Hearing Research 429 (2023) 108704

Contents lists available at ScienceDirect

Hearing Research

journal homepage: www.elsevier.com/locate/heares

Research Paper

# Intelligibility benefit for familiar voices is not accompanied by better discrimination of fundamental frequency or vocal tract length

Emma Holmes[a,b,*], Ingrid S. Johnsrude[b,c]

[a] Department of Speech Hearing and Phonetic Sciences, UCL, London WC1N 1PF, UK
[b] Brain and Mind Institute, University of Western Ontario, London, Ontario N6A 3K7, Canada
[c] School of Communication Sciences and Disorders, University of Western Ontario, London, Ontario N6G 1H1, Canada

ARTICLE INFO

*Article history:*
Received 30 May 2022
Revised 11 November 2022
Accepted 19 January 2023
Available online 20 January 2023

*Keywords:*
Speech
Voice
Familiar
Discrimination
Vocal tract length
Pitch

ABSTRACT

Speech is more intelligible when it is spoken by familiar than unfamiliar people. If this benefit arises because key voice characteristics like perceptual correlates of fundamental frequency or vocal tract length (VTL) are more accurately represented for familiar voices, listeners may be able to discriminate smaller manipulations to such characteristics for familiar than unfamiliar voices. We measured participants' ($N = 17$) thresholds for discriminating pitch (correlate of fundamental frequency, or glottal pulse rate) and formant spacing (correlate of VTL; 'VTL-timbre') for voices that were familiar (participants' friends) and unfamiliar (other participants' friends). As expected, familiar voices were more intelligible. However, discrimination thresholds were no smaller for the same familiar voices. The size of the intelligibility benefit for a familiar over an unfamiliar voice did not relate to the difference in discrimination thresholds for the same voices. Also, the familiar-voice intelligibility benefit was just as large following perceptible manipulations to pitch and VTL-timbre. These results are more consistent with cognitive accounts of speech perception than traditional accounts that predict better discrimination.



## 1. 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. 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) are present—demonstrating that familiar voices also help us to retrieve linguistic content. A familiar-voice intelligibility benefit has been documented for different types of familiar voice (naturally familiar and lab-trained) masked by different types of competing sound (Barker and Newman, 2004; Domingo et al., 2020; Holmes et al., 2018, 2021; Johnsrude et al., 2013; Kreitewolf et al., 2017; Levi et al., 2011; Newman and Evers, 2007; Nygaard and Pisoni, 1998; Nygaard et al., 1994; Souza et al., 2013; Yonan and Sommers, 2000). Although, we do not fully understand why this familiar-voice intelligibility benefit arises, and theories of speech perception make different predictions. If the benefit relies on acoustic representations that are more precise for familiar than unfamiliar voices (for example, based on representations of voice pitch and/or formant positions), we would expect listeners to be better at discriminating relevant acoustic dimensions for familiar than unfamiliar voices. Whereas, if the benefit relies on more efficient cognitive processing, familiar voices could be more intelligible without better discrimination of voice acoustics for familiar than unfamiliar voices.

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. For example, through a hypothetical process is known as (implicit) "talker normalization" (Nearey, 1998; Sussman, 1986). The finding that familiarity with a talker's voice influences the perception of speech content (i.e., rendering the words that are spoken more intelligible), challenges this view (see, for example, Lachs et al. 2003, Nygaard et al. 1994, Pisoni, 1997, Remez et al. 1997). Greater intelligibility for familiar voices can be explained under exemplar-based, or 'episodic' accounts of speech perception (Goldinger, 1996, 1998), which posit that specific details about particular instances

* Corresponding author: Department of Speech Hearing and Phonetic Sciences, Chandler House, 2 Wakefield Street, London, WC1N 1PF, UK.
*E-mail address:* emma.holmes@ucl.ac.uk (E. Holmes).

https://doi.org/10.1016/j.heares.2023.108704

of speech are stored in memory. Under these accounts, memory traces for familiar voices might allow participants to better match their acoustic properties. The familiar-voice intelligibility benefit can also be explained under the prototype account (Lavner et al., 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 strongly to the prototype representation and—if the prototype is similar enough to the familiar voice (which could be possible for immediate family members known since birth, although seems unlikely for lab-trained voices and for most of the voices that become naturally familiar over our lifetimes, given that we become familiar with multiple voices that have different acoustics)—this prototype could allow acoustic properties to be better recovered for familiar than unfamiliar voices. However, specifically which details of familiar voices are stored and subsequently utilized to benefit speech intelligibility are unclear.

According to the source-filter model of speech production (Chiba and Kajiyama, 1941; Fant, 1960) voice acoustics are the product of the vocal source (vocal-fold vibration) filtered through the vocal tract. Vocal-fold vibration rate affects the perceived pitch of a vocalisation. The length of the vocal tract (including the laryngeal cavity, the pharynx and the oral cavity) determines its resonance characteristics, which manifest as the position of formants in frequency space. This is perceived as a specific timbre (hereafter, referred to as VTL-timbre). These two prominent characteristics—pitch and VTL-timbre—determine whether a voice is heard as male or female, adult or child (e.g., Smith and Patterson 2005), and are important cues to voice identity (Holmes et al., 2018; LaRiviere, 1975; Lavner et al., 2000; Lavner et al., 2001; van Dommelen, 1987, 1990). Even though voice recognition is possible when speech is converted into sine-wave speech (in which fundamental frequency is absent) (Remez et al., 2007; Sheffert et al., 2012), large changes to a voice's fundamental frequency or formant positions degrade voice recognition compared to when these attributes are preserved (Holmes et al., 2018). While there is evidence for a dissociation in how these voice attributes contribute to voice recognition and to the familiar-voice intelligibility benefit, large changes to fundamental frequency and formant positions also reduce the magnitude of the familiar-voice intelligibility benefit (Holmes et al., 2018). Thus, it seems plausible that these characteristics could be critical for the familiar-voice intelligibility benefit.

One possible explanation for the familiar-voice benefit is that people are better at predicting attributes (such as pitch and VTL-timbre) of a familiar voice than an unfamiliar voice—which may allow them to better understand speech spoken by familiar people when it is masked by other sounds. For these attributes to be useful, they do not necessarily need to carry phonetic information directly. For example, people might utilise precise predictions about pitch and VTL-timbre for a familiar voice to help them focus their attention on that person's voice, which would help them to segregate that stream from competing speech or otherwise track and comprehend it when other sounds are present. This explanation aligns with several theories of speech recognition: better predictions could be underpinned by more stored 'episodes' for familiar voices (Goldinger, 1996, 1998), or because a familiar voice could be closer to the prototype representation (Lavner et al., 2001). If either of these explanations are correct, then we would expect that precise representations for familiar voices should allow 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* be associated with better discrimination thresholds for familiar voices—is that familiarity reduces the cognitive demands imposed by speech perception and comprehension (Heald and Nusbaum, 2014; Holmes and Johnsrude, 2020). Familiar voices may be processed more efficiently than unfamiliar voices. For example, understanding speech in familiar voices may rely on different memory systems than unfamiliar voices: long-term memory representations of familiar voices may render speech processing more efficient. Under the ideal adaptor framework (Kleinschmidt and Jaeger, 2015), this could involve retrieving the appropriate generative model for 'normalizing' a particular talker's speech as soon as their identity is recognized from their voice. Unfamiliar voices would not be stored in memory so their mapping to linguistic units would either need to be computed from previous speech (e.g., via extrinsic normalization: see Nearey 1998) which is cognitively demanding, or they would undergo an inappropriate (i.e., less ideal) mapping to a prototype that would be slower to apply. Or, the acoustic attributes that are phonetically diagnostic for a familiar voice may be retrieved when their voice is heard, and this may reduce the range of phonetic hypotheses, reducing working memory load (Heald and Nusbaum, 2014).

Holmes and Johnsrude (2020) presented closed-set sentences (Kidd et al., 2008) in a familiar voice with different kinds of maskers that were roughly equivalent in terms of energetic masking. They reasoned that, if the familiar-voice benefit was due to acoustic characteristics of the familiar voice, the benefit should be obtained regardless of masker type. Instead, they found that the benefit depended on the degree to which the masker shared cognitive processes with the target: The more the masker resembled the target, cognitively, the greater the masking release. Specifically, they observed a large familiar-voice benefit when the masker was a competing talker speaking the same language as the target, a significantly smaller benefit when the masker was a competing talker speaking an language that was unfamiliar to the listener, but no benefit when the masker was unintelligible modulated noise. These results imply that voice familiarity may not help listeners process that voice *per se* (given that such a benefit would be the same regardless of the content of competing sounds), but helps listeners to resist interference from the content of a masker—for example, by making speech processing faster or less cognitively demanding. Crucially, under accounts based on cognitive demand, familiarity would *not* be expected to affect JNDs for discriminating voice attributes in quiet.

Previous studies rule out several other explanations for the familiar-voice intelligibility benefit. For example, if participants are more likely to guess at possible words when a voice is familiar (i.e., a shift in report criterion, or bias), this could inflate intelligibility scores, since some guesses may be correct, whereas words that are not guessed are always incorrect. This potential confound is removed when closed-set materials (i.e., matrix tasks) are used: In closed-set tests, participants report the same number of words on every trial, so they must always guess if they are uncertain. A familiar-voice intelligibility benefit is still robustly observed even when such closed-set tests are used (Domingo et al., 2020; Domingo et al., 2019; Holmes et al., 2018; Holmes and Johnsrude, 2020; Holmes et al., 2021; Johnsrude et al., 2013; Kreitewolf et al., 2017). Another advantage of closed-set tests is that transitional probabilities between words in sentences are strictly controlled and the materials are identical across familiarity conditions. This, therefore, rules out any explanation based on listeners being more able to predict upcoming *words* in sentences spoken by familiar people. In other words, context effects are strictly controlled. Finally, the fact that sentences (rather than words) are scored in some studies (Holmes et al., 2018; Holmes and Johnsrude, 2020; Holmes et al., 2021) means that

faster guessing cannot help, since guesses are likely to be incorrect, so would not improve sentence accuracy. Also, 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.

## 2. Material and methods

### 2.1. 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 h in person each week (median = 17.0 h, 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 et al., 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 $r > 0.47$. In fact, the effect size of the familiar voice benefit to intelligibility found by Johnsrude et al. (2013) was even larger ($d = 1.44$), 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.

### 2.2. 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.

### 2.3. 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 s (s.d. = 0.3). The levels of the digital recordings of the sentences were normalised to the same root-mean-square power.

We manipulated fundamental frequency ($f_0$), corresponding to the percept of voice pitch, and formant positions (i.e., formant frequencies), contributing to the percept of VTL-timbre, using the 'Change Gender' function in Praat (version 5.4.04; www.fon.hum.uva.nl/praat). Using Praat, we shifted the 'median pitch' (in Hertz) of the sentence upwards, which changes the fundamental frequency ($f_0$) of the sentence (and also the frequencies of the harmonics). We simulated a change to VTL by applying a multiplication factor to the formant frequencies (in Hertz), which changes the formant spacing and therefore affects the timbre. To ensure that distortions introduced by either manipulation were not cues for discrimination, we created new 'unshifted' versions of the sentences by shifting the formants upwards, then applying the inverse manipulation (to approximate the VTL of the original sentence), then subsequently shifting median pitch up and then down again.

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

### 2.4. Procedure

First, participants completed the voice discrimination task. On each trial, participants heard three different sentences spoken by the same talker, presented sequentially. The three sentences could be spoken by the familiar talker or by one of the two unfamiliar talkers. The first sentence was presented in its 'unshifted' version. Either the second or third sentence was the manipulated version (i.e., different pitch or VTL-timbre than the original recording) and the remaining sentence was the 'unshifted' version. In a two-alternative forced-choice task, participants had to indicate which of the two sentences (second or third) had been manipulated. We used a weighted up-down (1 up 1 down) adaptive procedure (Kaernbach, 1991) with a step size of 0.1% to estimate each participant's 90% JND for discriminating manipulations to pitch and VTL-timbre. The weighting ratio for the adaptive procedure was 1:9, which means that the feature of interest (i.e., pitch or VTL-timbre) increased by 0.9% when the response was incorrect and decreased by 0.1% when the response was correct. The starting value for each run was 1.15% above the original median pitch or VTL-timbre and the procedure stopped after 8 reversals. For each talker, we adapted pitch and VTL-timbre separately, producing 6 separate runs (3 talkers x 2 manipulations) that we interleaved.

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


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

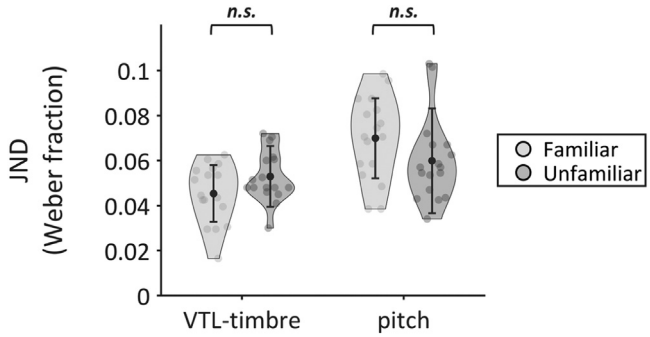


**Fig. 2.** Pitch discrimination for familiar and unfamiliar voices ($N$ = 17). Just-noticeable difference (JND), expressed as a Weber fraction, for discriminating pitch and acoustic correlates of vocal tract length (VTL-timbre). Error bars show ± 1 standard error of the mean.

formant spacing discrimination threshold ('VTL-manipulated' condition).

In the speech intelligibility task, participants heard two sentences spoken simultaneously by different talkers. They had to identify the 4 remaining words from the sentence that began with either "Bob" or "Pat" (counterbalanced across participants), by clicking buttons on a screen (Fig. 1). We included two familiarity conditions: (1) the target sentence was spoken by the participant's partner and the masker sentence was spoken by an unfamiliar talker ("Familiar Target" condition), or (2) both sentences were spoken by unfamiliar talkers ("Both Unfamiliar" condition). Both the target and masker sentences were always manipulated in the same way (i.e., VTL-manipulated, pitch-manipulated, or unshifted; manipulations were applied at each individual's discrimination thresholds for each attribute, separately for each of the three voices they heard). Given that Johnsrude et al. (2013) found an interaction between familiarity and target-to-masker ratio (TMR), we presented targets and maskers at 4 different TMRs: −6, −3, 0, and +3 dB. To discourage participants from using absolute sound level as a cue for the target talker, we roved the overall level of the combined sentences at four levels between ± 1.5 dB. All trial types were randomly interleaved. Participants completed 768 trials, with a short break every 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 voice 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. We used a yes-no procedure 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 (which could be particularly useful when the voices were manipulated, if none of the voices sounded familiar). Whereas, by eliminating the direct comparison between two voices, the yes-no task assessed familiarity with the partner's voice independently from the other voices in the set.

### 2.5. Analyses

The JNDs were calculated as the median of the last five reversals in the adaptive procedure. 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 (after the Name word) 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. Analyses were conducted using SPSS. Data are available at the following link: https://osf.io/b72d5/?view_only=2250063289014d289c6b4a1c4784d8d7

## 3. Results

### 3.1. Discrimination thresholds

Fig. 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 to VTL-timbre (mean = 0.049, s.d. = 0.008) [$F(1, 16) = 16.86$, $p = 0.001$, $\omega_p^2 = 0.47$].

Overall, there was no evidence for a difference in JNDs between familiar (mean = 0.058, s.d. = 0.013) and unfamiliar (mean = 0.056, s.d. = 0.011) voices [$F(1, 16) = 0.07$, $p = 0.80$, $\omega_p^2 = -0.05$]. Furthermore, t-tests on the simple effects indicated no significant difference in JNDs between familiar and unfamiliar voices for pitch [$t(16) = 1.48$, $p = 0.16$, $d_z = 0.36$] or VTL-timbre [$t(16) = 1.78$, $p = 0.10$, $d_z = 0.43$].

However, there was a significant crossover interaction between the Familiarity and Manipulation factors [$F(1, 16) = 8.12$, $p = 0.011$, $\omega_p^2 = 0.28$] (see Fig. 2), with slightly (but not significantly) better VTL-timbre JNDs for familiar than unfamiliar voices, and slightly (but not significantly) worse pitch JNDs for familiar than unfamiliar voices. The interaction was explained by a significant difference

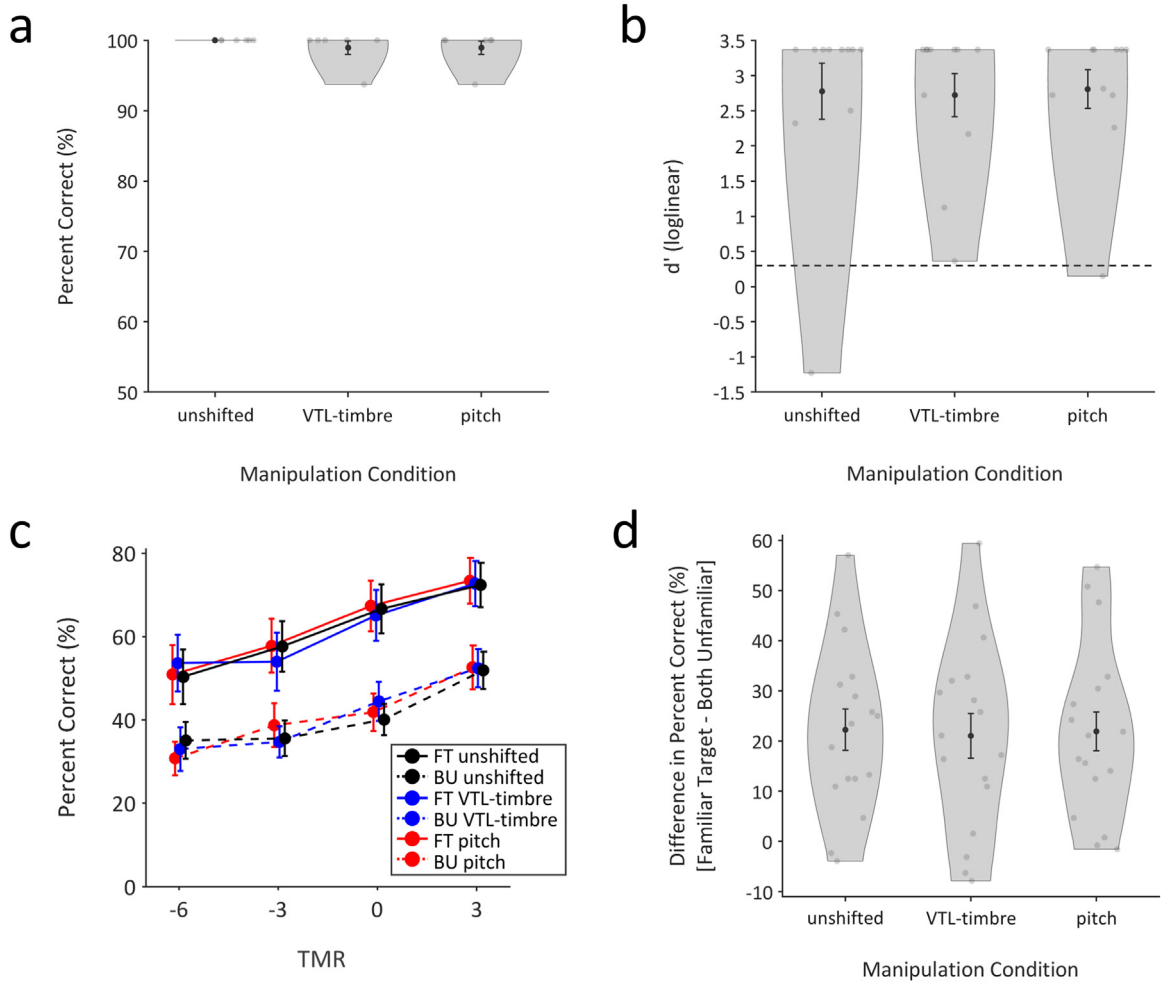


**Fig. 3.** Explicit voice recognition and speech intelligibility for voices with their original characteristics, voices manipulated in acoustic correlates of vocal tract length (VTL-timbre), and voices manipulated in pitch. (a) Percentage of correct responses in the two-alternative forced-choice (2AFC)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

between VTL-timbre and pitch JNDs for familiar [$t(16) = 5.50$, $p < 0.001$, $d_z = 1.33$], but not unfamiliar [$t(16) = 1.30$, $p = 0.21$, $d_z = 0.32$], voices. This pattern of results shows that participants can discriminate smaller changes to VTL-timbre than pitch for familiar voices, but there is no difference in the ability to discriminate changes to VTL-timbre and pitch for unfamiliar voices.

### 3.2. Explicit voice recognition

Participants were able to identify their partner's voice with high accuracy in all voice manipulation conditions. Fig. 3a illustrates percent correct on the 2AFC explicit recognition task (range = 87.5–100.0%). The data violated the assumption of normality (skewed distributions reflecting very high performance and $p < 0.05$ in Shapiro-Wilk test), so we compared percent correct across the three Manipulation conditions using a Friedman test. The effect of Manipulation was not significant [$X^2(2) = 2.00$, $p = 0.37$].

Fig. 3b illustrates d′ for the subset of participants who completed the yes-no version of the explicit recognition task. These data also violated the assumption of normality (skewed distributions and $p < 0.05$ in Shapiro-Wilk test), so we compared recognition (d′) across the three Manipulation conditions using a Friedman test. The effect of Manipulation was not significant [$X^2(2) = 0.95$, $p = 0.62$].

### 3.3. Speech intelligibility

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

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

Fig. 3d illustrates the speech-intelligibility benefit across the four manipulation conditions, collapsed across TMRs. One-sample t-tests for each Manipulation condition showed that the familiar-

**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 |

voice benefit was significantly greater than zero in all four conditions ($t \geq 4.61$, $p < 0.001$).

### 3.4. Correlations between measures

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

First, we investigated whether participants who showed a greater familiar-voice intelligibility benefit (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 intelligibility benefit and the difference in thresholds—either for pitch thresholds ($r = 0.11$, $p \sim 1.00$) or VTL-timbre thresholds ($r = -0.14$, p ~ 1.00). Second, we investigated whether participants who showed a greater familiar-voice intelligibility benefit showed a greater difference between pitch and VTL-timbre thresholds for the familiar voice. We found no relationship between the familiar-voice intelligibility benefit and the difference in thresholds between pitch and VTL-timbre for the familiar voice ($r = -0.28$, $p \sim 1.00$).

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.

## 4. 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 and Newman, 2004; Domingo et al., 2020, 2019; Holmes et al., 2018; Holmes and Johnsrude, 2020; Johnsrude et al., 2013; Kreitewolf et al., 2017; Levi et al., 2011; Newman and Evers, 2007; Nygaard and Pisoni, 1998; Nygaard et al., 1994; Souza et al., 2013; Yonan and Sommers, 2000). The magnitude of the familiar-voice benefit to intelligibility in the current experiment (10–25%) is of a similar magnitude to that observed 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 or VTL-timbre 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.

### 4.1. 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. These results are broadly consistent with previous results from Zaltz et al. (2018), who also found a non-significant trend towards better discrimination thresholds for VTL than pitch when using an unfamiliar voice. To our knowledge, these effects have never previously been studied for familiar voices.

The magnitude of the discrimination thresholds reported here (Weber fraction for pitch of 0.065 and for VTL-timbre of 0.049, which correspond to 1.09 and 0.82 semitones, respectively) differ from those in previous studies, but this can be explained by differences in the design. Here, we measured 90% thresholds, whereas previous studies typically report ~70% thresholds. For example, when measuring 70.7% thresholds for sentence stimuli, Zaltz et al. (2018) found a Weber fraction of 0.037 for discriminating pitch and 0.034 for discriminating VTL, which are slightly lower than the thresholds we report here. Another difference between Zaltz et al. (2018) and the current study was they *lowered* pitch and VTL compared to original stimuli, whereas we raised them—and thresholds for detecting a higher-than-usual pitch may be higher than thresholds for detecting a lower-than-usual pitch (e.g., see Salzberg 1980). Other studies have used shorter speech segments, such as syllables, and have found higher thresholds than we report here. For example, Gaudrain and Başkent (2015) used triplets of consonant-vowel syllables and found a Weber fraction of 0.167 for discriminating pitch and 0.098 for discriminating VTL. The increased content and duration of sentences used in the current study may have contributed to lower thresholds compared to those previously reported for syllables.

We expected to find better thresholds for familiar than unfamiliar voices, but did not: Discrimination thresholds did not differ 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 the formant positions 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 novel during this task. The lack of a significant difference also cannot 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. Instead, these results suggest that experience does not affect auditory acuity. Indeed, musicians—who are assumed to be deeply familiar with the pitch ranges and timbres of their instruments—have pitch discrimination thresholds for

their instruments that are no better than their thresholds for other instruments they have not learnt to play, which is consistent with this conclusion (Holmes et al., 2022).

Variability in discrimination thresholds across participants did not relate 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. Broadly speaking, this independence between discrimination and intelligibility is consistent with previous work showing that, in cochlear implant users, pitch and VTL-timbre JNDs do not correlate with the speech intelligibility benefit participants gain from separating a target and competing voice in pitch and VTL-timbre (Boghdady et al., 2019; El Boghdady et al., 2021). In addition, we found that manipulations to pitch or VTL-timbre did not affect speech intelligibility (discussed in more detail in the next section), which is also consistent with the hypothesis that better intelligibility for a familiar voice does not depend on precise representations of acoustic characteristics.

The use of 90% thresholds in this study ensured that the manipulations were clearly noticeable on the vast majority of trials (if thresholds closer to 50% had been used, these manipulations would have been less extreme). Keeping performance in the discrimination task off ceiling (100%) also allowed us to equate perceptual discriminability for pitch and VTL-timbre. Even though 90% thresholds relate to a flatter part of the psychometric function than 50% or 70% thresholds, we have no reason to believe that our measurements were too noisy to detect significant differences between familiar and unfamiliar voices, since we were still able to observe a significant main effect of the type of manipulation, and a crossover interaction between familiarity and which voice attribute was manipulated.

Although we found no main effect of familiarity on discrimination thresholds, the difference in Weber fraction (difference in JND threshold) between VTL-timbre and pitch was significantly larger for familiar than for unfamiliar voices. This interaction manifested as a crossover interaction, with slightly better VTL-timbre discrimination thresholds, and slightly worse pitch discrimination thresholds, for familiar compared to unfamiliar voices. This interaction might reflect a small (but significant) shift in the perceptual weight assigned to the properties of a voice when a voice is familiar, compared to unfamiliar. Participants might rely more heavily on VTL than pitch for familiar voices because vocal tract length is stable within talkers, whereas pitch varies within talkers: the frequency difference between the first two formants is a cue to the content of a phoneme (e.g., /a/ compared to /i/), but the average spacing between formants (which was modified by the manipulations we applied here) is relatively stable across utterances from an individual voice. Consistent with the idea that VTL-timbre is a more reliable voice characteristic than pitch, Holmes et al. (2018) found that large changes to VTL-timbre eliminate the ability to recognise familiar voices, whereas perceptually equivalent changes to pitch reduce recognition by a significantly smaller amount. Given that VTL is more stable within a talker than pitch is, changes in VTL-timbre may be more salient for familiar voices than changes in pitch are. This idea is discussed in more detail in Section 4.3, below.

Pitch discrimination thresholds were slightly (although not significantly) worse for familiar compared to unfamiliar voices. This effect, if reliable, might be explained by a categorisation attractor effect. In other words, to facilitate generalisation across utterances by the same talker, listeners may become less sensitive to pitch variations for familiar voices, which could help them to group different utterances as belonging to the same voice (Lavan et al., 2019).

Although it is possible that the shift in balance towards better VTL-timbre and worse pitch thresholds for familiar voices contributes to the intelligibility benefit, the effect is small, so is unlikely to explain the large intelligibility benefit of 20–25% in sentence-report accuracy. Also, we found no correlation across participants between the magnitude of the threshold difference and the magnitude of the intelligibility benefit.

### 4.2. Perceptually detectable manipulations to pitch and VTL-timbre have no detectable effect on intelligibility or recognition of familiar voices

In this experiment, we replicated the benefit to speech intelligibility from a familiar target voice (Barker and Newman, 2004; Domingo et al., 2020; Holmes et al., 2018; Johnsrude et al., 2013; Kreitewolf et al., 2017; Levi et al., 2011; Newman and Evers, 2007; Nygaard and Pisoni, 1998; Nygaard et al., 1994; Souza et al., 2013; Yonan and Sommers, 2000). 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. In contrast, even when the voice was manipulated in pitch or VTL-timbre to the participant's discrimination threshold, the intelligibility benefit was preserved, and participants could still reliably recognize their partner's voice. Because we used each participant's 90% discrimination threshold, these manipulations were perceptually salient. We conclude that representations of familiar voices are robust to small, but perceptible, manipulations of pitch and VTL-timbre of 4–10% and 2–6%, respectively.

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 et al., 1995). Pitch fluctuates throughout the duration of a spoken sentence when a talker speaks emotively (Bänziger and 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 been found to affect intelligibility and recognition (Holmes et al., 2018)—but these manipulations were about 5 times bigger than the manipulations here. Therefore, listeners use information about pitch and VTL-timbre to recognise and understand familiar talkers, but do not seem to rely on highly precise representations of pitch and VTL-timbre that are close to the discrimination threshold.

### 4.3. Implications for accounts of speech processing

For many years, we have known that talker attributes (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, 1997). 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 episodic memories for previously-heard words. 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. Given that we become familiar with many voices over our lifetimes that each have distinct acoustics, this explanation seems unlikely to explain the familiar-voice benefit to intelligibility, which has been observed for friends and lab-trained voices as well as long-term spouses. Nevertheless, even if this explanation was plausible, it would also predict better discrimination of acoustic properties for familiar than unfamiliar voices—if the prototype was more similar in its acoustic properties to familiar voices—which is inconsistent with our results.

Our results suggest that the familiar-voice benefit to intelligibility must arise from processes unrelated to better discrimination of pitch or VTL-timbre. One possibility is that other attributes underlie the familiar-voice benefit to intelligibility. This might include predictions related to idiosyncratic phonetics, pitch contour, intonation, harmonic-to-noise ratio, rate, rhythm, or other individual voice characteristics. Note that the benefit cannot simply be explained by glimpsing, because energetic masking was equivalent across familiar and unfamiliar conditions; listeners may, however, use top-down expectations to make better use of glimpses for familiar than unfamiliar voices. Predictions about rate and rhythm would be consistent with speech rate effects on vowel perception (Maslowski et al., 2019), reported benefits of knowing 'when' to listen (e.g., Gaudrain and Carlyon 2013, Holmes et al. 2018, Kitterick et al. 2010) and studies showing that cortical tracking of the amplitude envelope of speech contributes to intelligibility (e.g., Riecke et al. 2018, Zoefel et al. 2018). Pitch and VTL-timbre seemed like the most likely characteristics *a priori*: although they do not carry phonetic information directly, they have been shown to contribute to voice recognition (Holmes et al., 2018; LaRiviere, 1975; Lavner et al., 2000, 2001; van Dommelen, 1987, 1990) and could contribute to speech intelligibility if they are incorporated into phonetic representations alongside other characteristics (as would be predicted by episodic accounts, described above), if they are used for talker normalization, or if they are used to predict where in the acoustic signal to direct attention. However, our results speak against this interpretation. Also, Holmes et al. (2018) found that manipulating both pitch and VTL-timbre by a large amount (approximately 5 times larger than the manipulations of the current study) reduced but did not eliminate the familiar-voice benefit to intelligibility—consistent with the idea that other voice attributes could be used to realize the intelligibility benefit.

Another explanation for our results—which is consistent with the results of Holmes and Johnsrude (2020)—is that familiar voices are more intelligible because they help listeners resist interference from a competing talker. In other words, familiarity with a voice may affect the active cognitive processes engaged in speech perception (e.g., Friston et al., 2021; Heald and Nusbaum, 2014; Holmes and Johnsrude, 2020). For example, familiar and unfamiliar voices may undergo similar normalization, but processing may be more efficient or use fewer cognitive resources for familiar than unfamiliar voices (Nygaard and Pisoni, 1998; Yonan and Sommers, 2000). Greater efficiency could enable listeners to better understand speech when an interfering masker is present, but would not make familiar voices presented alone more perceptually discriminable than unfamiliar voices.

## 5. 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 the same familiar 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 amongst participants; and, third, the benefit to intelligibility for familiar voices manipulated in pitch and VTL-timbre at the 90% discrimination threshold was similar in magnitude to the benefit for voices with the original pitch and VTL-timbre. We did find a trend towards better acuity for VTL-timbre and worse acuity for pitch, for familiar compared to unfamiliar voices. This is consistent with a listener placing greater reliance on VTL-timbre than pitch when a voice is familiar.

## Data availability

Data are available at the following link: https://osf.io/b72d5/.

## CRediT authorship contribution statement

**Emma Holmes:** Conceptualization, Methodology, Software, Investigation, Formal analysis, Visualization, Writing – original draft. **Ingrid S. Johnsrude:** Conceptualization, Writing – review & editing, Supervision, Funding acquisition.

## Acknowledgments

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 Commun. 46 (3–4), 252–267. doi: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. doi:10.1016/j.cognition.2004.06.001.

Boghdady, N.El, Gaudrain, E., Başkent, D, 2019. Does good perception of vocal characteristics relate to better speech-on-speech intelligibility for cochlear implant users?a). J. Acoust. Soc. Am. 145 (1), 417. doi:10.1121/1.5087693.

Chiba, T., Kajiyama, M., 1941. The Vowel: Its Nature and Structure. Tokyo-Kaiseikan, Tokyo, Japan.

Domingo, Y., Holmes, E., Johnsrude, I.S., 2020. The benefit to speech intelligibility of hearing a familiar voice. J. Exp. Psychol. Appl. 26 (2), 236–247. doi: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. J. Acoust. Soc. Am. 146 (5), 3487–3494. doi:10.1121/1.5133628.

El Boghdady, N., Langner, F., Gaudrain, E., Başkent, D., Nogueira, W., 2021. Effect of spectral contrast enhancement on speech-on-speech intelligibility and voice cue sensitivity in cochlear implant users. Ear Hear. 271–289. doi:10.1097/AUD.0000000000000936.

Fant, G., 1960. Acoustic Theory of Speech Production. The Hague, Netherlands.

Faul, F., Erdfelder, E., Buchner, A., Lang, A.-G., 2009. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav. Res. Methods 41 (4), 1149–1160. doi: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. Hear. Res. 399, 107998. doi:10.1016/j.heares.2020.107998.

Gaudrain, E., Başkent, D., 2015. Factors limiting vocal-tract length discrimination in cochlear implant simulations. J. Acoust. Soc. Am. 137 (3), 1298. doi:10.1121/1.4908235.

Gaudrain, E., Carlyon, R.P., 2013. Using Zebra-speech to study sequential and simultaneous speech segregation in a cochlear-implant simulation. J. Acoust. Soc. Am. 133 (1), 502. doi:10.1121/1.4770243.

Goldinger, S.D., 1996. Words and voices: episodic traces in spoken word identification and recognition memory. J. Exp. Psychol. Learn. Mem. Cogn. 22 (5), 1166–1183. doi:10.1037/0278-7393.22.5.1166.

Goldinger, S.D., 1998. Echoes of echoes? An episode theory of lexical access. Psychol. Rev. 105 (2), 251–279.

Hautus, M.J., 1995. Corrections for extreme proportions and their biasing effects on estimated values of d′. Behav. Res. Methods Instrum. Comput. 27 (1), 46–51. doi:10.3758/BF03203619.

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

Hillenbrand, J.M., Getty, L.A., Clark, M.J., Wheeler, K., 1995. Acoustic characteristics of American english vowels. J. Acoust. Soc. Am. 97 (5), 3099–3111. doi:10.1121/1.411872.

Holmes, E. (2018). *Speech recording videos*. 10.5281/zenodo.1165402

Holmes, E., Domingo, Y., Johnsrude, I.S., 2018a. Familiar voices are more intelligible, even if they are not recognized as familiar. Psychol. Sci. 29 (10), 1575–1583. doi:10.1177/0956797618779083.

Holmes, E., Johnsrude, I.S., 2020. Speech spoken by familiar people is more resistant to interference by linguistically similar speech. J. Exp. Psychol. Learn. Mem. Cogn. 46 (8), 1465–1476.

Holmes, E., Kinghorn, E.E., McGarry, L.M., Busari, E., Griffiths, T.D., Johnsrude, I.S., 2022. Pitch discrimination is better for synthetic timbre than natural musical instrument timbres despite familiarity. J. Acoust. Soc. Am. 152 (1), 31–42. doi:10.1121/10.0011918.

Holmes, E., Kitterick, P.T., Summerfield, A.Q., 2018b. Cueing listeners to attend to a target talker progressively improves word report as the duration of the cue-target interval lengthens to 2,000ms. Atten. Percept. Psychophys. 80 (6), 1520–1538. doi:10.3758/s13414-018-1531-x.

Holmes, E., To, G., Johnsrude, I., 2021. How long does it take for a voice to become familiar? Speech intelligibility and voice recognition are differentially sensitive to voice training. Psychol. Sci. 32 (6), 903–915. doi:10.31234/osf.io/bm2uq.

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. Psychol. Sci. 24 (10), 1995–2004. doi:10.1177/0956797613482467.

Kaernbach, C., 1991. Simple adaptive testing with the weighted up-down method. Percept. Psychophys. 49 (3), 227–229. doi:10.3758/BF03214307.

Kidd, G., Best, V., Mason, C.R., 2008. Listening to every other word: examining the strength of linkage variables in forming streams of speech. J. Acoust. Soc. Am. 124 (6), 3793–3802. doi:10.1121/1.2998980.

Kitterick, P.T., Bailey, P.J., Summerfield, A.Q., 2010. Benefits of knowing who, where, and when in multi-talker listening. J. Acoust. Soc. Am. 127 (4), 2498–2508. doi:10.1121/1.3327507.

Kleinschmidt, D.F., Jaeger, T.F., 2015. Robust speech perception: recognize the familiar, generalize to the similar, and adapt to the novel. Psychol. Rev. 122 (2), 148–203. doi:10.1161/CIRCRESAHA.116.303790.The.

Kreitewolf, J., Mathias, S.R., von Kriegstein, K., 2017. Implicit talker training improves comprehension of auditory speech in noise. Front. Psychol. 8, 1584. doi: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. Rethinking Implicit Memory, pp. 215–235. doi:10.1093/acprof:oso/9780192632326.003.0010 In.

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

Lavan, N., Burston, L.F.K., Garrido, L., 2019. How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. Br. J. Psychol. 110 (3), 576–593. doi:10.1111/bjop.12348.

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

Lavner, Y., Rosenhouse, J., Gath, I., 2001. The prototype model in speaker identification by human listeners. Int. J. Speech Technol. 4 (1), 63–74. doi: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? J. Acoust. Soc. Am. 130 (6), 4053–4062. doi:10.1121/1.3651816.

Maslowski, M., Meyer, A.S., Bosker, H.R., 2019. How the tracking of habitual rate influences speech perception. J. Exp. Psychol. Learn. Mem. Cogn. 45 (1), 128–138. doi:10.1037/XLM0000579.

Nearey, T.M., 1998. Static, dynamic, and relational properties in vowel perception. J. Acoust. Soc. Am. 85 (5), 2088. doi:10.1121/1.397861.

Newman, R.S., Evers, S., 2007. The effect of talker familiarity on stream segregation. J. Phon. 35 (1), 85–103. doi:10.1016/j.wocn.2005.10.004.

Nygaard, L.C., Pisoni, D.B., 1998. Talker-specific learning in speech perception. Percept. Psychophys. 60 (3), 355–376. doi:10.3758/BF03206860.

Nygaard, L.C., Sommers, M.S., Pisoni, D.B., 1994. Speech perception as a talker-contingent process. Psychol. Sci. 5 (I), 42–46.

Peterson, G.E., 1961. Parameters of vowel quality. J. Speech Hear. Res. 4, 10–29. doi:10.1044/jshr.0401.10.

Pisoni, D.B., Johnson, K., Mullennix, J.W., 1997. Some thoughts on “normalization” in speech perception. In: Talker Variability in Speech Processing. Academic Press, San Diego, pp. 9–32.

Remez, R.E., Fellowes, J.M., Nagel, D.S., 2007. On the perception of similarity among talkers. J. Acoust. Soc. Am. 122 (6), 3688–3696. doi:10.1121/1.2799903.

Remez, R.E., Fellowes, J.M., Rubin, P.E., 1997. Talker identification based on phonetic information. J. Exp. Psychol. Hum. Percept. Perform. 23 (3), 651–666. doi:10.1037/0096-1523.23.3.651.

Riecke, L., Formisano, E., Sorger, B., Başkent, D., Gaudrain, E., 2018. Neural entrainment to speech modulates speech intelligibility. Curr. Biol. 161–169. doi:10.1016/j.cub.2017.11.033.

Salzberg, R.S., 1980. The effects of visual stimulus and instruction on intonation accuracy of string instrumentalists. Psychol. Music 8 (2), 42–49. doi:10.1177/030573568082005.

Sheffert, S.M., Pisoni, D.B., Fellowes, J.M., Ramez, R.E., 2012. Learning to recognise talkers from natural, sinewave, and reversed speech samples. J. Exp. Psychol. Hum. Percept. Perform. 28 (6), 1447–1469. doi:10.1016/j.biotechadv.2011.08.021.Secreted.

Smith, D.R.R., Patterson, R.D., 2005. The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. J. Acoust. Soc. Am. 118 (5), 3177–3186. doi:10.1121/1.2047107.

Souza, P.E., Gehani, N., Wright, R., McCloy, D., 2013. The advantage of knowing the talker. J. Am. Acad. Audiol. 24 (January 2013), 689–700. doi:10.3766/jaaa.24.8.6.

Sussman, H.M., 1986. A neuronal model of vowel normalization and representation. Brain Lang. 28 (1), 12–23. doi:10.1016/0093-934X(86)90087-8.

van Dommelen, W.A., 1987. The contribution of speech rhythm and pitch to speaker recognition. Lang. Speech. 30 (4), 325–338. doi:10.1177/002383098703000403.

van Dommelen, W.A., 1990. Acoustic parameters in human speaker recognition. Lang. Speech. 33 (3), 259–272.

Yonan, C.A., Sommers, M.S., 2000. The effects of talker familiarity on spoken word identification in younger and older listeners. Psychol. Aging 15 (I), 88–99. doi:10.1037/0882-7974.15.1.88.

Zaltz, Y., Goldsworthy, R.L., Kishon-Rabin, L., Eisenberg, L.S., 2018. Voice discrimination by adults with cochlear implants: the benefits of early implantation for vocal-tract length perception. JARO J. Assoc. Res. Otolaryngol. 19 (2), 193. doi:10.1007/S10162-017-0653-5.

Zoefel, B., Archer-Boyd, A., Davis, M.H., 2018. Phase entrainment of brain oscillations causally modulates neural responses to intelligible speech. Curr. Biol. 28 (3), 401–408. doi:10.1016/j.cub.2017.11.071, e5.