Familiar voices are more intelligible, even if they are not recognized as familiar

Holmes, E.[1], Domingo, Y.[1], & Johnsrude, I. S.[1,2]

[1] Brain and Mind Institute, University of Western Ontario, Canada

[2] School of Communication Sciences and Disorders, University of Western Ontario, Canada

Corresponding author: Emma Holmes; E-mail: emma.holmes@ucl.ac.uk; Phone: +44 7597 967397; Mailing address: Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London, 12 Queen Square, London WC1N 3BG, U.K.

Abstract

We can recognize familiar people by their voices, and familiar talkers are more intelligible than unfamiliar talkers when competing talkers are present. However, whether the acoustic voice characteristics that permit recognition and those that benefit intelligibility are the same or different is unknown. Here, we recruited pairs of participants who had known each other for 6 months or longer, and manipulated the acoustic correlates of two voice characteristics (vocal tract length and glottal pulse rate). These had different effects on explicit recognition of, and the speech-intelligibility benefit realized from, familiar voices. Furthermore, even when explicit recognition of familiar voices was eliminated, they were still more intelligible than unfamiliar voices—demonstrating that familiar voices do not need to be explicitly recognized to benefit intelligibility. Processing familiar-voice information appears therefore to depend on multiple, at least partially independent, systems that are recruited depending on the perceptual goal of the listener.

Introduction

When we converse with other people, we become familiar with their voices, and this enables us to subsequently recognize those people by voice. Historically, the components of speech that convey talker-identity information (‘the carrier’) were considered separately from those that convey the spoken message (‘the content’; Halle, 1985; Joos, 1948). Indeed, brain activity differs when participants attend to speech content or the speaker’s identity (von Kriegstein, Kleinschmidt, Sterzer, & Giraud, 2005), showing that information about the carrier is encoded at least partially separately from the content. Intriguingly, however, familiar-voice information can aid intelligibility of degraded speech content. In the presence of a competing talker, listeners find speech more intelligible if it is spoken by a familiar than unfamiliar talker (Domingo, Holmes, & Johnsrude, submitted; Johnsrude et al., 2013; Kreitewolf, Mathias, & von Kriegstein, 2017; Levi, Winters, & Pisoni, 2011; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994; Yonan & Sommers, 2000). Thus, experience with a carrier aids in identification of content. However, the acoustic characteristics that underlie the benefit to speech intelligibility from a familiar voice—and whether they are the same as those that are critical for recognizing a voice as familiar—are currently unknown.

Speech spoken by different talkers varies on several dimensions. The source-filter model of speech production (Fant, 1960; Chiba & Kajiyama, 1941) assumes that the acoustics of speech result from the action of the articulatory filter upon the vocal source, which is created through vocal-fold vibration. The rate of vocal-fold vibration (which is also known as the glottal pulse rate) is related to the mass of the vocal folds. The rate of vibration determines the fundamental frequency (f_0) of the speech signal. This source is dynamically filtered by the vocal tract, which differs in length and shape between different talkers. These properties of the vocal tract determine the resonances, or formants, of speech, which are frequency-specific concentrations of sound energy. Both f_0 and formant spacing are somewhat variable within talkers. Although vocal-tract characteristics are relatively fixed within a talker, the shape of the vocal cavity changes when talkers alter the positions of the articulators (e.g., lips and tongue) to create different sounds (e.g., Hillenbrand, Getty, Clark, & Wheeler, 1995). The length of the vocal tract also changes the location (spacing) of the formants in lawful ways (Turner et al., 2009). The length and tension of the vocal folds can be controlled by the talker; for example, f_0 contour differs between statements and questions (Eady & Cooper, 1986) and instantaneous f_0 fluctuates throughout a sentence when a talker speaks emotively (Bänziger & Scherer, 2005). Nevertheless, average f_0 and formant spacing both differ reliably between different people, due to physical constraints, and are informative about the gender (Titze, 1989) and size (Smith et al., 2005) of a talker.

These two cues (f_0 and formant spacing) also contribute to listeners' judgements of talker identity. They both influence the perceived similarity of unfamiliar talkers (f_0: Baumann & Belin, 2009; Gaudrain, Li, Ban, & Patterson, 2009; Matsumoto, Hiki, Sone, & Nimura, 1973; Murry & Singh, 1980; Walden, Montgomery, Gibeily, Prosek, & Schwartz, 1978; formant spacing: Baumann & Belin, 2009; Gaudrain et al., 2009; Matsumoto et al., 1973; Murry & Singh, 1980). In addition, they allow listeners to recognize familiar people from their voices ($f_{0:}$ Abberton & Fourcin, 1978; LaRiviere, 1975; Lavner, Gath, & Rosenhouse, 2000; Lavner, Rosenhouse, & Gath, 2001; van Dommelen, 1987, 1990; formant spacing: LaRiviere, 1975; Lavner et al., 2000, 2001). Lavner et al. (2000) found that changing formant positions or f_0 reduced familiar-talker recognition, but recognition was more greatly affected by changes to formant positions than by changes to f_0—thus suggesting that vocal tract features contribute more than glottal source features to familiar-talker recognition. This previous work is specific to the acoustic cues that allow listeners to recognize talkers as familiar; the acoustic cues that allow listeners to find familiar voices more intelligible have not been explored. Given that brain activity differs when participants attend to speech content or the speaker's identity (von

Kriegstein et al., 2005), it seems plausible that the acoustic cues that underlie the speech-intelligibility benefit for familiar voices may be different to those underlying recognition.

We recruited pairs of participants who had known each other for 6 months or longer. We used a closed-set (rather than open-set) task to assess speech intelligibility, so that differences between familiar and unfamiliar voice conditions could not be attributed to a difference in the tendency to guess when uncertain. Each participant recorded sentences from the “BUG” speech corpus (Kidd, Best, & Mason, 2008), where every sentence is of the form ““<Name> <verb> <number> <adjective> <noun>” (e.g., “Bob bought five green bags”). We investigated whether manipulating the acoustic correlates of glottal pulse rate (i.e., f_0) or of vocal tract length (VTL; i.e. formant spacing) reduced the ability to recognise the voice as familiar and/or the speech-intelligibility benefit gained from a familiar compared to unfamiliar target talker in the presence of a competing talker.

Methods

Participants

We recruited 11 pairs of participants (7 male, 15 female) who had known each other for 0.5–9.0 years (median = 2.0 years, interquartile range = 1.5) and who spoke regularly (> 5 hours per week). Pairs of participants were friends or couples. Seven were opposite-sex pairs and three were same-sex (female-female) pairs. Twenty-one participants completed the entire experiment. This sample size is sufficient to detect within-subjects effects of size f = 0.41 with 0.95 power (Faul et al., 2007); Johnsrude et al. (2013) reported a familiar-talker benefit to speech intelligibility of size f = 0.72, which should be detectable with the current sample. The 21 participants were aged 19–24 years (median = 22.5 years, interquartile range = 2.6) and were native Canadian English speakers who reported no history of hearing difficulty. Participants had average pure-tone hearing levels of 15 dB HL or better in each ear (at four octave frequencies between 0.5 and 4 kHz). The experiment was cleared by Western University's Health Sciences Research Ethics Board. Informed consent was obtained from all participants.

Apparatus

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

Acoustic stimuli were recorded using a Sennheiser e845-S microphone connected to a Steinberg UR22 sound card (Steinberg Media Technologies). During the listening tasks, acoustic stimuli were presented through the Steinberg UR22 sound card (Steinberg Media Technologies) and were delivered binaurally through Grado Labs SR225 headphones.

Stimuli

Each participant recorded 480 sentences from the Boston University Gerald (BUG) corpus (Kidd et al., 2008), which follow the structure: "<Name> <verb> <number> <adjective> <noun>". In the sub-set used in the experiment, there were two names ('Bob' and 'Pat'), eight verbs ('bought', 'found', 'gave', 'held', 'lost', 'saw', 'sold', 'took'), eight numbers ('two', 'three', four', 'five', 'six', 'eight', 'nine', 'ten'), eight adjectives ('big', 'blue', 'cold', 'hot', 'new', 'old', 'red', 'small'), and eight nouns ('bags', 'cards', 'gloves', 'hats', pens', 'shoes', 'socks', 'toys'). An example is "Bob bought three blue bags". To ensure that all sentences were spoken at similar rates—and thus the five words from two different sentences would overlap when used in the speech intelligibility task—we played videos indicating the desired pace for each sentence (Holmes, 2018) while participants completed the recordings. The sentences had an average duration of 2.5 seconds (s = 0.3). The levels of the digital recordings of the sentences were normalised to the same root mean square (RMS) power.

Sentences were processed using the 'Change Gender' function in Praat (Boersma & Weenink, 2013). Fundamental frequency (f_0) was changed by shifting the 'median pitch' of the sentence upwards. Changes in vocal tract length (VTL) were simulated by shifting the frequencies of the formants upwards by a percentage, which also increases their spacing. We created ‘unshifted’ versions by shifting the median pitch and formants upwards, then downwards again by the same amount, to restore the median pitch and formant positions of the original sentence. The reason for creating ‘unshifted’ versions was to preserve any distortions introduced by the signal processing, but maintain the original f_0 and formant values.

We aimed to manipulate f_0 and VTL by approximately the same perceptual amount, so that any differences in the extent to which the two attributes influenced task performance was not due to differences in perceptual discriminability of the two cues. To this aim, we estimated listeners’ thresholds for discriminating f_0 and VTL and used a multiple of this just-noticeable-difference threshold in the main experiment. We wanted to make the manipulations large, so we multiplied the median threshold (across participants) by 5, which was the largest manipulation possible before the sentences became distorted by the signal processing algorithm. We estimated the thresholds for discriminating changes to f_0 and VTL in a group of 5 participants who did not take part in the main experiment. These participants performed a two-alternative forced-choice (2AFC) task with a weighted (9:1) up-down adaptive procedure (Kaernbach, 1991) that estimated the 90% threshold for discriminating f_0 and VTL manipulations of the familiar voice (i.e., the participant’s partner’s voice). On each trial, participants heard three different sentences spoken by their partner’s voice, presented sequentially. The first sentence was presented with the original f_0 and VTL (unshifted version). Either the second or third sentence was the manipulated version and the remaining sentence was unshifted, like the first sentence. Participants indicated whether the second or third sentence was manipulated. We used separate, but interleaved, runs for f_0 and VTL, each with a starting manipulation value of 1.15% above the original recording. The procedure stopped after 8 reversals and threshold values were calculated as the median of the last 5 reversals (f_0: 8.05%; VTL: 5.35%). We set the manipulation magnitude at five times the median threshold from the group of 5 participants, which produced stimuli with median pitches (corresponding to f_0) that were 40.25% higher than that of the original sentences and sentences with formant frequencies (corresponding to VTL) that were 26.75% higher than those of the original sentences. We refer to these stimuli as f_0-manipulated and VTL-manipulated stimuli, respectively. We created 'both-manipulated' sentences by shifting median pitch by 40.25% and formants by 26.75%.

During the experiment, each participant heard sentences spoken by their familiar partner and sentences spoken by two unfamiliar talkers, who were the partners of other participants in the experiment, sex matched to the familiar talker. The advantage of this aspect of the design was that acoustic stimuli were counterbalanced across the familiar and unfamiliar voice conditions; so that, across the group, these two types of condition were acoustically as similar as possible. Each voice was presented to one participant (i.e. their partner) as a familiar talker and to two other participants as an unfamiliar talker. The only exception was the participant whose partner did not complete the experiment. This voice was presented as unfamiliar twice, but never as familiar. For the same reason, two other voices were presented once as familiar and only once as unfamiliar.

Procedure

Participants completed two tasks: a speech intelligibility task and an explicit recognition task. Half completed the speech intelligibility task first and the other half completed the explicit recognition task first. Each task included three voice-manipulation conditions: (1) the original f_0 and VTL were preserved (unshifted condition), (2) f_0 was manipulated (f_0-manipulated condition), (3) VTL was manipulated (VTL-manipulated condition), and (4) f_0 and VTL were both manipulated in combination (both-manipulated condition).

In the speech intelligibility task, participants heard two sentences spoken simultaneously by different talkers. They identified the four remaining words of a sentence that began with a particular target name ("Bob" or "Pat"), by clicking buttons on a screen. On each trial, either the target sentence was spoken by the participant's partner and the masker sentence was spoken by an unfamiliar talker ("Familiar Target" condition), or both sentences were spoken by unfamiliar talkers ("Both Unfamiliar" condition). The target and masker sentences were always spoken by different talkers but were both manipulated in the same way (i.e. VTL-manipulated, f_0-manipulated, both-manipulated, or unshifted). Target and masker sentences were presented at two different target-to-masker ratios (TMRs): -6 and +3 dB. For all participants, acoustic stimuli were presented at a comfortable listening level (approximately 67 dB(A) SPL), which was roved over a range of 3 dB. All trial types (2 familiarity conditions x 4 manipulation conditions x 2 TMRs) were randomly interleaved. Participants completed 768 trials (i.e., 32 trials in each condition), 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.

In the explicit recognition task, listeners heard one sentence on each trial. The sentence could be spoken by the participant's partner or by one of the two unfamiliar voices. We used the same four voice manipulations as in the speech intelligibility task (VTL-manipulated, f_0-manipulated, both-manipulated, or unshifted). Participants were told that some of the sentences had been manipulated and were instructed to report whether they thought each sentence was spoken by their partner or not, regardless of any manipulation. Participants completed 84 trials (21 for each manipulation condition).

At the end of the experiment, we checked that participants could accurately discriminate between sentences that had been manipulated in f_0 and/or correlates of VTL and sentences in which the original f_0 and correlates of VTL had been preserved. On each trial, participants heard three different sentences spoken by their partner, presented sequentially. On each trial, all three sentences were spoken by either the familiar talker or one of the two unfamiliar talkers. The first sentence was always presented in its 'unshifted' version, as a reference. Of the two remaining sentences, one was the manipulated version and the other was the 'unshifted' version. In a 2AFC task, participants had to indicate whether the second or third sentence had been manipulated. Participants completed 48 trials, with 16 in each of the three manipulation conditions (VTL-manipulated, f_0-manipulated, or both-manipulated).

Analyses

We calculated sensitivity (d') for the explicit-recognition data using loglinear correction (Hautus, 1995), so chance d' is 0.3. For the speech intelligibility task, we calculated the percentage of sentences in which participants reported all four words (after the name) correctly.

To assess the familiar-talker benefit to speech intelligibility, we compared percent correct between the Familiar Target and Both Unfamiliar conditions. In both conditions, participants had to report words from a target sentence in the presence of a masker sentence that was spoken by a different (unfamiliar) talker. The masker voices were identical in the two conditions—the only difference between these two conditions was whether the target sentence was spoken by a familiar talker or by one of the unfamiliar talkers. We also analysed whether performance on the speech intelligibility and explicit recognition tasks were affected by the manipulation condition (VTL-manipulated, f_0-manipulated, both-manipulated, or unshifted).

To assess whether there was a relationship between recognition performance and speech-intelligibility benefit (e.g. to assess whether there is a greater intelligibility benefit for voices that are better recognized), we calculated Spearman's rank correlation coefficients between performance in the explicit recognition task and the magnitude of the speech-intelligibility benefit for the familiar voice (i.e., the difference in percent correct between the Familiar Target and Both Unfamiliar conditions). We did this separately for each manipulation condition.

Results

Results from the manipulation discrimination task showed that participants could discriminate changes in f_0 (mean [$\bar{x}$] = 91.6%, standard deviation [s] = 18.5), VTL ($\bar{x}$ = 95.9%, s = 18.2), and both cues

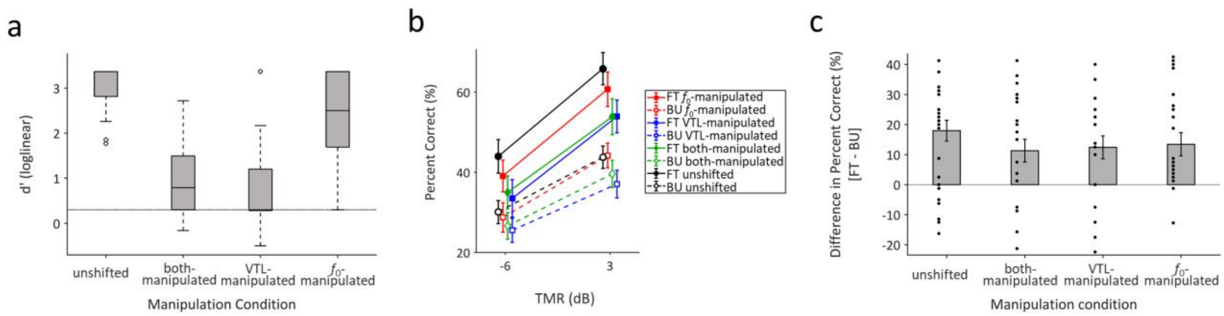

Fig 1. Explicit recognition and speech intelligibility (N=21). (a) Sensitivity (d′) in the Explicit Recognition task. Open circles illustrate data from participants who were outliers. (b) Percent of trials in which participants reported the words from the target sentence correctly in the Speech Intelligibility task (c) 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. Error bars represent ±1 standard error of the mean. Filled circles display results from individual participants. See the Results section for a description of significant differences between conditions. FT = Familiar Target; BU = Both Unfamiliar.

combined ($\bar{x}$ = 94.7, s = 22.3) with high accuracy. One participant achieved below-chance performance (12.5%) on the discrimination task, but performed similarly to the other participants in the explicit recognition and speech intelligibility tasks, so we included this participant in the analyses (excluding this participant did not affect the pattern of results).

Explicit recognition

As shown in Figure 1a, sensitivity (d′) in the explicit recognition task depended strongly on condition. Sensitivity was much lower in VTL-manipulated and both-manipulated conditions than in the unshifted and f_0-manipulated conditions. The d′ data violated the assumption of normality (skewed distributions and p < .05 in Shapiro-Wilk test), so non-parametric tests are reported.

We compared d′ across the four manipulation conditions using Wilcoxon signed-rank tests. Participants were significantly better at recognizing their partner's voice in the unshifted condition compared to all others ($Z \geq 2.67$, $p \leq .008$). They were also better in the f_0-manipulated condition than in both conditions in which VTL was manipulated ($Z \geq 3.62$, $p < .001$). Sensitivity (d′) did not differ between the two conditions in which VTL was manipulated (VTL-manipulated and both-manipulated; $Z = .71$, $p = .48$).

Sign tests, evaluating d′ scores against chance level (0.3), showed that participants were unable to recognize their partner's voice (i.e., chance sensitivity) in the two VTL-manipulated conditions (VTL-manipulated: $S = 8$, $p = .38$; both-manipulated: $S = 13$, $p =.38$) but were significantly better than chance in the unshifted ($S = 21$, $p <.001$) and f_0-manipulated ($S = 18$, $p =.001$) conditions.

To investigate whether the manipulations affected recognition differently for male and female voices we conducted a 2x4 Mixed ANOVA (Sex x Manipulation). We found no main effect of voice sex [$F(1, 19) = 1.13$, $p = .30$, $\omega = .01$] and no significant interaction between Sex and Manipulation condition [$F(1, 19) = .26$, $p = .62$, $\omega = -.04$].

Speech intelligibility

Baseline performance in the Both Unfamiliar condition was similar across the four manipulation conditions (Figure 1b). Therefore, for each manipulation, we calculated the familiar-voice speech-intelligibility benefit by subtracting percent correct in the Both Unfamiliar condition from percent correct in the Familiar Target condition.

The data met the assumptions of normality, as assessed by the Shapiro-Wilk test and by observing box-plots and Q-Q plots. We analyzed the data using a two-way within-subjects ANOVA with the factors Manipulation (unshifted, f_0-manipulated, VTL-manipulated, both-manipulated) and TMR (-6, +3). The main effect of Manipulation was significant [$F(3, 60) = 3.69$, $p = .017$, $\omega = .11$]. Planned comparisons showed that the familiar-voice benefit in the unshifted condition was significantly larger than in all other conditions ($p \leq .036$). The familiar-voice benefit did not differ significantly between any of the other conditions ($p \geq .31$). Participants received a significantly greater familiar-voice benefit at +3 dB TMR ($\bar{x}$ = 10.1, s = 13.7) than at -6 dB TMR ($\bar{x}$ = 17.4, s = 19.6) [$F(1, 20) = 9.17$, $p = .007$, $\omega = .27$]. The interaction between Manipulation and TMR was not significant [$F(3, 60) = .24$, $p = .87$, $\omega = -.04$].

Figure 1c illustrates the familiar-voice benefit to speech intelligibility across the four manipulations, collapsed across TMRs. One-sample t-tests for each manipulation showed that the familiar-voice benefit was significantly greater than zero in all four conditions ($p \leq .007$).

We split the data by whether the voices were male or female and conducted a 2x4 (Sex x Manipulation) Mixed ANOVA on the magnitude of the speech-intelligibility benefit for the familiar voice. There was no main effect of voice sex [$F(1, 19) = 1.65$, $p = .21$, $\omega = .03$] and no significant interaction between Sex and Manipulation [$F(1, 19) = 1.92$, $p = .18$, $\omega = .04$].

Voice manipulations affected recognition and intelligibility differently

There was no significant relationship between recognition performance and the speech-intelligibility benefit for any of the four manipulations ($r \leq .34$, $p \geq .13$). Thus, speech-intelligibility benefit for a familiar voice does not appear to relate to the ability to explicitly recognize that person from their voice.

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 percent improvement in speech intelligibility from the familiar talker into *z*-scores and entered the data into a 2-way within-subjects ANOVA. We tested the two-way interaction between Task (speech intelligibility and explicit recognition) and Manipulation (unshifted, f_0-manipulated, VTL-manipulated, and both-manipulated). The interaction was significant [$F(3, 60) = 35.35$, $p < .001$, $\omega = .62$], confirming that the pattern across manipulations indeed differed between the two tasks.

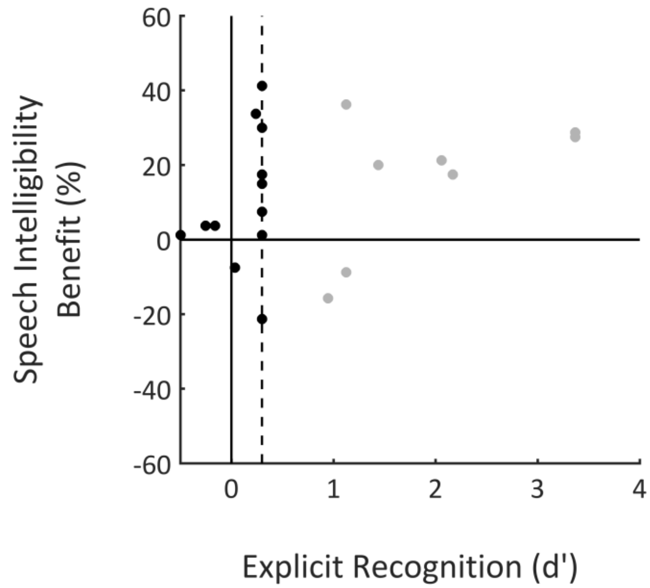

Fig 2. VTL-manipulated condition: Relationship between explicit recognition d′ and the magnitude of the speech-intelligibility benefit for the familiar voice (i.e., Familiar Target – Both Unfamiliar). The vertical dashed line indicates chance performance (d′ = 0.3) in the explicit recognition task. Each point illustrates one participant. Points that are coloured in black represent participants who scored at or below chance level in the explicit recognition task for the VTL-manipulated condition.

To further examine whether participants were able to gain a speech-intelligibility benefit from distorted voices that they were not able to explicitly recognize, we selected a sub-set of participants (N = 13) whose sensitivity was at or below chance (d′ ≤ 0.3) in the VTL-manipulated condition of the explicit recognition task (Figure 2). We performed a sign test for these 13 participants to determine whether the speech-intelligibility benefit for the VTL-manipulated familiar voice differed from zero. Indeed, these participants gained a speech-intelligibility benefit for the VTL-manipulated familiar voice that was significantly greater than zero (median = 7.50%,

$S = 11$, $p = .022$). This result demonstrates that participants are able to gain a speech-intelligibility benefit from a distorted familiar voice, even when they are not able to explicitly recognize that voice as familiar.

Discussion

When the acoustic correlates of VTL were manipulated (27% shift in formant frequencies), participants could no longer recognize a familiar voice, but still found it more intelligible than sex-matched unfamiliar voices. In contrast, when f_0 was manipulated (shifted by 40%) participants could still recognize the familiar voice as well as finding it more intelligible. Importantly, the patterns of results for these two manipulations differed significantly from each other, to the point that participants who were unable to recognize the VTL-modified familiar voice still found it more intelligible than unfamiliar voices. Thus, the two abilities rely on (at least partially) distinct cognitive (and possibly neural) substrates. If you are using voice acoustics to recognize someone you know, VTL information seems to be much more important than pitch information. If, however, you are using voice acoustics to understand a familiar talker better, pitch and VTL information play a partial role, but neither are critical.

In the face-recognition literature, a distinction has been drawn between identity and expression processing (for a review, see Calder & Young, 2005). Patients with prosopagnosia are able to identify emotional expressions in faces, despite impaired recognition of facial identity (Humphreys et al., 1993). Similarly, patient studies have revealed a double dissociation between voice-identity processing and speech processing (e.g., Van Lancker & Canter, 1982).

The ‘auditory face’ model (Belin et al., 2004), which is based on an influential model of face perception (Bruce & Young, 1986), has been used to describe voice perception. This model suggests that voice perception is multi-dimensional, with different systems specialised for identity, speech recognition and emotional expression identification. The dissociation between explicit recognition and the speech-intelligibility benefit in the current study is intriguing, because it predicts that patients who are impaired in their ability to recognize voices might still find familiar voices more intelligible when they are masked by a competing talker. Our results are consistent with the idea that familiar-voice information may feed into (at least partially) separate voice recognition and speech analysis systems.

The acoustic correlates of VTL appear to be critical for explicit recognition, whereas f_0 contributes to a lesser extent. This finding is consistent with the results of other studies that compared the contributions of f_0 and VTL to explicit recognition (Lavner et al., 2000; Gaudrain et al., 2009). The current results extend those previous findings by showing that the greater influence of acoustic correlates of VTL on voice recognition cannot be explained by differences in perceptual discriminability of the two sets of acoustic features. We approximately equated the discriminability of the manipulations by selecting manipulation magnitudes from discrimination (just-noticeable difference) thresholds in a separate group of participants. Thus, we conclude that recognition of a voice as familiar is more robust to perceived differences in f_0 than to perceived differences in correlates of VTL. Gaudrain et al. (2009) speculate that greater within-talker variation in f_0 than VTL could explain the smaller contribution of f_0 to talker recognition. Here, the average within-talker variability was 39.30% (s = 21.19) for f_0 and 0.39% (s = 0.06) for formant spacing. The majority (N = 12) of the talkers had f_0 ranges less than our f_0 manipulation of 40.25%, whereas all had formant spacing ranges substantially less than our formant manipulation of 26.75%. Thus, based on our recorded sentences, it seems plausible that differences in within-talker variability explains the greater effect of the VTL than the f_0 manipulation on recognition.

Although the VTL manipulation eliminated the ability to recognize a voice as familiar, it did not eliminate the ability to gain a speech-intelligibility benefit from the familiar voice. Manipulating f_0 and acoustic correlates of VTL decreased speech intelligibility (compared to the unshifted condition) similarly. There was no additional decrement when both cues were manipulated together compared to when f_0 or VTL were manipulated alone. It is important for the interpretation of our results that speech intelligibility in the Both Unfamiliar condition was similar across the manipulations (see Figure 1b), meaning that the baselines used to calculate the familiar-voice benefit were at a similar place on the psychometric function for all manipulation conditions. Thus, the difference in the familiar-target benefit to intelligibility is real, rather than an artifact of differences in baseline performance.

The manipulations we used were as large as we could impose without distorting the recordings, and were almost as large as the average difference between male and female voices (Titze, 1989). Given that even these manipulations failed to eradicate the intelligibility difference, listeners must rely on acoustic information other than average f_0 and the formant ratio to better understand speech spoken by a familiar talker when a competing talker is present. For example, f_0 contour, formant patterns, harmonic-to-noise ratio, intonation, and rhythm might be important for the familiar-talker benefit to intelligibility. However, the same cues were present in the VTL-manipulated stimuli in the explicit recognition task, and participants performed at chance. Therefore, these cues are not sufficient for recognizing a voice as familiar.

In a separate group of participants (N = 18), we repeated the experiment using smaller manipulations of f_0 and acoustic correlates of VTL. For each listener, we manipulated f_0 and acoustic correlates of VTL at the listener's 90% threshold for discriminating manipulations to those cues (i.e., manipulations were shifts of one just-noticeable difference unit, not five; the range of thresholds were 1.7–6.3% for VTL and 3.9–9.9% for f_0). Although these manipulations were perceptually discriminable (by definition), we found no effect of the manipulations on the ability to recognize the voice as familiar or on the magnitude of the speech-intelligibility benefit for the familiar voice. This result demonstrates that larger deviations to a familiar voice are required to reduce explicit recognition and the speech-intelligibility benefit for familiar voices.

Across both experiments, we replicated the familiar-voice benefit to speech intelligibility (Domingo et al., submitted; Johnsrude et al., 2013; Kreitewolf et al., 2017; Levi et al., 2011; Nygaard & Pisoni, 1998; Nygaard et al., 1994; Yonan & Sommers, 2000) when the original f_0

and information about the original VTL of the familiar voice was preserved. The familiar-voice intelligibility benefit is similar in magnitude in the current experiments (10–25%) as Johnsrude et al. (2013) found for spouses' voices (10–20%), which is consistent with recent data indicating that even 6 months of experience with a friend or partner's voice is sufficient to yield a large intelligibility benefit (Domingo et al., submitted).

Overall, our results demonstrate a large improvement in speech intelligibility when participants listened to a friend's voice in the presence of a competing talker than when they listened to a stranger's voice. This benefit was relatively robust to large manipulations of f_0 and acoustic correlates of VTL. Indeed, participants gained an intelligibility benefit from a manipulated familiar voice even when they were no longer able to explicitly recognize that voice as familiar. The findings demonstrate a dissociation between explicit recognition of a familiar voice and the speech-intelligibility benefit gained from a familiar voice in the presence of a competing talker. The findings imply that different mechanisms may be involved in processing familiar-voice information, depending on the context in which the information is used.

Acknowledgements

This work was supported by funding 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 would like to thank Grace To and Shivaani Shanawaz for assisting with data collection.

Author Contributions

E.H. and I.S.J. designed the research. E.H. and Y.D. collected the data. E.H. analysed the data. E.H., Y.D., and I.S.J. wrote the paper.

Declaration of Conflicting Interests

The authors declare no conflicts of interest with respect to the authorship or the publication of this article.

References

Abberton, E., & Fourcin, A. J. (1978). Intonation and speaker identification. *Language and Speech*, *21*(4), 305–318.

Anderson, A. K., & Phelps, E. A. (2000). Expression without recognition: Contributions of the human amygdala to emotional communication. *Psychological Science*, *11*(2), 106–111. https://doi.org/10.1111/1467-9280.00224

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

Baumann, O., & Belin, P. (2009). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research*, *74*(1), 110–120. https://doi.org/10.1007/s00426-008-0185-z

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. Trends in Cognitive Sciences, 8(3), 129–135. http://doi.org/10.1016/j.tics.2004.01.008

Boersma, P., & Weenink, D. (2013) Praat: doing phonetics by computer [Computer program]. Version 5.4.04, retrieved 26 January 2015 from http://www.praat.org/

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology, 77*, 305–327.

Calder, A. J., & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience, 6*(8), 641–651. http://doi.org/10.1038/nrn1724

Chiba, T., & Kajiyama, M. (1941). The vowel, its nature and structure, Tokyo-Kaiseikan Pub Co., Tokyo.

Domingo, Y., Holmes, E., & Johnsrude, I. S. (submitted). The benefit to speech intelligibility of hearing a familiar voice.

Eady, S. J., & Cooper, W. E. (1986). Speech intonation and focus location in matched statements and questions. *The Journal of the Acoustical Society of America, 80*(2), 402–15.

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,* 175-191.

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

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

Halle, M. (1985). Speculations about the representation of words in memory. *Phonetic Linguistics*.

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

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

Holmes, E. (2018). Speech Recording Videos (Version v1.0.0). Zenodo. http://doi.org/10.5281/zenodo.1165402

Humphreys, G. W., Donnelly, N., & Riddoch, M. J. (1993). Expression is computed separately from facial identity, and it is computed separately for moving and static faces:

Neuropsychological evidence. *Neuropsychologia*, *31*(2), 173–181. https://doi.org/10.1016/0028-3932(93)90045-2

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

Joos, M. (1948). Acoustic Phonetics. *Language*, *24*(Suppl. 2), 5–136. https://doi.org/10.2307/522229

Kaernbach, C. (1991). Simple adaptive testing with the weighted up-down method. *Perception & Psychophysics, 49*(3), 227–229. https://doi.org/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. *The Journal of the Acoustical Society of America*, *124*(6), 3793–3802. https://doi.org/10.1121/1.2998980

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

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

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

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

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

Acoustical Society of America, 130(6), 4053–62. https://doi.org/10.1121/1.3651816

Matsumoto, H., Hiki, S., Sone, T., & Nimura, T. (1973). Multidimensional representation of personal quality of vowels and its acoustical correlates. *IEEE Transactions on Audio and Electroacoustics*, *21*(5), 428–436. https://doi.org/10.1109/TAU.1973.1162507

Murry, T., & Singh, S. (1980). Multidimensional analysis of male and female voices. *The Journal of the Acoustical Society of America*, *68*(5), 1294–1300. https://doi.org/7440851

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*(I), 42–46.

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America*, *85*(4), 1699–1707. https://doi.org/10.1121/1.397959

Turner, R. E., Walters, T. C., Monaghan, J. J. M., & Patterson, R. D. (2009). A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data. *The Journal of the Acoustical Society of America, 125*(4), 2374–2386. http://doi.org/10.1121/1.3079772

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

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

Van Lancker, D. R., & Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and Cognition*, *1*(2), 185–195. https://doi.org/10.1016/0278-2626(82)90016-1

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

Walden, B. E., Montgomery, A. A., Gibeily, G. J., Prosek, R. A., & Schwartz, D. M. (1978). Correlates of psychological dimensions in talker similarity. *Journal of Speech, Language, and Hearing Research*, *21*, 265–275.

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*(I), 88–99. https://doi.org/10.1037/0882-7974.15.1.88