Research Article

Psychological Science
1–13

DOI: 10.1177/0956797621991137
www.psychologicalscience.org/PS

How Long Does It Take for a Voice to Become Familiar? Speech Intelligibility and Voice Recognition Are Differentially Sensitive to Voice Training

Emma Holmes[1], Grace To[1], and Ingrid S. Johnsrude[1,2]
[1]The Brain and Mind Institute, The University of Western Ontario, and [2]School of Communication Sciences and Disorders, The University of Western Ontario

Abstract

When people listen to speech in noisy places, they can understand more words spoken by someone familiar, such as a friend or partner, than someone unfamiliar. Yet we know little about how voice familiarity develops over time. We exposed participants (N = 50) to three voices for different lengths of time (speaking 88, 166, or 478 sentences during familiarization and training). These previously heard voices were recognizable and more intelligible when presented with a competing talker than novel voices—even the voice previously heard for the shortest duration. However, recognition and intelligibility improved at different rates with longer exposures. Whereas recognition was similar for all previously heard voices, intelligibility was best for the voice that had been heard most extensively. The speech-intelligibility benefit for the most extensively heard voice (10%–15%) is as large as that reported for voices that are naturally very familiar (friends and spouses)—demonstrating that the intelligibility of a voice can be improved substantially after only an hour of training.

Keywords

attention, speech perception, auditory perception, memory, learning, open data

Received 1/31/20; Revision accepted 11/13/20

We encounter familiar people every day. Most commonly, these are friends, partners, and family members—but we also encounter people whom we know less well, such as work colleagues or television and radio presenters. As we get to know someone new, we develop the ability to recognize their identity from their voice. We are also better able to understand words spoken by familiar people than people we have never met (Domingo et al., 2019, 2020; Holmes et al., 2018; Kreitewolf et al., 2017; Levi et al., 2008; Nygaard & Pisoni, 1998; Nygaard et al., 1994; Yonan & Sommers, 2000; Zheng et al., 2013). However, our understanding of how long it takes to become familiar with new voices is relatively limited. Here, we investigated the extent to which recognition and intelligibility of a voice improve after different lengths of voice training. (Note that throughout this article, "voice training" and "trained voice" refer to training that familiarized participants with specific voices, not to voice training that the speakers themselves received or the voice in which they spoke.)

Voice Familiarity Improves Speech Intelligibility

Speech can be difficult to understand when several people speak at the same time ("cocktail party problem"; Cherry, 1953, p. 976). Yet when a competing talker is present, large intelligibility benefits have been demonstrated for voices that are highly familiar, such as a spouse the participant has been living with for more than 18 years (Johnsrude et al., 2013) or the

Corresponding Author:
Emma Holmes, University College London, Wellcome Centre for Human Neuroimaging
E-mail: emma.holmes@ucl.ac.uk

participant's mother (Barker & Newman, 2004). Even friends are substantially more intelligible than unfamiliar people (Domingo et al., 2020; Holmes et al., 2018). In fact, Domingo et al. (2020) found no significant difference in the magnitude of the intelligibility benefit for the voices of friends known for at least 1.5 years and the voices of long-term spouses (> 5 years). People who had known each other for less than 1.5 years were not included in the study, which raises the question of how much training on a voice is required to derive the maximum intelligibility benefit.

Several experiments have shown that when presented in conjunction with a masker, voices that participants have been trained on in the lab over 2 to 9 days have better intelligibility than unfamiliar voices (Kreitewolf et al., 2017; Nygaard & Pisoni, 1998; Nygaard et al., 1994; Yonan & Sommers, 2000). However, the magnitude of the speech-intelligibility benefit for artificially trained voices seems to be smaller than for naturally familiar voices: Voice training improves participants' ability to report words by up to 10% (Nygaard et al., 1994) when speech intelligibility is measured in terms of percentage of correct responses or by 0.52 dB (Kreitewolf et al., 2017) when a threshold is estimated on the basis of manipulations of the target-to-masker ratio (TMR). For friends' voices, the benefit has been estimated as 10% to 15% (Domingo et al., 2019, 2020; Holmes et al., 2018) or 5 dB to 9 dB (Holmes & Johnsrude, 2020; Johnsrude et al., 2013). A direct comparison of these studies is difficult because they tested intelligibility with different maskers (white noise in Nygaard et al., 1994; speech-shaped noise in Kreitewolf et al., 2017; a single competing talker in the experiments with familiar voices), and baseline performance in the unfamiliar condition differed across studies. Nevertheless, these findings imply that improved intelligibility of familiar voices is not an all-or-none phenomenon but instead may depend on the length of exposure or the setting in which voices are encountered (trained or natural). Our first aim was to assess whether brief voice training could produce speech-intelligibility benefits and, if so, how speech intelligibility relates to the length of time that participants have been trained on that voice.

Recognition of Familiar Voices

Participants are also able to explicitly recognize voices they have been trained on in the lab (e.g., Doddington, 1985; Nygaard & Pisoni, 1998; Nygaard et al., 1994; Yonan & Sommers, 2000). Studies measuring speech-intelligibility benefits from trained voices have typically measured recognition of those voices at several times during training. For example, Nygaard and Pisoni (1998) found that the ability to identify 10 talkers improved steadily over 9 days of training. Yonan and Sommers (2000) presented participants with 120 sentences spoken by four different talkers on 2 consecutive days; on each day of training, participants were also tested on voice identification for 80 sentences. Young adults' performance was almost perfect after only 1 day of training.

It is unclear how new voices become recognizable following shorter exposures and whether improvements in speech intelligibility parallel improvements in recognition. Our second aim in this study was to compare explicit recognition of a voice with any speech-intelligibility benefit for the same voice. The acoustic features—fundamental frequency and acoustic correlates of vocal-tract length—that are used to recognize voices and to derive the intelligibility benefit from them are at least partially overlapping (Holmes et al., 2018; Nygaard & Pisoni, 1998; Nygaard et al., 1994; Remez et al., 1997; Sheffert et al., 2002). However, Holmes et al. (2018) demonstrated that a familiar voice can benefit intelligibility even if it is not explicitly recognizable. Given that speech intelligibility and voice recognition are partially dissociable, it is plausible that the intelligibility benefit from and the recognition of a previously heard voice may develop at different rates.

Statement of Relevance

Many people find it difficult to understand speech in noisy places. Yet voice familiarity provides a large benefit to intelligibility. We investigated the duration of voice exposure required to improve intelligibility. Speech presented with a competing talker was more intelligible when it was spoken in a voice that was previously heard for 10 to 60 min than in a novel voice. Training for 60 min provided an intelligibility benefit of 10% to 15%, commensurate with the large benefit that has been reported for naturally familiar voices, such as those of friends and spouses. These findings demonstrate that speech intelligibility can be dramatically improved with as little as 1 hr of training, highlighting the great potential of such training for improving intelligibility in everyday settings. This may particularly benefit older people and people with hearing loss, who experience particular difficulty listening in noisy settings, and people whose occupations require accurate speech perception in noisy surroundings, such as aircraft pilots.

Type of Training

The way that voices are trained (i.e., the type of training) has been proposed to influence how voices are learned. Case et al. (2018) examined whether similarity of encoding and retrieval conditions (face-to-face interactions compared with repeating prerecorded sentences) affects implicit learning, although they found no difference. The acoustic background against which novel voices are heard might affect voice learning, but this has not yet been tested. Previously, some researchers have trained participants in quiet contexts (e.g., Nygaard & Pisoni, 1998; Nygaard et al., 1994), whereas others have trained participants in noisy contexts (e.g., Kreitewolf et al., 2017). However, the speech-intelligibility benefits for participants trained on voices in quiet and noisy contexts have not been compared.

Our third aim was to compare two different training conditions: voices presented alone and voices presented in the presence of multitalker babble. We might expect benefits for voices trained in noise if background noise increases cognitive load during training (Mattys et al., 2012), making participants work harder to recognize the voices (Best et al., 2018) and therefore learn the voices more quickly—leading to a larger intelligibility benefit and better recognition. We also might expect benefits for participants trained on voices in noise if noise guides listeners to learn parts of a voice that are most distinct from background noise (Mattys et al., 2005), consistent with transfer-appropriate processing (Morris et al., 1977) and the encoding-specificity hypothesis (Tulving & Thomson, 1973)—which could help participants to better understand speech or recognize a voice when it is masked by similar sounds but would likely have no effect on intelligibility or recognition in quiet.

On the other hand, we might expect benefits for voices trained in quiet if increased cognitive load because of background noise means there are fewer resources available to encode voice information (Rabbitt, 1968)—leading to a larger intelligibility benefit and better recognition following training in quiet. We also might expect benefits for voices trained in quiet if increased background noise masks voice characteristics useful for recognition or intelligibility, such as fundamental frequency or formant frequencies—which might produce distinct effects on intelligibility and recognition, depending on which voice characteristics are masked (Holmes et al., 2018).

The Current Study

Here, we investigated how different amounts of experience with a voice affect recognition and intelligibility. We trained participants on three voices, each speaking for a different amount of time. We then tested participants' ability to identify those voices, and sentences spoken in the same voices, in the presence of another talker. We also compared performance with these trained voices to performance with novel voices that were not heard during training (trained and novel voices were counterbalanced across participants). Half of the participants were trained on the voices with babble noise presented simultaneously, whereas the other half heard the voices alone (in quiet) during training.

Method

Participants

We recruited 53 participants. Of these, three did not complete the study. The remaining 50 participants were between the ages of 18 and 28 years (*Mdn* = 18.6 years, interquartile range = 1.1); 12 participants were men, 36 were women, and two preferred not to disclose gender. A sample size of 50 provides 80% power to detect within-subjects effects (i.e., among four familiarity conditions) of $f \geq .17$, between-subjects effects (i.e., between two training groups) of $f \geq .32$, and within-between-subjects interactions of $f \geq .17$ (Faul et al., 2007). The familiar-voice benefit to speech intelligibility found in previous studies has a large effect size (f = 0.72 in Johnsrude et al. (2013) and f = 0.88 in Holmes et al. (2018), and familiarity effects of this size should be detectable with power of about 100% in the current design.

All participants were native Canadian English speakers and had no history of hearing difficulties. They had average pure-tone hearing levels (HLs; measured at four octave frequencies between 0.5 and 4 kHz) better than 15 dB HL in each ear. The study was approved by Western University's Health Sciences Research Ethics Board, and all participants gave informed consent.

Design

The study contained four parts: familiarization, training, an explicit-recognition test, and a speech-intelligibility test. Schematics of the four tasks are displayed in Figure 1. Participants completed all four parts in a single session that lasted approximately 4 hr. Participants could take breaks between blocks within each task and were encouraged to take longer breaks between tasks.

Each participant heard three voices during familiarization and training. During familiarization, participants heard 10 sentences spoken by each of the three talkers (randomly interleaved). During training, one of the voices was heard speaking 78 sentences, another 156 sentences, and the third 468 sentences. In the explicit-recognition and intelligibility tests, they heard the same

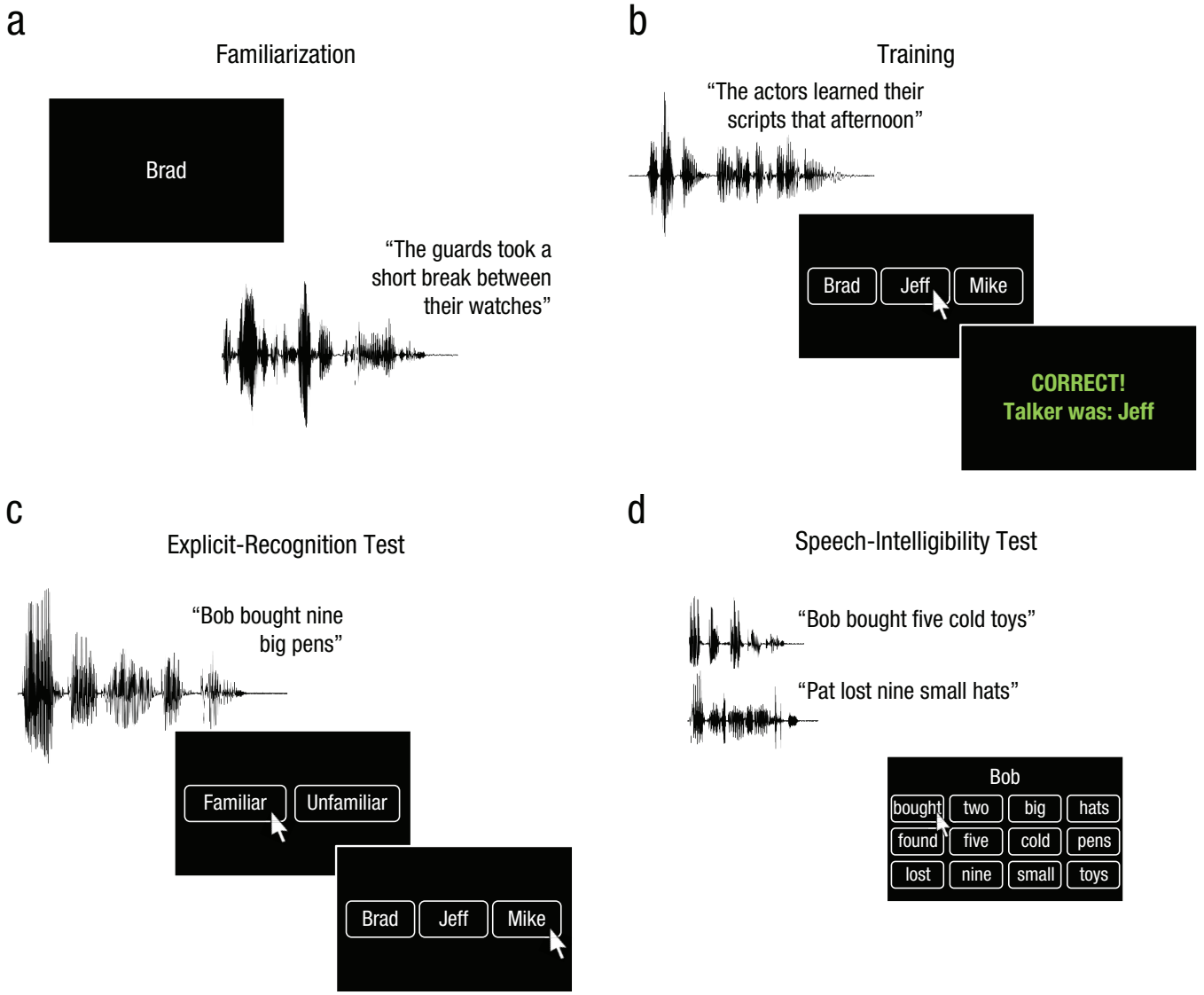

Fig. 1. Schematics of the four phases of the study. In the familiarization phase (a), participants saw a name on the screen (e.g., "Brad") and heard a single meaningful spoken sentence (e.g., "The guards took a short break between their watches") presented without any extraneous noise (i.e., in quiet). Each sentence was spoken by one of three talkers (30 trials, 10 per condition). In the training phase (b), participants heard a single meaningful spoken sentence and had to identify which of three voices spoke each sentence. One of the voices was heard speaking 78 sentences, another 156 sentences, and the third 468 sentences (702 trials in total). Participants received feedback about whether their response was correct and were shown the correct talker name. For half of participants, all of the sentences were presented with babble noise; for the other half, all were presented in quiet. In the explicit-recognition test (c), participants judged whether closed-set sentences were spoken by familiar (three previously heard) or unfamiliar (two novel) talkers. Each sentence was presented in quiet (105 trials, 21 per talker). If the talker was familiar, participants then had to select the name that matched the voice. In the speech-intelligibility test (d), participants simultaneously heard two closed-set sentences, each of which was spoken by a different talker. They had to attend to the sentence that began with a particular name (here, "Bob") and report the other four words from the sentence by clicking one button from each of four columns. Note that only three rows of words are shown here for clarity, but the study always contained eight rows, corresponding to all of the words in the corpus (see Table 1). The competing sentence was always spoken by one of the two unfamiliar talkers, and the target sentence was spoken either by one of the three previously heard talkers or by the other unfamiliar talker (640 trials, 160 per condition).

three voices and two other voices they had not previously heard. Across participants, the five talkers were counterbalanced across familiarity conditions.

Half of participants (n = 25) heard sentences alone during training (i.e., in quiet) and the other half heard the training sentences in the presence of babble noise.

Apparatus

Acoustic stimuli were recorded using a Sennheiser e845-S microphone (Sennheiser Electronic, Wedemark, Germany) connected to a Steinberg UR22 sound card (Steinberg Media Technologies, Hamburg, Germany) in

Table 1. Words From the Boston University Gerald (BUG) Corpus Used in the Speech-Intelligibility Test

Column 1	Column 2	Column 3	Column 4
bought	two	big	bags
found	three	blue	cards
gave	four	cold	gloves
held	five	hot	hats
lost	six	new	pens
saw	eight	old	shoes
sold	nine	red	socks
took	ten	small	toys

Note: In the speech-intelligibility test, participants simultaneously heard two different five-word sentences, one beginning with "Bob" and the other with "Pat," and they had to report the other four words from the sentence by clicking one of eight options from each of the four columns shown.

a single-walled, sound-attenuating booth (Model CL-13 LP MR; Eckel Industries, Morrisburg, Ontario, Canada). Stimuli were recorded in monophonic sound at a sampling rate of 44100 Hz.

During the study, participants sat in a comfortable chair in the same single-walled, sound-attenuating booth facing a 24-in. LCD monitor (either ViewSonic VG2433SMH or Dell G2410t). Acoustic stimuli were presented through the sound card and were delivered diotically through Grado Labs SR225 headphones (Grado Labs, Brooklyn, NY). Acoustic stimuli were presented at a comfortable listening level—approximately 67 dB(A) sound-pressure level. We maintained the same overall presentation level throughout the study (regardless of whether participants heard a single sentence, two sentences, or one sentence mixed with babble noise), so the level of the target sentence differed between the recognition and intelligibility tests and also between the higher and lower TMR conditions in the speech-intelligibility test.

Stimuli

Two different sentence corpora were used in this study: one for familiarization and training and another for testing. For familiarization and training, we used meaningful sentences based on the sentence corpus used by Rodd et al. (2005), such as "The boy was able to climb the mountain." We also created new sentences based on the syntactic and syllabic structure of the existing sentences. We used 354 meaningful sentences in total (177 from Rodd et al., 2005), which are listed in the Supplemental Material available online. Of these, 351 were presented during training. During familiarization, the remaining three sentences were presented to all participants, along with a subset of 27 sentences (a different subset for different participants) that were also presented during training. We used such everyday naturalistic sentences because we wanted to approximate the natural phonetic, phonological, and semantic variety encountered when one becomes familiar with a talker in everyday life.

For testing, we wished to harness the psychometric accuracy of a closed-set procedure. We used the word set from the Boston University Gerald (BUG) corpus (Kidd et al., 2008), recorded as sentences by our participants. These sentences each contain five words, in the form "*Name verb number adjective noun*." An example is "Bob found three old socks."

Word report for open-set, everyday sentences is problematic. If participants are biased to report guesses, they will report more words, which could lead to higher intelligibility values when the total number of words reported is not taken into account. Guessing in open-set tests is likely to lead to correct responses when sentences are semantically meaningful or when there are few lexical neighbors (Sommers et al., 1997). In contrast, the BUG matrix task requires participants to generate exactly four responses (one for each word after the name) on each trial. Intelligibility is therefore unconfounded by guessing. We used two name words ("Bob" and "Pat"). The other words each had eight possible options, which are displayed in Table 1. We created a subset of 384 sentences to be recorded from the BUG word set. The probabilities of each pair of words occurring together within a sentence were equated across the set.

We recorded five male talkers (20–24 years old) speaking the same 738 sentences. All talkers had a Canadian accent and had no speech impediments. 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. Participants saw the written sentence on the screen and were instructed to speak each word at the same time that a vertical bar passed the beginning of the written word. They were told to speak the sentences as naturally as possible. The recorded familiarization and training sentences each had an average duration of 3.1 s (SD = 0.7). The recorded test sentences each had an average duration of 2.4 s (SD = 0.2). The levels of recorded sentences were normalized for root-mean-square power. The babble noise was a mixture of 12 male and female talkers speaking different sentence material.

Procedure

Familiarization. During familiarization (Fig. 1a), participants passively listened to 30 unique meaningful

sentences. Of these, 10 sentences were spoken by each of three talkers. As participants listened to each sentence, a name word—"Brad," "Jeff, " or "Mike"—was displayed on the screen. Participants were asked to associate each of these names with the talker's voice. After the sentence had ended, participants clicked a mouse to hear the next sentence. Across participants, the sentences spoken by each talker were counterbalanced, as were the name words assigned to each voice. The familiarization procedure lasted approximately 10 min.

Training. In the training phase (Fig. 1b), participants completed 702 trials. Each trial contained one sentence in one of the three voices. They heard one talker (*most familiar*) speak 468 sentences (i.e., 67% of sentences), another talker (*moderately familiar*) speak 156 sentences (i.e., 22% of sentences), and the remaining talker (*least familiar*) speak 78 sentences (i.e., 11% of sentences). We selected the number of sentences to roughly correspond to 60, 20, and 10 min of training (respectively), which we predicted would be sufficient to observe differences among training conditions. Participants heard 351 unique sentences during training: Each sentence was heard twice (once spoken by the moderately familiar or least-familiar talker and once or twice spoken by the most-familiar talker). Across participants, the voices and sentences assigned to the three familiarity conditions were counterbalanced.

Half of participants (n = 25) heard the training sentences alone (i.e., in quiet), as during familiarization, whereas the other half heard the same sentences in the presence of simultaneous babble noise, which was presented at a signal-to-noise ratio of 0 dB. The sentence-babble mixtures were presented diotically. The babble noise began 250 ms before the sentence began and ended 250 ms after the sentence had been spoken.

After each spoken sentence had ended, a pop-up box appeared on the screen prompting participants to indicate the name associated with the talker who spoke the sentence. Feedback was provided after each response: Text presented on the screen told participants whether they had answered correctly or incorrectly and displayed the correct name; the text was colored green if participants had answered correctly and red if they had answered incorrectly. Training lasted approximately 1.5 hr and was divided into six blocks, each containing 117 trials.

Explicit-recognition test. The explicit-recognition test was presented after training. It tested whether participants recognized previously heard (*trained*) and novel voices as familiar and unfamiliar and whether participants could identify the names associated with the trained voices. During the explicit-recognition test (Fig. 1c), participants heard BUG sentences spoken by the three trained talkers and two novel talkers. After each sentence, they indicated whether or not they had heard the talker during the training phase. If they indicated they had heard the talker during training, they were prompted to indicate the name of the talker. No feedback was provided. The recognition task contained 105 trials, each containing a unique BUG sentence: 21 sentences were spoken by each of the five talkers. Across participants, the sentences assigned to the five talkers (and conditions) were counterbalanced.

Speech-intelligibility test. Finally, participants completed a speech-intelligibility test. On each trial of this test, participants simultaneously heard two BUG sentences, each of which was spoken by a different talker. The BUG sentences were different from those presented during the explicit-recognition test. The five words contained within the two simultaneously presented sentences were always different. Participants were instructed to listen to the target sentence that began with a specified name word ("Bob" or "Pat") and report the remaining four words from that sentence by clicking words from a list of options on the screen, in any order (see Fig. 1d). Participants completed 640 trials; each of the two name words—either "Bob" or "Pat"—was the target for half of the trials (in separate blocks counterbalanced across participants).

The target sentence could be spoken in any of the five voices, representing four training conditions—most familiar, moderately familiar, least familiar, and unfamiliar (two voices). The masker sentence was always in one of the two unfamiliar voices, and when the target was in an unfamiliar voice, the other unfamiliar voice spoke the masker sentence. Equal numbers of trials (160) were administered in the four conditions. Half of the masker sentences within each condition were spoken by one of the unfamiliar talkers; the other half were spoken by the other unfamiliar talker. This aspect of the design ensured that the masker talkers in all four conditions were identical. Within each condition, we presented the sentences at two different TMRs: −6 and +3 dB.

Analyses

For the explicit-recognition test, hits (correct responses) were defined as trials in which participants heard one of the trained talkers and identified the correct name. Misses were defined as trials in which participants responded that the trained talker was unfamiliar or selected an incorrect name. Hits and misses were calculated separately for the three trained talkers. Correct-rejection and false-alarm rates were calculated from the 42 trials in which participants heard a novel voice.

Sensitivity (Hautus, 1995) was calculated for each of the three trained talkers. Chance d' was 0.3, and the maximum attainable d' was 4.3.

For the speech-intelligibility test, we calculated the percentages of sentences that were reported correctly in each of the 16 conditions (4 familiarity conditions × 2 TMR conditions × 2 training groups). We calculated the familiar-voice-intelligibility benefit as the difference in the percentage of correct responses between the unfamiliar baseline condition and each of the three conditions in which a trained talker was the target (most familiar, moderately familiar, and least familiar).

To examine whether the pattern of results across manipulations differed significantly between the speech-intelligibility and explicit-recognition tasks, we converted d' from the explicit-recognition task and the familiar-voice benefit to intelligibility (the difference in the percentage of correct responses between each of the familiar conditions and the unfamiliar baseline condition) into z scores, which were calculated separately for the two training groups.

We conducted four planned mixed analyses of variance (ANOVAs). One ANOVA was conducted for each task (training, explicit recognition, and speech intelligibility) to compare training groups (between subjects) and familiarity conditions (within subjects). For the speech-intelligibility test, we included an additional factor of TMR (within subjects). A fourth ANOVA directly compared performance on the explicit-recognition test and speech-intelligibility test. We did not correct for multiple ANOVAs, given that all of these were planned analyses. Where Mauchly's test of sphericity was significant, we report statistics with Greenhouse-Geisser correction.

We also conducted a two-way within-subjects ANOVA to look for voice learning over the course of the study, comparing the percentage of correct responses between the beginning and end of the speech-intelligibility task. These results are also uncorrected and should be treated as exploratory because this analysis was unplanned. For this analysis, we took the first 20 trials from each familiarity-by-TMR condition (i.e., 40 trials total) and the last 20 trials from each familiarity-by-TMR condition (i.e., 40 trials total). The ANOVA compared effects of familiarity and trial position (beginning or end), collapsing across the two TMRs and training groups.

In all instances, we calculated effect sizes and confidence intervals (CIs) for effect sizes using MOTE (Buchanan et al., 2018).

Results

Training

Recognition performance in the training conditions was high (Fig. 2a). The most-familiar voice was recognized correctly on 98.3% (SE = 0.2) of trials, the moderately familiar voice was recognized correctly on 94.4% (SE = 1.0) of trials, and the least-familiar voice was recognized correctly on 94.3% (SE = 1.2) of trials. A two-way mixed ANOVA investigating whether performance during training differed across groups (quiet, babble; between subjects) and familiarity conditions (most familiar, moderately familiar, least familiar; within subjects) revealed no effect of group, $F(1, 48) = 0.36$, $p = .55$, $\omega_p^2 = -.01$, 95% CI = [.00, 1.00],[1] and no significant interaction between group and familiarity, $F(2, 96) = 0.21$, $p = .81$, $\omega_p^2 = -.01$, 95% CI = [.00, 1.00]. There was, however, a significant effect of familiarity, $F(1.4, 67.3) = 12.64$, $p < .001$, $\omega_p^2 = .19$, 95% CI = [.01, .28], with better performance in the most-familiar condition than in the moderately familiar condition, $t(49) = 4.29$, $p < .001$, $d_z = 0.61$, 95% CI = [0.30, 0.91], and in the least-familiar condition, $t(49) = 3.56$, $p = .001$, $d_z = 0.50$, 95% CI = [0.21, 0.80]. Performance did not differ between the moderately familiar and least-familiar conditions, $t(49) = 0.16$, $p = .88$, $d_z = 0.2$, 95% CI = [–0.26, 0.30].

Explicit recognition

The percentage of correct responses in the explicit-recognition test was good but below ceiling. The most-familiar voice was recognized correctly on 73.2% of trials, the moderately familiar voice was recognized correctly on 73.1% of trials, and the least-familiar voice was recognized correctly on 73.4% of trials. The unfamiliar voices were correctly recognized as unfamiliar on 84.4% of trials. Hits and false-alarm rates are displayed in Table 2.

Figure 2b shows the d' values in the explicit-recognition test. We conducted a two-way mixed ANOVA with the factors training group (quiet, babble) and familiarity (most familiar, moderately familiar, least familiar). Again, there was no effect of training group, $F(1, 48) < 0.01$, $p = .95$, $\omega_p^2 = -.02$, 95% CI = [.00, 1.00], and no interaction, $F(2, 96) = 2.58$, $p = .08$, $\omega_p^2 = .02$, 95% CI = [.00, .08]. There was also no effect of familiarity, $F(2, 96) = 0.12$, $p = .89$, $\omega_p^2 = -.01$, 95% CI = [.00, .08]. Collapsing across training groups, we compared recognition d' in each familiarity condition with chance level ($d' = 0.3$) using sign tests. Participants were able to identify all three voices with above-chance accuracy ($S \geq 40$, $p < .001$).

Speech intelligibility

Intelligibility data are shown in Figure 3. A three-way mixed ANOVA with the factors training group (quiet, babble), familiarity (most familiar, moderately familiar, least familiar, unfamiliar), and TMR (–6 dB, +3 dB; within subjects) revealed no effect of group, $F(1, 48) =$

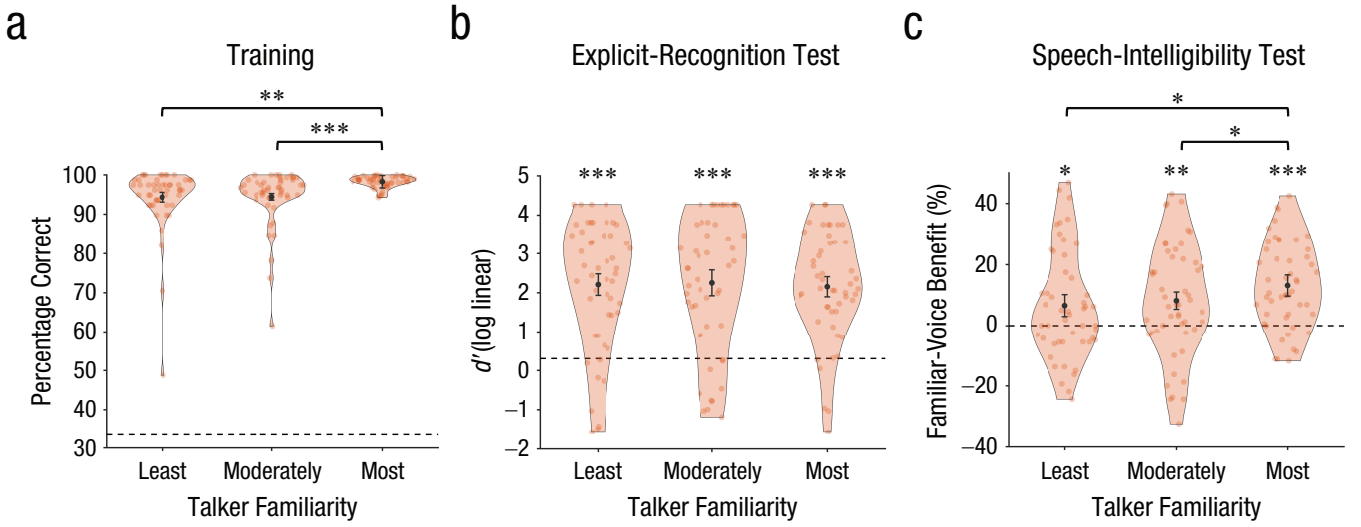

Fig. 2. Task performance in each of the three familiarity conditions. In (a), the percentage of correct responses in the training phase is shown. The dashed line indicates chance performance (33%). In (b), d' values in the explicit-recognition test are displayed. The dashed line indicates chance d'. In (c), the familiar-voice benefit to intelligibility is shown, collapsed across target-to-masker ratios. The familiar-voice benefit is the difference in the percentage of correct responses between each of the familiar conditions and the unfamiliar baseline condition. The shaded region in each plot shows the density of the data; data from individual participants is indicated by transparent dots. The black dots represent means, and error bars show 95% within-subjects confidence intervals (Morey, 2008). Asterisks indicate the significance of performance in each condition compared with chance (*$p < .05$, **$p < .01$, ***$p < .001$).

0.20, $p = .66$, $\omega_p^2 = -.02$, 95% CI = [.00, 1.00], and no interactions involving group—Group × TMR: $F(1, 48) = 1.19$, $p = .28$, $\omega_p^2 < .01$, 95% CI = [.00, .09]; Group × Familiarity: $F(3, 144) = 0.17$, $p = .91$, $\omega_p^2 = -.01$, 95% CI = [.00, 1.00]; Group × Familiarity × TMR: $F(3, 144) = 0.15$, $p = .93$, $\omega_p^2 = .00$, 95% CI = [.00, 1.00].

There was an interaction between TMR and familiarity, $F(3, 144) = 7.47$, $p < .001$, $\omega_p^2 = .03$, 95% CI = [.00, .10]. The overall pattern across familiarity conditions was generally preserved at the two TMRs. All familiar voices were more intelligible than the unfamiliar voices, −6 dB TMR: $t(49) \geq 2.11$, $p \leq .040$, $d_z = 0.30$; +3 dB TMR: $t(49) \geq 2.89$, $p \leq .006$, $d_z = 0.41$, and the moderately familiar voice did not differ from the least-familiar voice, −6 dB TMR: $t(49) = 0.26$, $p = .79$, $d_z = 0.04$, 95% CI = [−0.24, 0.31]; +3 dB TMR: $t(49) = 1.03$, $p = .31$, $d_z = 0.15$, 95% CI = [−0.13, 0.42]. At both TMRs, the most-familiar voice was more intelligible than the moderately familiar voice, although this difference was significant only at the higher TMR, −6 dB TMR: $t(49) = 1.61$, $p = .11$, $d_z = 0.23$, 95% CI = [−0.05, 0.51]; +3 dB TMR: $t(49) = 2.56$, $p = .014$, $d_z = 0.36$, 95% CI = [0.07, 0.65].

Consistent with these comparisons, results showed that the main effect of familiarity was significant, $F(2.6, 125.4) = 10.49$, $p < .001$, $\omega_p^2 = .12$, 95% CI = [.03, .23]. Overall, intelligibility was significantly better at +3 dB than at −6 dB TMR, $F(1, 48) = 140.96$, $p < .001$, $\omega_p^2 = .59$, 95% CI = [.38, .74].

Dissociation between recognition and intelligibility

In the analyses above, there was a significant effect of familiarity for speech intelligibility but not for explicit recognition. To examine whether the pattern of results across familiarity conditions differed between the speech-intelligibility and explicit-recognition tasks, we converted d' from the explicit-recognition task (Fig. 2b)

Table 2. Mean Hit and False-Alarm Rates in the Explicit-Recognition Test

	Hits			
Training group	Most-familiar condition	Moderately familiar condition	Least-familiar condition	False alarms
Quiet	.73 (.06)	.67 (.06)	.79 (.06)	.16 (.04)
Babble	.73 (.05)	.79 (.06)	.68 (.07)	.17 (.04)

Note: Values in parentheses are ±1 *SEM*.

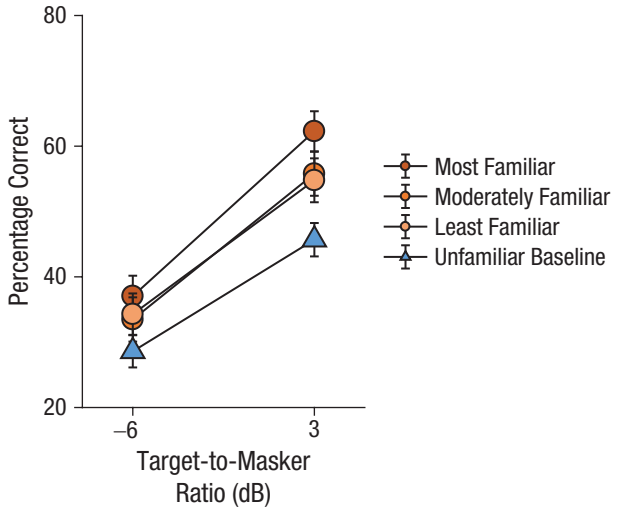

Fig. 3. Percentage of correct responses on the speech-intelligibility test as a function of target-to-masker ratio and familiarity condition. Error bars show ±1 *SEM*.

and the speech-intelligibility-benefit scores (Fig. 2c) for each of the three familiar voices into *z* scores and entered the data into a two-way within-subjects ANOVA. Given that effects involving training group were never significant, we pooled the groups for this analysis. The two-way interaction between task (explicit recognition, speech intelligibility) and familiarity (most familiar, moderately familiar, least familiar) was significant, $F(1.8, 86.7) = 4.55$, $p = .017$, $\omega_p^2 = .01$, 95% CI = [.00, .08]. We confirmed with follow-up one-way ANOVAs that this was driven by a significant effect of familiarity on speech-intelligibility-benefit scores, $F(2, 98) = 4.27$, $p = .017$, $\omega_p^2 = .04$, 95% CI = [.00, .13], and a nonsignificant effect of familiarity on explicit-recognition scores, $F(2, 98) = 0.15$, $p = .87$, $\omega_p^2 = .00$, 95% CI = [.00, 1.00]. These results confirm that the pattern of findings across familiarity conditions differed between the explicit-recognition and speech-intelligibility tasks.

No evidence for voice learning during the intelligibility task

In an exploratory analysis, we investigated whether participants achieved better intelligibility at the end of the intelligibility task than at the beginning—which could potentially indicate voice learning during the intelligibility task (Fig. 4). We replicated the effect of familiarity, $F(3, 147) = 8.42$, $p < .001$, $\omega_p^2 = .04$, 95% CI = [.00, .10], but found no main effect of trial position, $F(1, 49) = 0.01$, $p = .91$, $\omega_p^2 = .00$, 95% CI = [.00, 1.00], and no interaction between trial position and familiarity, $F(3, 147) = 0.37$, $p = .77$, $\omega_p^2 = .00$, 95% CI = [.00, 1.00]. Thus, we found no evidence that intelligibility improved throughout the duration of the task for any of the voices; voice learning instead seems to have been restricted to the training phase of our study.

Discussion

Our results demonstrate that recognition and improved speech intelligibility for previously heard (trained) voices emerge rapidly. Even for the least-familiar talker, to which participants were exposed for approximately 10 min, we found successful voice recognition and a significant intelligibility benefit. We found a different pattern of results for explicit recognition and intelligibility, confirmed by a significant interaction between task and familiarity. Explicit recognition did not differ among the three voices that had been trained for different durations. However, speech intelligibility was best for the most-familiar voice and significantly lower for the moderately familiar and least-familiar voices. Thus, whereas recognition was relatively stable over the range of exposures we tested, intelligibility was better after longer durations of training.

We found a significant intelligibility benefit for all three trained (i.e., familiar) talkers, despite the fact that training focused on voice identification rather than reporting speech content. This demonstrates that when people learn about voices, they learn characteristics that subsequently can be exploited to enhance intelligibility, even when intelligibility is not challenged during learning. Our results demonstrate that training on voice identification provides rapid learning that improves intelligibility of the trained voices after as little as 10 min of voice training.

The benefit that we observed for the most-familiar talker (10%–15%) is similar in magnitude to that previously reported for naturally familiar voices with the same masker and task (Domingo et al., 2020; Holmes & Johnsrude, 2020). This highlights the great potential of voice training for improving intelligibility in everyday settings. Voice training may be particularly beneficial for older people or people with hearing loss who experience particular difficulty listening in noisy settings (e.g., Dubno et al., 1984). There is already some evidence that older people (Yonan & Sommers, 2000) and older people with confirmed hearing loss (Souza et al., 2013) benefit from familiar-voice information. Unlike participants in previous experiments that tested the intelligibility of trained talkers after 2 or more days of training (Kreitewolf et al., 2017; Nygaard & Pisoni, 1998; Nygaard et al., 1994; Yonan & Sommers, 2000), our participants were trained on the voice of the most-familiar talker for only about 1 hr. This duration of training is comparable with everyday situations in which people become familiar with new colleagues or

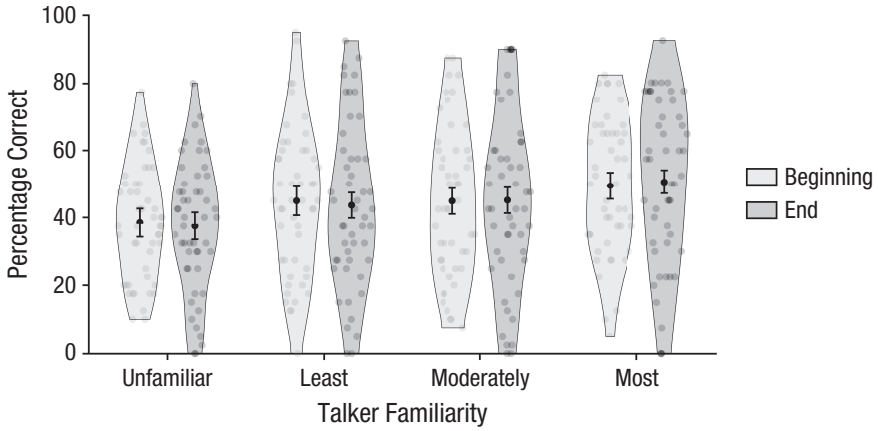

Fig. 4. Percentage of correct responses at the beginning and end of the speech-intelligibility task, separately for each familiarity condition. The shaded region in each plot shows the density of the data; data from individual participants is indicated by transparent dots. The black dots represent means, and error bars show ±1 *SEM*.

with radio and television presenters. The short duration of the training used here suggests that voice training could be an accessible way to improve speech intelligibility in everyday settings. Such benefits may also be relevant for people in occupations that require accurate speech perception when other sounds are present—aircraft pilots, for example. Whether familiar voices are associated with less effort as well as better intelligibility could also be investigated in future work.

These results also contribute to an emerging idea that recognition of familiar voices, and the intelligibility benefit gained from them, are at least partially dissociable. The distinction between intelligibility and recognition touches on a long debate about whether indexical properties of a voice (i.e., consistent aspects of an individual's speech across utterances; see Remez et al., 2007) are separate from properties that convey speech content (i.e., the words that were spoken). It was once thought that speech was stripped of indexical information in order to understand lexical content (e.g., Abercrombie, 1967; Bricker & Pruzansky, 1976; Liberman et al., 1967). However, the fact that speech is more intelligible when it is spoken by familiar people (e.g., Nygaard & Pisoni, 1998; Nygaard et al., 1994; Remez et al., 1997) demonstrates that indexical properties of speech are used to access speech content. In addition, a functional MRI study demonstrated distinct neural activations, depending on whether participants attended to the identity of a talker or to the words they spoke (von Kriegstein et al., 2005). Fewer studies have explored whether people use familiar-voice information differently depending on whether the goal is to recognize someone's identity or to understand the words they are speaking. One previous study (Holmes et al., 2018) provided evidence for a dissociation by showing that recognition and intelligibility of familiar voices are differentially sensitive to acoustic characteristics: Some participants gained an intelligibility benefit for a friend's voice after its vocal-tract length had been manipulated so it was not explicitly recognizable. Here, we provide new evidence for this dissociation by showing that recognition and intelligibility improve at different rates as voices become familiar through training.

The dissociation between intelligibility and recognition is unlikely to be fully explained by differences in difficulty. Although recognition of all three trained talkers was moderately high and did not differ across voices, these results cannot be explained by a ceiling effect: The average d' was 2.2, whereas the maximum attainable d' was 4.3, and the percentage of correct responses was also below ceiling at 73%. Thus, recognition could have differed among conditions but did not. Although average performance on the recognition test (73%) was better than average performance on the intelligibility test, it was not substantially better than intelligibility at the most-favorable TMR (+3 dB TMR in the most-familiar condition: 64%)—and the difference in intelligibility between the most-familiar and moderately familiar conditions was more distinct at the most-favorable TMR. In contrast, there was no evident trend toward better explicit recognition of the most-familiar or moderately familiar talker compared with the least-familiar talker.

We do not find it surprising that the explicit-recognition test showed performance below ceiling, despite participants' near-perfect performance during training. Training was a three-alternative forced-choice task with a chance rate of 33%, whereas the chance rate for the explicit-recognition task was lower than this. During

familiarization and training, participants may have learned characteristics that distinguished the three trained voices from each other, but those characteristics may not have enabled them to distinguish these voices from novel voices in the larger (five-talker) pool of voices. Further, the sentences presented during the recognition test were from a different corpus than those presented during training. Participants may have become very good at identifying a talker's voice in naturalistic open-set sentences, but this learning did not fully transfer to the closed-set BUG sentences. For these reasons, we do not consider the difference between training and test performance to be an interesting or important feature of our results.

Participants performed the explicit-recognition test before the intelligibility test, but they were exposed to the three trained voices equally often in both tests, so this exposure cannot account for differential effects in intelligibility among the three trained voices. Also, exposure to the novel voices in the explicit-recognition test should only reduce the magnitude of the familiar-voice benefit (not improve it). Thus, it could only work against the effect we were trying to measure, which we found to be significant for all three trained talkers. In addition, if additional voice exposure affected performance in the intelligibility test, then we would expect to find a different pattern of results between the first and last trials of the intelligibility test, but we found no evidence for this.

Recognition and intelligibility were similar regardless of whether participants were trained to recognize the talkers in quiet or in simultaneously presented babble noise—reinforcing the idea that similar training and test conditions are unnecessary for familiar-voice learning (Case et al., 2018). These results are also consistent with the conclusion mentioned above: that intelligibility is enhanced when a voice is learned, even when intelligibility is not challenged during learning, as all of our participants found it easy to report the sentences in quiet.

We selected quiet and babble for the training conditions because they both differed from a single competing talker, which is the masker we presented during the intelligibility test. A drawback of using exactly the same masker (i.e., a single competing talker) during testing and training is that it is difficult to determine whether any training benefit arises because of similarity between training and testing conditions or because of practice effects with highly similar stimuli during training and testing, which would not generalize to other conditions. Of the two training conditions we tested, babble is more similar to a single competing talker because it provides both informational and energetic masking, albeit less informational masking than a single competing talker (Brungart, 2001; Brungart et al., 2001). Our data provide no evidence for the encoding-specificity hypothesis (Tulving & Thomson, 1973) or for the idea that the additional challenge associated with listening in babble during training improves recognition or intelligibility—otherwise, we should have found better intelligibility and recognition for participants who were trained in babble than for those who were trained in quiet. It is possible that our training-in-babble noise—at 0 dB TMR—was not sufficiently challenging to promote enhanced learning, although detrimental effects of noise on recognition memory have been found at similar TMRs and at more positive TMRs in previous studies (Koeritzer et al., 2018; Rabbitt, 1968). Nevertheless, these findings demonstrate that the benefit to intelligibility is relatively robust and is not overly sensitive to voice-training conditions.

The training task required participants to identify which of three talkers spoke on each trial. There was no difference in accuracy for participants trained in quiet and those trained in background babble. This suggests that performance during the explicit-recognition task, when participants had to say whether or not a given voice had been heard during training, would not have been affected if babble had been added. This is consistent with findings of previous studies showing that familiar voices are still recognizable in the presence of noise (Best et al., 2018; Clarke et al., 1966; Wenndt & Mitchell, 2012).

In summary, we showed that a relatively small amount of training (~10 min) is sufficient for listeners to identify someone's identity from their voice and to realize a benefit to intelligibility when a competing talker is present. Nevertheless, we found that recognition and intelligibility develop over different timescales. Recognition was similar for the three voices participants were trained on for different lengths of time (~10–60 min), but intelligibility was best for the voice participants were trained on the most (~1 hr). Overall, our results demonstrate the great potential of voice training for improving intelligibility in everyday settings.

Transparency

Action Editor: M. Natasha Rajah
Editor: Patricia J. Bauer

Author Contributions

E. Holmes, G. To, and I. S. Johnsrude designed the study. G. To recorded the stimuli and collected the data. E. Holmes analyzed the data and drafted the manuscript. G. To and I. S. Johnsrude helped edit the manuscript. All of the authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by the Canadian Institutes of Health Research (Operating Grant No. MOP 133450) and the Natural Sciences and Engineering Research Council of Canada (Discovery Grant No. 327429-2012).

Open Practices

Data for this study have been made publicly available via OSF and can be accessed at https://osf.io/2gaem. Stimuli (voice recordings) cannot be shared publicly because of concerns about confidentiality. The design and analysis plans for the study were not preregistered. This article has received the badge for Open Data. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD

Emma Holmes https://orcid.org/0000-0002-0314-6588

Acknowledgments

We thank Brian Gygi for sharing the babble sound file that we used for this study.

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797621991137

Note

1. It is not uncommon for ω_p^2 to be negative, because it corrects for bias. However, effect sizes cannot be less than 0. In this article, we report ω_p^2 and associated CIs from MOTE, which outputs CIs between 0 and 1.

References

Abercrombie, D. (1967). *Elements of general phonetics*. Aldine.

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

Best, V., Ahlstrom, J. B., Mason, C. R., Roverud, E., Perrachione, T. K., Kidd, G., Jr., & Dubno, J. R. (2018). Talker identification: Effects of masking, hearing loss, and age. *The Journal of the Acoustical Society of America*, *143*(2), 1085–1092. https://doi.org/10.1121/1.5024333

Bricker, P. D., and Pruzansky, S. (1976). Speaker recognition. In N. J. Lass (Ed.), *Contemporary Issues in Experimental Phonetics* (pp. 295–326). Academic Press.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101–1109. https://doi.org/10.1121/1.1345696

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*(5), 2527–2538. https://doi.org/10.1121/1.1408946

Buchanan, E. M., Padfield, W. E., Van Nuland, A., Wikowsky, A., & Gillenwaters, A. (2018). *MOTE: The Shiny app to calculate effect sizes and their confidence intervals*. https://osf.io/tds83

Case, J., Seyfarth, S., & Levi, S. V. (2018). Short-term implicit voice-learning leads to a Familiar Talker Advantage: The role of encoding specificity. *The Journal of the Acoustical Society of America*, *144*(6), EL497–EL502. https://doi.org/10.1121/1.5081469

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229

Clarke, F. R., Becker, R. W., & Nixon, J. C. (1966). Characteristics that determine speaker recognition. ESD-TR-66-636. *Technical Documentary Report. United States. Air Force. Systems Command. Electronic Systems Division*.

Doddington, G. R. (1985). Speaker recognition—identifying people by their voices. *Proceedings of the IEEE*, *73*(11), 1651–1664. https://doi.org/10.1109/PROC.1985.13345

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

Dubno, J. R., Dirks, D. D., & Morgan, D. E. (1984). Effects of age and mild hearing loss on speech recognition in noise. *The Journal of the Acoustical Society of America*, *76*(1), 87–96. https://doi.org/10.1121/1.391011

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

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

Holmes, E. (2018). *Speech recording videos* (Version 1.0.0) [Computer code]. Zenodo. https://doi.org/10.5281/zenodo.1165402

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

Holmes, E., & Johnsrude, I. S. (2020). Speech spoken by familiar people is more resistant to interference by linguistically similar speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(8), 1465–1476. https://doi.org/10.1037/xlm0000823

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

Kidd, G., Jr., Best, V., & Mason, C. R. (2008). Listening to every other word: Examining the strength of linkage variables in forming streams of speech. *The Journal of the Acoustical Society of America*, *124*(6), 3793–3802. https://doi.org/10.1121/1.2998980

Koeritzer, M. A., Rogers, C. S., Van Engen, K. J., & Peelle, J. E. (2018). The impact of age, background noise, semantic ambiguity, and hearing loss on recognition memory for spoken sentences. *Journal of Speech, Language, and Hearing Research*, *61*(3), 740–751. https://doi.org/10.1044/2017_JSLHR-H-17-0077

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

Levi, S., Winters, S., & Pisoni, D. B. (2008). A cross-language familiar talker advantage? *The Journal of the Acoustical Society of America*, *123*(5), 3331. https://doi.org/10.1121/1.2933847

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes*, *27*(7–8), 953–978. https://doi.org/10.1080/01690965.2012.705006

Mattys, S. L., White, L., & Melhorn, J. F. (2005). Integration of multiple speech segmentation cues: A hierarchical framework. *Journal of Experimental Psychology: General*, *134*(4), 477–500. https://doi.org/10.1037/0096-3445.134.4.477

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial in Quantitative Methods for Psychology*, *4*(2), 61–64.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

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

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Rabbitt, P. M. A. (1968). Channel-capacity, intelligibility and immediate memory. *The Quarterly Journal of Experimental Psychology*, *20*(3), 241–248. https://doi.org/10.1080/14640746808400158

Remez, R. E., Fellowes, J. M., & Nagel, D. S. (2007). On the perception of similarity among talkers. *The Journal of the Acoustical Society of America*, *122*(6), 3688–3696. https://doi.org/10.1121/1.2799903

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

Rodd, J. M., Davis, M. H., & Johnsrude, I. S. (2005). The neural mechanisms of speech comprehension: fMRI studies of semantic ambiguity. *Cerebral Cortex*, *15*(8), 1261–1269. https://doi.org/10.1093/cercor/bhi009

Sheffert, S. M., Pisoni, D. B., Fellowes, J. M., & Remez, R. E. (2002). Learning to recognize talkers from natural, sinewave, and reversed speech samples. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(6), 1447–1469. https://doi.org/10.1037/0096-1523.28.6.1447

Sommers, M. S., Kirk, K. I., & Pisoni, D. B. (1997). Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format. *Ear and Hearing*, *18*(2), 89–99. https://doi.org/10.1097/00003446-199704000-00001

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

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352–373. https://doi.org/10.1037/h0020071

von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. *Journal of Cognitive Neuroscience*, *17*(3), 367–376. https://doi.org/10.1162/0898929053279577

Wenndt, S. J., & Mitchell, R. L. (2012, March 25–30). *Familiar speaker recognition* [Conference session]. 37th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Kyoto, Japan.

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

Zheng, Z. Z., Vicente-Grabovetsky, A., MacDonald, E. N., Munhall, K. G., Cusack, R., & Johnsrude, I. S. (2013). Multivoxel patterns reveal functionally differentiated networks underlying auditory feedback processing of speech. *The Journal of Neuroscience*, *33*(10), 4339–4348. https://doi.org/10.1523/JNEUROSCI.6319-11.2013