on intelligibility or recognition of familiar voices***

408 In this experiment, we replicated the benefit to speech intelligibility from a familiar
409 target voice (Barker & Newman, 2004; Domingo et al., 2020; Holmes et al., 2018; Johnsrude

et al., 2013; Kreitewolf et al., 2017; Levi et al., 2011; Newman & Evers, 2007; Nygaard & Pisoni, 1998; Nygaard et al., 1994; Souza et al., 2013; Yonan & Sommers, 2000) when the original pitch and VTL-timbre of the familiar voice was preserved. Interestingly, we found that participants gained a familiar-voice benefit of a similar magnitude when the voice had been manipulated in pitch or VTL-timbre to the participant's discrimination threshold. Participants could still reliably recognize their partner's voice when it had undergone the same manipulations in pitch or VTL-timbre. By definition, these manipulations were perceptually discriminable, because we used each participant's 90% threshold. Therefore, these findings suggest that representations of familiar voices are robust to small, but perceptible, manipulations of pitch and VTL-timbre of 4–10% and 2–6%, respectively. If better perceptual discrimination contributed to the familiar-voice benefit to intelligibility, then we would expect that changing the pitch or VTL-timbre of a voice so it is at the 90% discrimination threshold should disrupt the intelligibility benefit for familiar voices.

The robustness of the familiar-voice intelligibility benefit to variations in pitch and VTL-timbre may arise because, in natural listening situations, voice characteristics fluctuate over time. For example, when the same talker produces different vowel sounds, the shape of their vocal cavity changes due to changes in the positions of the articulators, which causes differences in the locations of the formants (Hillenbrand, Getty, Clark, & Wheeler, 1995). Pitch fluctuates throughout the duration of a spoken sentence when a talker speaks emotively (Bänziger & Scherer, 2005). Thus, to recognise a person from their voice or to understand the words they are saying in everyday listening situations—when pitch and VTL-timbre vary naturally—some flexibility is necessary. Although it is desirable for listeners to detect such variations in vocal characteristics, it would not be advantageous for natural within-talker variability in pitch and VTL-timbre to remove the intelligibility benefit for familiar voices or the ability to recognize a voice as familiar. Larger manipulations to pitch and VTL-timbre have

435 been found to affect intelligibility and recognition (Holmes et al., 2018)—but these
436 manipulations were about 5 times bigger than the manipulations here. Therefore, listeners
437 use information about pitch and VTL-timbre to recognise and understand familiar talkers, but
438 do not seem to rely on highly precise representations of pitch and VTL-timbre that are close
439 to the discrimination threshold.

440 ***Other processes likely underlie the familiar-voice intelligibility benefit***

441 Overall, our results suggest that the familiar-voice benefit to intelligibility must arise
442 from processes other than better discrimination of pitch or VTL-timbre. One possibility is
443 that other types of predictions underlie the familiar-voice benefit to intelligibility—such as
444 predictions about pitch contour, intonation, harmonic-to-noise ratio, or rhythm. Pitch and
445 VTL-timbre seemed like the most likely characteristics *a priori*, given they have been shown
446 to contribute to voice recognition (Holmes et al., 2018; LaRiviere, 1975; Lavner et al., 2000,
447 2001; van Dommelen, 1987, 1990). However, Holmes et al. (2018) found that manipulating
448 both pitch and VTL-timbre by a large amount (approximately 5 times larger than the
449 manipulations of the current study) reduced but did not eliminate the familiar-voice benefit
450 to intelligibility—consistent with the idea that other voice attributes could be used to realize
451 the intelligibility benefit.

452 Another explanation for our results—which is consistent with the results of Holmes
453 and Johnsrude (2020)—is that familiar voices are more intelligible because they help listeners
454 resist interference from a competing talker. Holmes and Johnsrude (2020) tested the
455 intelligibility of familiar and unfamiliar voices in different masking conditions. If familiar voices
456 are more intelligible because (any of) their vocal attributes are more predictable, then there
457 should be a familiar-voice benefit in all masking conditions that have similar acoustic
458 properties: the content of the masker should not matter. However, Holmes and Johnsrude
459 (2020) found a significant difference in the familiar-voice benefit to intelligibility among masking

conditions: they found no benefit in modulated noise, a moderate benefit when the masker spoke a sentence in an incomprehensible language, and the largest benefit when the masker spoke an English sentence. These results demonstrate that properties of the masker, rather than the target, affect the intelligibility benefit. The results reported by Holmes and Johnsrude (2020) suggest that better predictions about voice characteristics for familiar than unfamiliar people are unlikely to underlie the familiar-voice benefit to intelligibility, and instead implies that familiarity with a voice affects the active cognitive processes engaged in speech perception (Heald & Nusbaum, 2014). Our results are consistent with the explanation that familiar voices are more intelligible because they help listeners resist interference from a competing talker. This account does not predict that familiar voices presented alone should be more perceptually discriminable than unfamiliar voices, but it predicts that words spoken by familiar people will be more intelligible when a competing talker is present (because greater efficiency enables listeners to better resist interference from the competing talker)—which is exactly what we found.

Implications for accounts of speech processing

For many years, we have known that talker attributes (sometimes termed 'indexical properties') influence the perception of speech content (e.g., the words that are spoken) (Nygaard et al., 1994; Remez et al., 1997). Speech is never heard outside of a particular (talker) context—and speech content and talker information are intermingled in the acoustic signal (see, for example, Lachs et al., 2003). The finding that speech spoken by familiar people is more intelligible than speech spoken by unfamiliar people suggests that talker information is not simply stripped away from the acoustic signal, but rather contributes to speech recognition (Pisoni, 1996). Yet, how familiarity with an interlocutor's voice affects the process of speech recognition has remained unclear.

Under the episodic account of speech recognition (Goldinger, 1996, 1998), words are recognized by comparing the acoustic signal against stored memories, which contain all instances of spoken words that a listener has heard throughout their life: each instance is stored as an episodic memory. If the familiar-voice benefit arises because exposure to someone's voice increases the number of their words that are stored in memory, then we would expect discrimination of voice characteristics to be better for familiar than unfamiliar voices: Listeners would have accumulated episodic memories for familiar voices, and may have few if any episodic memories that are similar to a novel (unfamiliar) voice. Therefore, our results are difficult to reconcile with the episodic account of speech recognition.

Our results also speak against the idea that stored representations of the pitch or VTL-timbre of a familiar voice assist the perceptual normalization process (e.g., Peterson, 1961): if such representations were present, it would be surprising if they were not used to facilitate discrimination.

Under prototype theory (Lavner et al., 2001), we might assume that the familiar-voice benefit arises because the acoustics of familiar voices contribute more to the prototype representation than do unfamiliar voices—because participants have had more exposure to the familiar voice. However, under this explanation, the prototype should be more similar in its acoustic properties to familiar voices, so we would also expect to find better discrimination of acoustic properties for familiar than unfamiliar voices.

For the reasons described above, our results are more consistent with the idea that familiar and unfamiliar voices undergo similar normalization, but processing is more efficient or uses fewer cognitive resources for familiar than unfamiliar voices (Nygaard & Pisoni, 1998; Yonan & Sommers, 2000). This is more consistent with conceptualizations that treat speech

perception as an active process that engages cognition (Friston et al., 2021; e.g., Heald & Nusbaum, 2014).

Conclusions

We predicted that natural familiarity with voices would lead to better thresholds for discriminating pitch or VTL-timbre, but we found no strong evidence for an advantage. Yet, participants received a large (20-25%) intelligibility benefit for same voices when a competing talker was present. Based on our results, it seems unlikely that better representations of pitch or VTL-timbre underlie the familiar-voice benefit to intelligibility that has been robustly observed: first, we found no significant benefit to auditory acuity for familiar voices across the group of participants; second, the magnitude of the familiar-voice benefit did not correlate with the difference in discrimination thresholds among participants; and, third, manipulations to pitch and VTL-timbre at the 90% discrimination threshold did not reduce the magnitude of the intelligibility benefit. We did find a significant crossover interaction between familiarity and manipulated characteristic (pitch or VTL-timbre) for discrimination, indicating that acuity for pitch and VTL-timbre is influenced by voice familiarity, but in opposite directions. This is consistent with a listener placing greater reliance on VTL-timbre than pitch when a voice is familiar.

Author Contributions

E.H. and I.J. designed the experiment and wrote the paper. E.H. collected and analysed the data.

Acknowledgements

The experiment was supported by funding to I.J. from the Canadian Institutes of Health Research (CIHR; Operating Grant: MOP 133450) and the Natural Sciences and Engineering

Research Council of Canada (NSERC; Discovery Grant: 327429-2012). We thank Grace To and Shivaani Shanawaz for their help preparing stimuli and collecting data.

References

Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, *46*(3–4), 252–267. https://doi.org/10.1016/j.specom.2005.02.016

Barker, B. a., & Newman, R. S. (2004). Listen to your mother! The role of talker familiarity in infant streaming. *Cognition*, *94*, 45–53. https://doi.org/10.1016/j.cognition.2004.06.001

Chiba, T., & Kajiyama, M. (1941). *The vowel: Its nature and structure*. Tokyo, Japan: Tokyo-Kaiseikan.

Domingo, Y., Holmes, E., & Johnsrude, I. S. (2020). The benefit to speech intelligibility of hearing a familiar voice. *Journal of Experimental Psychology: Applied*, *26*(2), 236–247. https://doi.org/10.1037/xap0000247

Domingo, Y., Holmes, E., Macpherson, E., & Johnsrude, I. S. (2019). Using spatial release from masking to estimate the magnitude of the familiar-voice intelligibility benefit. *The Journal of the Acoustical Society of America*, *146*(5), 3487–3494. https://doi.org/10.1121/1.5133628

Fant, G. (1960). *Acoustic Theory of Speech Production*. Netherlands: The Hague.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Friston, K. J., Sajid, N., Quiroga-Martinez, D. R., Parr, T., Price, C. J., & Holmes, E. (2021). Active Listening. *Hearing Research*, *399*, 107998. https://doi.org/10.1016/j.heares.2020.107998

552 Gaudrain, E., Li, S., Ban, V. S., & Patterson, R. D. (2009). The role of glottal pulse rate and
553 vocal tract length in the perception of speaker identity. *Proceedings of the Annual*
554 *Conference of the International Speech Communication Association, INTERSPEECH*, 148–151.

555 Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and
556 recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,
557 *22*(5), 1166–1183. https://doi.org/10.1037/0278-7393.22.5.1166

558 Goldinger, S. D. (1998). Echoes of echoes? An episode theory of lexical access. *Psychological*
559 *Review*, *105*(2), 251–279.

560 Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on
561 estimated values of d′. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51.
562 https://doi.org/10.3758/BF03203619

563 Heald, S. L., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Front*
564 *Syst Neurosci*, *8*(March), 35. https://doi.org/10.3389/fnsys.2014.00035

565 Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of
566 American English vowels. *Journal of the Acoustical Society of America*, *97*(5), 3099–3111.
567 https://doi.org/10.1121/1.411872

568 Holmes, E. (2018). *Speech recording videos*. https://doi.org/10.5281/zenodo.1165402

569 Holmes, E., Domingo, Y., & Johnsrude, I. S. (2018). Familiar voices are more intelligible, even
570 if they are not recognized as familiar. *Psychological Science*, *29*(10), 1575–1583.
571 https://doi.org/10.1177/0956797618779083

572 Holmes, E., & Johnsrude, I. S. (2020). Speech spoken by familiar people is more resistant to
573 interference by linguistically similar speech. *Journal of Experimental Psychology: Learning,*
574 *Memory, and Cognition*, *46*(8), 1465–1476. Retrieved from

https://doi.org/10.31234/osf.io/2ebrs

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, *24*(10), 1995–2004. https://doi.org/10.1177/0956797613482467

Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method. *Perception & Psychophysics*, *49*(3), 227–229. https://doi.org/10.3758/BF03214307

Kreitewolf, J., Mathias, S. R., & von Kriegstein, K. (2017). Implicit talker training improves comprehension of auditory speech in noise. *Frontiers in Psychology*, *8*, 1584. https://doi.org/10.3389/fpsyg.2017.01584

Lachs, L., McMichael, K., & Pisoni, D. B. (2003). Speech Perception and Implicit Memory: Evidence for Detailed Episodic Encoding of Phonetic Events. In *Rethinking Implicit Memory* (pp. 215–235). https://doi.org/10.1093/acprof:oso/9780192632326.003.0010

LaRiviere, C. (1975). Contributions of fundamental frequency and formant frequencies to speaker identification. *Phonetica*, *31*(3–4), 185–197. https://doi.org/10.1159/000259668

Lavner, Y., Gath, I., & Rosenhouse, J. (2000). Effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication*, *30*(1), 9–26. https://doi.org/10.1016/S0167-6393(99)00028-X

Lavner, Y., Rosenhouse, J., & Gath, I. (2001). The prototype model in speaker identification by human listeners. *International Journal of Speech Technology*, *4*(1), 63–74. https://doi.org/10.1023/A:1009656816383

Levi, S. V, Winters, S. J., & Pisoni, D. B. (2011). Effects of cross-language voice training on speech perception: whose familiar voices are more intelligible? *The Journal of the Acoustical*

598 *Society of America*, *130*(6), 4053–4062. https://doi.org/10.1121/1.3651816

599 Newman, R. S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal*
600 *of Phonetics*, *35*(1), 85–103. https://doi.org/10.1016/j.wocn.2005.10.004

601 Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception*
602 *& Psychophysics*, *60*(3), 355–376. https://doi.org/10.3758/BF03206860

603 Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-
604 contingent process. *Psychological Science*, *5*(I), 42–46.

605 Pisoni, D. B. (1996). *Some thoughts on "normalization" in speech perception*.

606 Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic
607 information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3),
608 651–666. https://doi.org/10.1037/0096-1523.23.3.651

609 Smith, D. R. R., & Patterson, R. D. (2005). The interaction of glottal-pulse rate and vocal-tract
610 length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of*
611 *America*, *118*(5), 3177–3186. https://doi.org/10.1121/1.2047107

612 Souza, P. E., Gehani, N., Wright, R., & McCloy, D. (2013). The advantage of knowing the
613 talker. *Journal of the American Academy of Audiology*, *24*(January 2013), 689–700.
614 https://doi.org/10.3766/jaaa.24.8.6

615 van Dommelen, W. A. (1987). The contribution of speech rhythm and pitch to speaker
616 recognition. *Language and Speech*, *30*(4), 325–338.
617 https://doi.org/10.1177/002383098703000403

618 van Dommelen, W. A. (1990). Acoustic parameters in human speaker recognition. *Language*
619 *and Speech*, *33*(3), 259–272.

620 Yonan, C. A., & Sommers, M. S. (2000). The effects of talker familiarity on spoken word
621 identification in younger and older listeners. *Psychology and Aging*, *15*(1), 88–99.
622 https://doi.org/10.1037/0882-7974.15.1.88

623