Highly accurate and robust identity perception from personally familiar voices.

Elise Kanber[1], Nadine Lavan[1,2], & Carolyn McGettigan[1]

[1]Department of Speech, Hearing, and Phonetic Sciences, University College London
[2]Department of Biological and Experimental Psychology, School of Biological and Chemical Sciences
Queen Mary University of London

Word Count (including abstract): 10,935

Correspondence to:
Carolyn McGettigan, Department of Speech, Hearing and Phonetic Sciences, University College London, 2 Wakefield Street, London WC1N 1PF, United Kingdom.

E-mail: c.mcgettigan@ucl.ac.uk

Or

Nadine Lavan, Department of Psychology, School of Biological and Chemical Sciences
Queen Mary University of London, Mile End Road, London E1 4NS, United Kingdom.
E-mail: n.lavan@qmul.ac.uk

Acknowledgements: This work was supported by a Research Leadership Award from the Leverhulme Trust (RL-2016-013) awarded to Carolyn McGettigan. Nadine Lavan is funded by a Sir Henry Wellcome Fellowship (220448/Z/20/Z).

Abstract

Previous research suggests that familiarity with a voice can afford benefits for voice and speech perception. However, even familiar voice perception has been reported to be error-prone, especially in the face of challenges such as reduced verbal cues and acoustic distortions. It has been hypothesised that such findings may arise due to listeners not being "familiar enough" with the voices used in laboratory studies, and thus being inexperienced with their full vocal repertoire. Extending this idea, voice perception based on highly familiar voices – acquired via substantial, naturalistic experience – should therefore be more robust than voice perception from less familiar voices. We investigated this proposal by contrasting voice perception of personally-familiar voices (participants' romantic partners) versus lab-trained voices in challenging experimental tasks. Specifically, we tested how differences in familiarity may affect voice identity perception from non-verbal vocalisations and acoustically-modulated speech. Large benefits for the personally-familiar voice over a less familiar, lab-trained voice were found for identity recognition, with listeners displaying both highly accurate yet more conservative recognition of personally-familiar voices. However, no familiar-voice benefits were found for speech perception in background noise. Our findings suggest that listeners have fine-tuned representations of highly familiar voices that result in more robust and accurate voice recognition despite challenging listening contexts, yet these advantages may not always extend to speech perception. Our study therefore highlights that familiarity is indeed a continuum, with identity perception for personally-familiar voices being highly accurate.

Keywords: voice, representations, personal familiarity, speech in noise, familiarity advantage

Introduction

In order to express intentions and adapt to different audiences or speaking situations, speakers constantly adjust the sound of their voices, such that a speaker will never produce exactly the same sound twice. As a consequence, the same person can in fact sound very different depending on the context (Latinus & Belin, 2011; Lavan, Burton, Scott & McGettigan, 2019), making voice identity recognition a challenging and error-prone process. In addition to having to cope with these highly variable and at times ambiguous vocal signals, listeners may also find themselves in challenging listening situations, such as only hearing brief snippets of a voice and/or hearing it in a noisy environment. Consequently, there are numerous reports that voice recognition can be poor, especially when listeners are not familiar with a voice: Most notably, earwitness memory - that is, situations in which listeners have to recognise a voice after often only brief and incidental exposure - has been highlighted as notoriously unreliable (Smith et al., 2019).

However, identity processing performance becomes more accurate when a voice is familiar to listeners (Latinus & Belin, 2011). Recently, a number of voice identity sorting tasks have been used to compare performance in familiar and unfamiliar listeners, highlighting familiarity advantages in voice identity processing. In these voice sorting tasks, listeners are usually presented with a number of short, naturally-varying voice recordings from two or more identities, which differ in numerous ways including speaking style, recording quality and linguistic content. Listeners are then asked to sort the recordings into clusters according to perceived identities. Familiar listeners tend to perceive close to the veridical number of identities, with relatively few errors occurring. In contrast, unfamiliar listeners tend to systematically fail in “telling people together”, by perceiving variable voice recordings of the same person as many different people. This manifests in unfamiliar listeners making several more clusters than the true number of identities in the task (Lavan, Burston, & Garrido, 2019; Lavan, Burston, et al., 2019, Stevenage, Symons, Fletcher & Coen, 2020). Similar evidence for familiarity advantages have also been reported for speaker discrimination tasks (Lavan, Scott & McGettigan, 2016). These findings have been interpreted as evidence that experience with familiar voices leads to the formation of robust, well-formed representations of those voices. Thus, familiar listeners can accommodate natural variations within the sound of a voice into a unified percept of that person, while unfamiliar listeners are more likely to misperceive such variations as evidence of multiple talkers. Such familiarity advantages are not restricted to identity processing but also emerge in other aspects of voice processing such as speech perception: Here, studies show that in the presence of competing talkers and noisy environments, listeners can understand the speech produced by a familiar person significantly better than the speech of an unfamiliar person (Kreitewolf, Mathias & von Kriegstein, 2017; Johnsrude et al., 2013; Nygaard & Pisoni, 1998).

Despite a general consensus that familiarity is advantageous for voice identity perception, the magnitude and nature of the familiarity advantages appear to vary. For example, Fontaine, Love, and Latinus (2017) found entirely different profiles of results in a voice identity perception task for lab-trained versus celebrity voices – in their study, listeners showed better voice identification performance for sounds created by averaging many exemplars of a voice, but this was only seen for the celebrity voices. Such findings give a first indication that the type and degree of familiarity is likely to have an impact on how well listeners can perceive identity from voices. Thus, differences in task instructions and types, listener characteristics (i.e. differing levels and types of familiarity), and stimulus properties have made it difficult thus far to fully understand when and how familiarity benefits identity perception.

Overall, it is also notable that despite the diversity in methods and stimuli, familiarity has primarily been modelled and tested in previous research by using famous/celebrity or lab-trained voices as stimuli. Familiarity with such lab-trained or famous voices is, however, often limited and constrained to certain contexts. Conversely, outside of laboratory settings, voices are nearly always experienced in rich social contexts, in which shared knowledge, experiences, and memories are formed and built upon - an aspect of personal familiarity that is often dramatically reduced or entirely absent for lab-trained or famous voices. This is not to say that it is impossible to recreate learning conditions in the lab that would result in a deep and robust familiarity, thus approximating personal familiarity. However, in most cases, the experience listeners have had with lab-trained or famous voices will differ qualitatively and quantitatively from personally-familiar voices. We therefore speculate that the type of familiarity (i.e. lab-trained or famous voices) most frequently studied in empirical research on voice identity perception may have led to an underestimation of the extent of human voice recognition capabilities by overlooking evidence at the upper end of familiarity - such as the familiarity we have with the personally-familiar voices of friends and family members.

In the current study, we therefore asked: What perceptual benefits can a highly personally-familiar voice afford a listener, compared to a lab-trained voice? Listeners completed two vocal identity tasks and one speech perception task, all of which included challenges to perception. The first task examined recognition in the presence of only minimal linguistic cues, through presenting brief filler sounds – these are vocalisations that are usually used to bridge a pause or hesitation in spontaneous speech (e.g. “um”, “mm”). The second vocal identity task involved presenting acoustically manipulated speech recordings to assess how listeners’ representations of voices are tuned to these acoustic cues, and the dependence of this tuning on familiarity. The final task examined sentence perception against multi-talker babble to assess familiarity advantages beyond identity perception and replicate the well-documented familiar talker benefit for speech intelligibility. Across all tasks, we predicted that personal familiarity would correspond to better performance - specific predictions for each task are outlined in the relevant sections below. The study design and analyses were preregistered on the Open Science Framework (https://osf.io/utche).

General Methods

Participants

Sixty-four participants in total (32 female, mean age = 27.95 years, SD = 6.50 years, range = 18-40 years) were recruited to take part in the study. The sample included sixteen couples (32 participants, 1 male and 1 female per couple, mean age = 26.31 years, SD = 6.10 years, range = 18-37 years) as well as 32 control participants (16 female, mean age = 29.22 years, SD = 6.66 years, range = 18-40 years). Couples visited the lab for a recording session, then participated in the perceptual tasks via the online testing platform Gorilla.sc (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2018). Control participants completed the perceptual tasks only, also via Gorilla.sc.

The couples had been in romantic relationships for a minimum of six months (mean length of relationship = 63.78 months, SD = 51.49 months, range = 6-204 months) and reported speaking to each other frequently (mean = 34.66 hours per week, range = 4 – 88 hours), thus we could assume these participants were highly familiar with their partner’s voice. One female participant did not complete the perceptual tasks after the initial recording session,

leaving a total of 31 couples group participants. Participants that failed the in-task vigilance checks (i.e. those scoring less than 75% or 6/8 correct per task) were furthermore excluded on a task-by-task basis. As several participants reported in debrief that the vigilance check in the speech perception task was confusing, participants that failed the checks only on that task were not excluded - however, participants who failed the checks in this task plus another were excluded from both affected tasks.

Control participants were recruited and tested to exclude the possibility that any reported effects were introduced by the specific voices used, for example the couples' voices systematically being more distinctive/memorable than the lab-trained voices. Each control participant was sex-matched to a couples group participant, so that each version of the experiment created for a member of a couple was repeated with a corresponding member of the control group. Where an individual participant's data were removed from the couples group, the corresponding control group participant's data were also removed, to maintain a one-to-one match between the voice identities presented to the two groups. We note that in order to minimise data loss through participant exclusion, control participants who failed the in-task vigilance checks (see below) were removed and replaced with new control participants until a full set of 31 usable datasets corresponding to the couples group was obtained.

All participants were native English speakers, had normal or corrected-to-normal vision, and reported no hearing difficulties. Couples were all speakers of Standard Southern British English (SSBE) such that accent would be controlled across all the voices used in the studies. Participants were recruited via the UCL Psychology Subject Pool and social media. On completion of the tasks, participants were compensated at a rate of £7.50/hr of participation. Ethical approval was obtained via the UCL research ethics committee (approval code: SHaPS-2018-CM-030) and informed consent given by all participants.

Materials

We obtained voice recordings from the 16 romantic couples (i.e. the 32 participants), 6 adult voices from the freely-available LUCID corpus of speech materials (3 female; Baker & Hazan, 2011), and 2 further adult voices recruited from within the Department of Speech, Hearing and Phonetic Sciences at UCL (1 female). Recordings of the couples were used in the experimental tasks to represent personally-familiar voices (i.e. the romantic partners), while the additional voice identities were used to represent, for each participant, 1 lab-trained identity plus 3 further unfamiliar identities used across the perceptual tasks (see Procedure).

Spontaneous speech

Spontaneous speech was elicited from the couples by asking them to perform the DIAPIX task (Baker & Hazan, 2011). This task involves pairs of participants engaging in an interactive "spot the difference" task: each individual receives only one image in a picture pair, and the aim is to locate all 12 differences between the pictures through discussion of their respective images. In a preliminary session, we recorded each couple discussing a total of three DIAPIX image pairs. The couple was seated in separate sound-attenuating chambers. Each participant wore Beyerdynamic DT297PV headsets fitted with cardioid microphones to enable discussion of their images, and so that we could record their speech (without interference from their partner). Speech was recorded and digitised at a sampling rate of 44100Hz. Both participants were required to click with their mouse at the location of each difference so these could be scored. Each session lasted as long as it took to find all 12 differences, or until a 10-minute timer ended.

Short excerpts (1.5-2s) of fluent and meaningful spontaneous speech, as well as conversational filler sounds (e.g. “um”, “mm”), were selected from each member of each couple, as well as from 6 additional SSBE speakers’ DIAPIX recordings (3 female, obtained via the LUCID corpus; Baker & Hazan, 2011). Fillers were selected on the basis that they were not lexical (e.g. “mmm”, “umm”, “uhuh” would be included; “yeah” or “yep” would not be included). All stimuli were saved as mono WAV files using PRAAT (Boersma & Weenink, 2010), normed for RMS amplitude, and finally converted into mp3 format for use on the online testing platform Gorilla.sc. These stimuli were used for the familiarisation and fillers task (Task 1) of the current study.

Read sentences

Sentence stimuli included:

- 50 items from the LUCID corpus (e.g. “My brother Paul ran towards the beach.”), produced by the couples (personally-familiar voices) and by the same 4 LUCID corpus speakers chosen for the spontaneous speech materials (lab-trained and unfamiliar voices). These sentences were used in the voice modulation task (Task 2) of the current study.
- 50 items from the co-ordinate response measure (CRM) database (Bolia, Nelson, Ericson, & Simpson, 2000), produced by the couples (personally-familiar voices) and two novel speakers. CRM sentences take the form “Ready [call sign], go to [colour] [number] now.” The call signs used were “Baron”, “Eagle”, and “Laker”, colours were “red”, “green”, “blue”, and “white”, and the numbers were one to eight. These items were used in the speech perception task (Task 3) of the current study.

All newly-recorded items were recorded in a sound-attenuating chamber, using a Røde NT-1A microphone connected to an RME fireface UC audio interface at a sampling rate of 44100Hz. Stimuli were normed for RMS amplitude and converted into mp3 format as required for online testing.

Assignment of voice identities to tasks

In all the tasks described for the couples group, recordings of each participant’s romantic partner represented the personally-familiar voice, while other, previously-unknown voices were used as lab-trained and unfamiliar identities. To control for basic acoustic cues across the identities, all voices used per participant were of the same sex as the romantic partner. The assignment of these unknown identities to the voice conditions was as follows:

- *Familiarisation of the lab-trained voice:* The participant’s romantic partner represented the personally-familiar voice. One of the 4 LUCID corpus speakers was used as the lab-trained voice (“Anna” or “Adam”), and one further LUCID speaker of the same sex was used as an unfamiliar identity (“Someone else”).
- *Tasks 1 & 2:* The participant’s romantic partner represented the personally-familiar voice. The familiarised LUCID corpus speaker was used as the lab-trained voice (“Anna” or “Adam”), plus a previously-unheard LUCID speaker was introduced as a new identity (“Clara” or “Charlie”)
- *Task 3:* The participant’s romantic partner represented the personally-familiar voice. A further novel, unfamiliar identity was introduced, using recordings from one of the speakers recruited from UCL Speech, Hearing and Phonetic Sciences.

We note that for each participant in the control group, the personally-familiar voice of one couples group member was presented as a second lab-trained identity, labelled either “Beth” or “Ben”.

Vigilance stimuli

A text-to-speech online tool (https://text2speech.us/) was used to generate computerised voices reading “Please press the left key”, and “Please press the right key.” These were used in vigilance trials (8 per task; 4 of each instruction) to check participants’ attention to the perceptual tasks.

Procedure

Online testing session

Approximately 1-2 weeks after recording the stimuli, each of the participants in the couples group completed the perceptual tasks independently (i.e. not in the presence of their partner) on the online testing platform Gorilla.sc (Anwyl-Irvine et al., 2018). A link to a personalised version of the study was sent to participants via email. Participants in the control group were recruited via the online recruitment platform Prolific.co (www.prolific.co) and also completed the tasks on Gorilla.sc. Participants set the volume of the stimuli to a comfortable listening level and were required to pass a headphone screening to ensure that participants were wearing headphones and able to hear the stimuli presented (Woods, Siegel, Traer, & McDermott, 2017). Each trial of the screening task involves judging which of three tones is the quietest. In each triplet, one tone is presented 180 degrees out of phase across the stereo channels. This makes the task simple with headphones, but difficult without, due to phase cancellation when listening over loudspeakers.

In each of the three main tasks, eight vigilance trials were included to ensure participants were paying sufficient attention to the audio stimuli. These trials required participants to press the left or right arrow keys on their keyboard in accordance with the audio instruction (see vigilance stimuli), instead of clicking a response option with their mouse. Participants that failed to respond correctly at least 75% of the time on these trials were excluded from the relevant task.

Familiarisation of the lab-trained voice

In order to directly compare the recognition of a lab-trained voice and one that is personally-familiar, listeners first needed to be trained to recognise a new voice before completing the perceptual tasks. Of the spontaneous speech excerpts extracted from the DIAPIX task recordings, 24 excerpts each were chosen for the personally-familiar voice and the lab-trained voice. For use in a passive exposure phase, these were arranged into two 12-excerpt sequences, with each sound clip separated by 1s of silence. For use in a test phase, a further 20 spontaneous speech stimuli were selected from all three identities (personally-familiar, lab-trained, unfamiliar).

In the familiarisation, participants in the couples group were first passively exposed to the lab-trained voice (introduced as either “Anna” or “Adam”; matched to their romantic partner’s sex), as well as re-acquainting themselves with their partner’s voice. Participants always heard the lab-trained voice first and their partner’s voice second. After listening to the two sequences of spontaneous speech from both identities, participants were tested on recognition of the two voices. The 60 test stimuli (20 each from the partner, the lab-trained voice, and an unfamiliar voice) were presented in a fully randomised order. Each trial consisted of a short voice clip, followed by three text response options: “My partner”, “Anna(/Adam)”, or “Someone else” - responses were made via a mouse-click to select one of these options. Audio-visual feedback (correct/incorrect) was given on every trial to aid learning of the new voice. This task lasted approximately 5-10 minutes. After this training, listeners were able to recognise the lab-trained voice with good accuracy (80.65% correct,

SD = 2.5%, chance = 33%). Control participants performed the same familiarisation task, however the "personally-familiar" voice was introduced as a lab-trained identity labelled either "Beth" or "Ben" for this group. Thus, these listeners learned to recognise two identities: "Anna"/"Adam" and "Beth/Ben". Recognition accuracy after training was also high in this group, for both lab-trained voices ("Beth"/"Ben": mean = 82.58%; mean "Anna"/"Adam" = 81.77%). Both voices were thus recognised with similar accuracy and ease, and at a comparable level to the recognition of "Anna"/"Adam" by the couples.

Following the training, participants either performed voice identity recognition from non-verbal vocalisations (Task 1) or voice identity recognition from acoustically modulated voices (Task 2) first. The order of Tasks 1 and 2 was counterbalanced across participants. Before the start of the first task, listeners were introduced to a novel and thus unfamiliar voice "Clara"/"Charlie" and presented with one example speech token from this speaker - this was their only exposure to this speaker before the task began. Note this was a different unfamiliar talker from the one used in the familiarisation.

The speech perception task (Task 3) was always completed last. For this task, a final unfamiliar talker was used but was not introduced to the participant, by name or otherwise. Participants did not receive any feedback on their performance during Tasks 1-3.

Task 1: Voice identity recognition from non-verbal vocalisations

In this task, listeners performed voice identity recognition from vocal stimuli with only minimal linguistic cues (i.e. filler sounds such as "umm", "uhh"). In general, it has been observed that voice identity perception from short, non-verbal stimuli is more challenging than from longer stimuli that include linguistic content (Schweinberger, Herholz & Sommer, 1997; Bricker & Pruzansky, 1966). Familiarity has, however, been found to produce benefits for voice identity perception, even under such challenging listening conditions. Evidence for familiarity advantages in these contexts comes from research using naturally-varying non-verbal vocalisations, such as spontaneous and volitional laughter, coughs and cries. For instance, Lavan, Scott, and McGettigan (2016) asked participants to perform a voice discrimination task on paired combinations of vowels and laughter. Listeners were either familiar (students hearing their lecturers' voices), or unfamiliar with the voices. In that study, familiar listeners were better able to discriminate between pairs of non-verbal vocalisations compared to unfamiliar listeners. Similarly, Zarate, Tian, Woods, and Poeppel (2015) demonstrated above-chance recognition of 5 lab-trained voices from non-speech vocalisations (e.g. laughs, coughs, cries, & grunts) after brief training. However, it should be noted that accuracy was overall low in both studies, and the non-verbal condition generated the worst performance for Zarate et al. (2015). In their discussion, Lavan and colleagues (2016) suggest that a lack of familiarity with the specific types of vocalisations used in the study (e.g. the lecturers' laughter) may have led to the observed impairments in their student participants (cf Lavan, Burston, et al., 2019).

Thus, while it has been shown that familiar listeners have an advantage for identification of known voices, limitations in the extent and/or content of their prior exposure to the test voices meant that identity perception was still error-prone under certain circumstances. In order to have a robust stored representation of a voice, a listener may need to have experience with the full range of vocalisations, produced in a variety of contexts (Lavan, Burton, et al., 2019). Accordingly, for individuals with whom we are personally-familiar (e.g. romantic partners, as in the current study), costs to performance should be reduced compared to lab-trained voices because stored representations should be built from more comprehensive exposure

to the speaker's vocal repertoire. Therefore, using non-verbal filler sounds as representative of vocalisations with minimal linguistic cues, we aimed to test this prediction.

Methods

Stimuli

20 filler sounds (mean duration = 0.59s) were extracted from the DIAPIX task recordings per identity (personally-familiar voice [lab-trained "Beth"/"Ben" for controls], lab-trained voice "Anna"/"Adam", and the unfamiliar voice Clara/Charlie) for this task, as well as 8 vigilance stimuli. The personally-familiar voice was always the romantic partner of one participant from the couples group. The lab-trained and unfamiliar voices were the same for all couples and control participants (where female participants heard male voice identities, and *vice versa*). Examples of stimuli used in this task can be found at: https://osf.io/g2jk6/.

Procedure

In this task, participants heard a total of 60 filler sounds produced by the three speakers (personally-familiar, lab-trained, unfamiliar) in a randomised order. On each trial, a filler sound was presented, followed by a prompt asking participants to select the identity they thought had produced it from three response options ("My partner", "Anna"/"Adam", "Clara"/"Charlie") via mouse-click. For control group participants, the three response options were "Beth"/"Ben", "Anna"/"Adam", and "Clara"/"Charlie". Vigilance trials required participants to respond with a keypress (left or right arrow key) instead of selecting a text response option with their mouse. This task lasted approximately 5 minutes.

Data Analysis

Unbiased hit rates (H_u scores) were calculated for each of the three familiarity conditions (personally-familiar, lab-trained, unfamiliar) to correct for any disproportionate usage of certain response categories (Wagner, 1993). Taking personally-familiar voice trials as an example case:

- Correct "My Partner" responses were defined as hits
- False alarms were defined as incorrectly responding with "My Partner" when hearing the lab-trained or unfamiliar voices.
- Correct "Anna"/"Adam" or "Clara"/"Charlie" responses were correct rejections
- Incorrect responses of "Anna"/"Adam" or "Clara"/"Charlie" to the participant's partner's voice were defined as misses.

H_u scores were arcsine transformed (Wagner, 1993). Data were analysed using linear mixed models (LMMs) via the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) in the *R* environment (R core team, 2013). For the LMMs, model estimates and associated confidence intervals are reported as an estimate of the size of relevant effects. The further estimates deviate from zero, the greater the effect. Confidence intervals that do not cross zero are significant. Following our pre-registered analysis plan, we analyse and report the findings of the couples and controls separately.

Results

Data from four couples group participants (and the corresponding members of the control group) were removed for failing the attention checks (i.e. scoring less than 6/8 on vigilance trials). Thus, 27 participants per group were retained for the statistical analyses.

Couples

To assess the impact of the three types of familiarity (personally-familiar; lab-trained; unfamiliar) on voice identity recognition performance based on the non-verbal filler sounds, an LMM was run with H_u scores for recognition performance as the outcome variable. In this confirmatory analysis, familiarity was entered into the model as a fixed effect, and random intercepts of participant and voice identity were added as random factors. Statistical significance was established via likelihood ratio tests comparing the full model that contained all fixed and random effects to a reduced model where the relevant effect had been dropped.

Familiarity had a significant effect on voice identity recognition ($\chi^2(2) = 20.33$, $p < .0001$), with post-hoc comparisons (via the *emmeans* package in R) indicating that listeners were significantly better at recognising their partner's voice (raw mean = 91.3%, SD = 9.7%) compared to the lab-trained ($p = .001$; raw mean = 64.4%, SD = 13.8%, E = -0.60 , CI = [-0.93, -0.28]) and unfamiliar identities ($p < .001$; raw mean = 47.2%, SD = 16.7, E = -0.69, CI = [-1.01, -0.37]; see Figure 1a). Figure 2 illustrates responses as a confusion matrix – this shows both a high hit rate and low false alarm rate for the personally-familiar voice, while the lab-trained and unfamiliar voices were more frequently confused with one another.

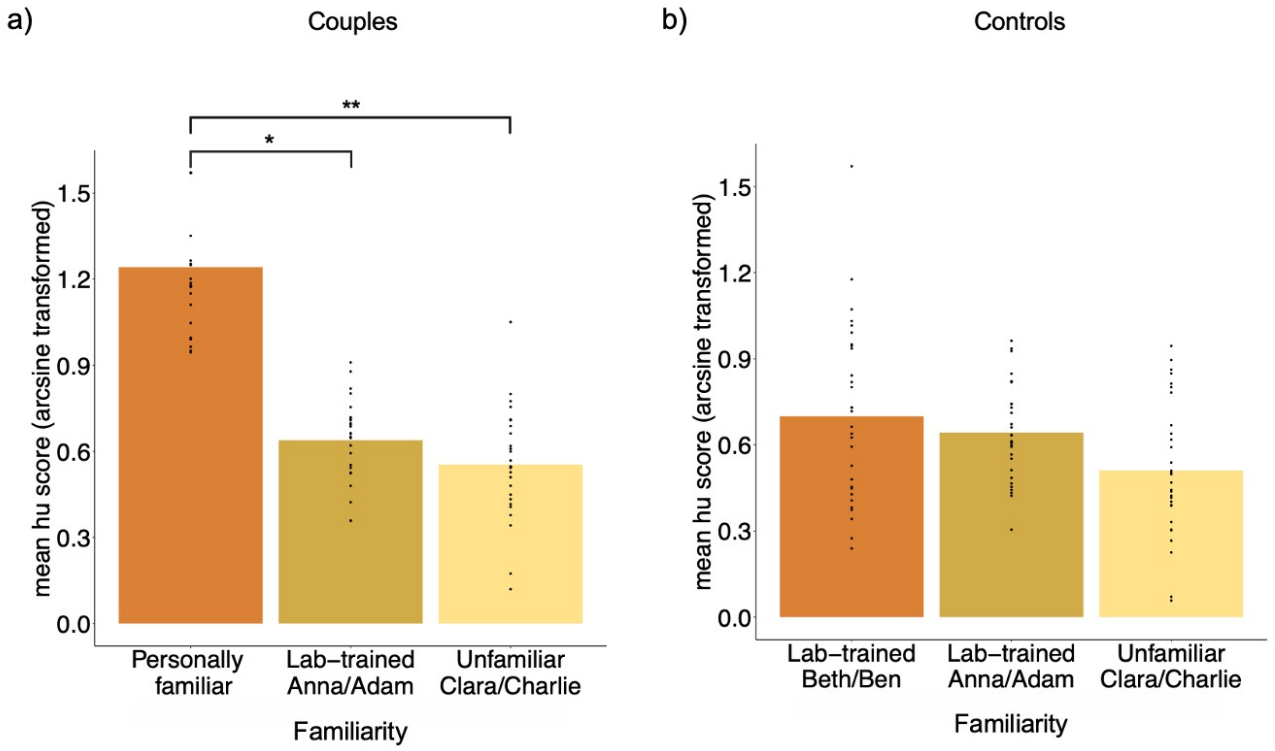

Figure 1. Bars display mean Hu scores (unbiased hit rates) for each of the three speakers in the fillers task (Task 1) for (a) the couples group and (b) the control group. Points represent individual participants' Hu scores for each speaker identity. ** $p < .001$, * $p = .001$. For colour figures, please see the online version of this article.

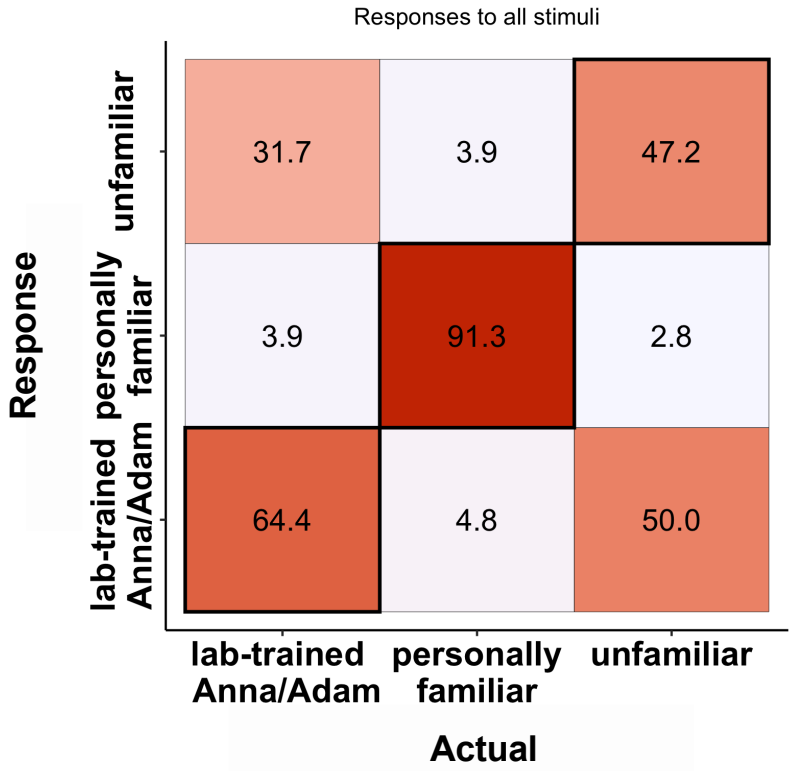

Figure 2. Confusion matrix displaying the couples group's responses per condition for the recognition of voice identity from non-verbal vocalisations (Task 1). Each cell shows the percentage of trials in which a presented voice ("Actual") was perceived as one of the three target identities ("Response"). Cells on the diagonal reflect correct responses (hits); darker reds indicate higher percentages. See the Supplementary Materials for the corresponding control group confusion matrix. Please see the online version of this article for colour figures.

Controls

If our observed results for the couples group were due to relative familiarity of the couples with the personally-familiar and lab-trained voices, and not due to systematic differences in distinctiveness or recognisability of these voices *per se*, there should be no significant differences in control participants' performance for these two identities in both vocal identity tasks (i.e. recognising identity from non-verbal filler sounds, and from modulated sentences).

To assess the impact of the three voice identities on recognition accuracy, an LMM was run with the same fixed and random effects, and model comparison, as reported for the couples group. Statistical significance was again established via likelihood ratio tests comparing the full model that contained all fixed and random effects, to a reduced model that did not include familiarity. Note that familiarity was still defined with 3 levels, corresponding to lab-trained "Beth"/"Ben" (i.e. personally-familiar for couples), lab-trained "Anna"/"Adam", and the unfamiliar voice, respectively.

Comparing the full model to the reduced model revealed no significant differences in performance between the three identities (two lab-trained (Beth/Ben: E = 0.70, CI = [0.59, 0.80]; Anna/Adam: E = -0.07, CI = [-0.36, 0.22]) and one unfamiliar (E = -0.22, CI = [-0.50, 0.07]) voice; $\chi^2(2) = 2.45$, p = .294; See Figure 1b). This shows that there was no overall difference in distinctiveness between the two lab-trained voices. This analysis therefore

shows that the effects observed for the couples group are a result of the familiarity with the personally-familiar partner's voice, and not artefacts of the stimuli used in our task.

Raw recognition accuracy for the two lab-trained voices was 65.2% (Anna/Adam; SD = 15.2%) and 62.7% (Beth/Ben; SD = 23.6%), and 44.1% for the unfamiliar voice (SD = 19.4%); however, there were frequent categorisation errors (see Supplementary Figure 1).

Discussion

Both couples and control group listeners in our study showed above-chance performance on all conditions of Task 1 (see Supplementary Materials). Participants in the couples group displayed near-perfect accuracy at recognising their romantic partner's voice from non-verbal filler sounds (raw accuracy = 91.3%). In comparison, performance for lab-trained voices in the current experiment was similar to previous work on non-verbal vocalisations, finding above-chance but error-prone performance overall (cf Zarate et al., 2015). While errors in the couples group were mainly associated with confusions between the lab-trained and unfamiliar voices (see Figure 2), the control listeners' confusions affected all three identities (see Supplementary Figure 2).

Our results therefore confirm our prediction that different kinds of familiarity affect how well listeners can perceive identity from these voices. While listeners excelled at perceiving voice identity from personally-familiar voices in this challenging task, they struggled more to accurately perceive identity from the lab-trained voices. Our data also suggest that brief exposure to a voice identity through training may be sufficient to distinguish a voice from other identities, but that this ability is vulnerable to interference from other voices (see Figure 2 and Supplementary Figure 2). Lab-trained identities may therefore be associated with less robust stored representations, reducing the degree to which a listener can recognise a speaker with limited cues to vocal identity (Fontaine, Love, & Latinus, 2017). In contrast, for personally-familiar voices, for which a robust representation has been formed through extensive and varied exposure, we observe highly accurate identity perception.

Task 2: Voice identity recognition in the context of acoustic modulation

The results of Task 1 showed that personally-familiar voices are recognised with higher accuracy from naturally-produced non-verbal utterances, due to the presence of a more robust perceptual representation of that voice. However, in addition to questions about overall accuracy of recognition, we can also ask questions about the content of that representation, by probing voice recognition in the presence of acoustic voice modulations. In a second perceptual task, we therefore presented listeners with short voice recordings in which acoustic cues had been modulated. Through this process, we were able to examine how listeners' representations are tuned to the manipulated acoustic cues, and how this tuning is affected by familiarity.

Previous work on identity perception has attempted to identify which acoustic cues are used by listeners to recognise identity in familiar voices. Lavner and colleagues (2000) tested listeners' recognition of personally-familiar identities (members of a kibbutz in which the listeners lived) based on recordings of vowels. Participants were asked to identify twenty voices from a list of twenty-nine possible speakers. Of those correctly identified, acoustically modified versions of the recorded vowels were then presented. These modifications included shifting individual formants and altering fundamental frequency, amongst other acoustic modulations. Modulation of vocal tract properties (i.e. formant frequencies) were

identified as being most disruptive for recognition, although different combinations and weightings of acoustic features were diagnostic for different individual voice identities.

Given the perceptual salience of glottal pulse rate (GPR, related to the fundamental frequency) and vocal tract length (VTL, related to formant frequencies) as cues for voice identity perception, Gaudrain, Li, Ban, and Patterson (2009) explored the degree to which these properties could be altered until listeners no longer recognised that two voice samples were produced by the same unfamiliar speaker. They found that VTL may be modulated to a smaller degree than GPR before listeners perceive changes in perceived identity.

Examining how the modulation of specific acoustic features affects unfamiliar listeners' judgements in voice identity perception tasks can therefore offer information about the relative importance of various acoustic cues to recognition, and can further illuminate the robustness and nature of the underlying representation of a personally-familiar voice. In the current task, we predicted that acoustic modulations would affect vocal identity recognition differentially for personally-familiar and lab-trained voices. However, we had no clear prediction of the direction of this effect: On the one hand, increased knowledge of one's partner's voice may allow a listener to accept larger modulations of voice acoustics without a cost to recognition. Alternatively, more in-depth knowledge of a speaker's vocal repertoire may restrict the range of acoustic properties that would be accepted as belonging to that personally-familiar voice, relative to a lab-trained identity.

Methods

Stimuli

This task used 50 read sentences extracted from the LUCID corpus materials, produced by the same identities as used in Task 1. Sentences were acoustically modulated with STRAIGHT (Kawahara & Irino, 2004) in the MATLAB environment (see Gaudrain, 2018) to simultaneously introduce changes in GPR and VTL in semitones (a semitone is a twelfth of an octave). GPR was altered by two or four semitones in either direction, and VTL by one or two semitones, so that with every upward semitone shift in VTL, there was an accompanying two-semitone downward shift in GPR, and vice versa (Gaudrain et al., 2009; see Figure 3a). The overall effect of the combined modulations was to create voices that sounded relatively more masculinised (i.e. lower pitch and longer vocal tract) and feminised (i.e. higher pitch and shorter vocal tract) than the original voice. Examples of each of the modulations steps, from 1 male and 1 female speaker, are publicly available on the open science framework (OSF) and can be accessed at: https://osf.io/g2jk6/. Once processed with STRAIGHT, 12 stimuli were randomly selected for each step for both the personally-familiar ("Beth"/"Ben") voice and the lab-trained ("Anna"/"Adam") voice – as there were only 50 recorded sentences available, two randomly selected items from each modulation step and from the unshifted voice recordings were repeated once each during the task. Six tokens per step were selected for the unfamiliar voice.

Procedure

In this task, participants were presented with the 150 modulated and unmodulated stimuli (60 each for the personally-familiar and lab-trained voices, 30 for the unfamiliar voice) in a fully randomised order. On each trial, a sentence was presented, followed by a prompt asking participants to select the speaker they thought they had heard from three text response options ("My partner", "Anna"/"Adam", "Clara"/"Charlie") via mouse-click. For controls, the three response options were "Beth"/"Ben", "Anna"/"Adam", and "Clara"/"Charlie". Vigilance trials required participants to follow an instruction to respond with

a keypress ("please press the left/right arrow key"), instead of selecting a text response option with their mouse. The task took approximately 15 minutes to complete.

Data Analysis

Unbiased hit rates (H_u scores) were calculated for each of the three familiarity conditions (personally-familiar, lab-trained, unfamiliar) to correct for any disproportionate usage of certain response categories (Wagner, 1993). H_u scores were arcsine transformed (Wagner, 1993). Data were analysed using linear mixed models (LMMs) via the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) in the *R* environment (R core team, 2013). For the LMMs, model estimates and associated confidence intervals are reported as an estimate of the size of relevant effects. The further estimates deviate from zero, the greater the effect. Confidence intervals that do not cross zero are significant. Following our pre-registered analysis plan, we analyse and report the findings of the couples and controls separately.

Results

Data from two couples group participants (and the corresponding members of the control group) were removed for failing the attention checks (i.e. scoring less than 6/8 on vigilance trials). Thus, 29 participants per group were retained for the statistical analyses.

Couples

Averaging across all modulation steps per speaker identity showed that the mean overall performance for the personally-familiar voice was 79.5% (SD= 14.0%), with mean scores on the individual modulation steps ranging from 58.6% - 98.9%. Mean overall performance for the lab-trained voice was 56.0% (SD= 10.0%) with mean scores on individual steps ranging from 43.9% - 71.2%. For the unfamiliar voice, mean overall performance was 50.4% (SD= 11.1%), ranging from 45.4% - 56.9% across the individual modulation steps.

To evaluate the effect of the acoustic modulations on recognition of the three identities, we analysed the interaction between degree of modulation (i.e. modulation "step"), and familiarity using LMMs. In this confirmatory analysis, the outcome measure was the H_u score for recognition performance; familiarity and degree of modulation were included as fixed effects, including the interaction between familiarity and degree of modulation. Participant and speaker identity were included as random effects. However, after accounting for the variance explained by participants, speaker identity did not explain any additional variance and was thus removed from the models. Statistical significance was again established by comparing the full model including the interaction, fixed, and random effect to a reduced model that included all of the same fixed and random effects, but did not include the interaction.

Comparing the full model to the reduced model indicated a significant interaction between familiarity and the degree of modulation ($\chi^2(8) = 40.68$, $p < .0001$; see Figure 3b).

Post-hoc pairwise comparisons (using *emmeans*) were run to assess the effect of increasing the degree of modulation on recognition of the three identities, and Bonferroni-corrected for 4 comparisons per voice identity (adjusted alpha = .0125). Performance for the personally-familiar voices was negatively affected by each additional step in both directions (all ps < .001; see Figure 3b). For the lab-trained voice, acoustic modulation only produced a significant decrease in performance for one comparison (unshifted vs. one step shift in negative direction; p = .0083). For the unfamiliar voice condition, acoustic modulation did not produce a significant difference in performance relative to the original voice (ps for all

comparisons > .013). These results suggest that acoustic manipulations had a bigger effect on performance for the personally-familiar voice identity than for the lab-trained and unfamiliar identities – this effect is in part due to performance being overall much better for personally-familiar voices, such that there was also greater scope for performance to decrease.

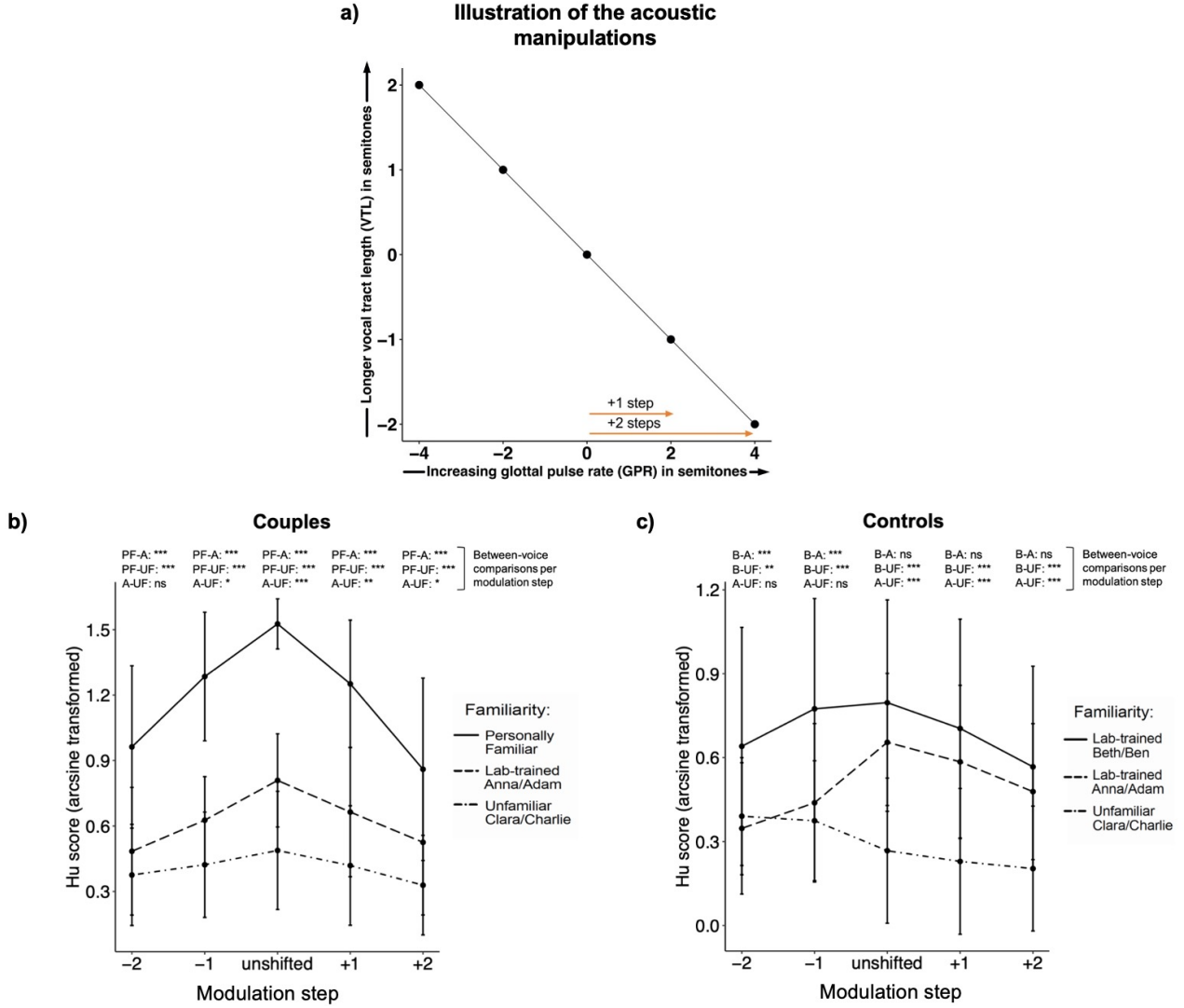

Figure **3** a) Acoustic manipulations made to the voices in Task 2. Points represent the five modulation steps used, plotted as combined shifts in glottal pulse rate (GPR) and vocal tract length (VTL) relative to the original voice recordings (i.e. 0,0). Increases in GPR (in semitones) correspond to sounds with higher subjective pitch. For vocal tract length, a positive shift in VTL (in semitones) gives the percept of a longer vocal tract. Orange (lighter) arrows show how the acoustic manipulations corresponded to the modulation "steps" described in the analyses. b) and c) Mean Hu scores are displayed per familiarity condition (Personally-familiar/Lab-trained "Beth"/"Ben", Lab-trained "Anna/Adam", unfamiliar) and modulation step (x-axis) for couples (left) and controls (right). Error bars display standard deviations around the mean. Asterisks denote significance of between-voice comparisons at each modulation step; PF = personally-familiar, A = Lab-trained "Anna"/"Adam", B = Lab-trained "Beth"/"Ben", UF = unfamiliar; *** p < .0001, ** p < .001, * p < .01, ns = not significant. Please see the online version of this article for colour figures.

A further set of post-hoc tests explored the effects of familiarity, via three Bonferroni-corrected pairwise comparisons at each modulation step (adjusted alpha = .0167). At all but one modulation step, performance was significantly different depending on familiarity with the speaker (personally-familiar > lab-trained, lab-trained > unfamiliar, personally-familiar > unfamiliar; ps < .0167). For the most masculinised condition (i.e. step -2), there was no significant difference between the lab-trained voice and the unfamiliar voice (p = .114). Differences in recognition accuracy between the personally-familiar voice and the two other conditions were smaller at the largest modulation steps (i.e. -2 and +2) compared to the unshifted condition (lab-trained (step -2): E = 0.240, CI = [0.05, 0.43], (step +2): E = 0.383, CI = [0.20, 0.57], unfamiliar (step -2): E = 0.452, CI = [0.26, 0.64], (step +2): E = 0.507, CI = [0.32, 0.69]), again suggesting that acoustic manipulations had a larger effect on personally-familiar voice recognition. A table detailing all post-hoc tests is included in the Supplementary Materials (see Supplementary Table 3).

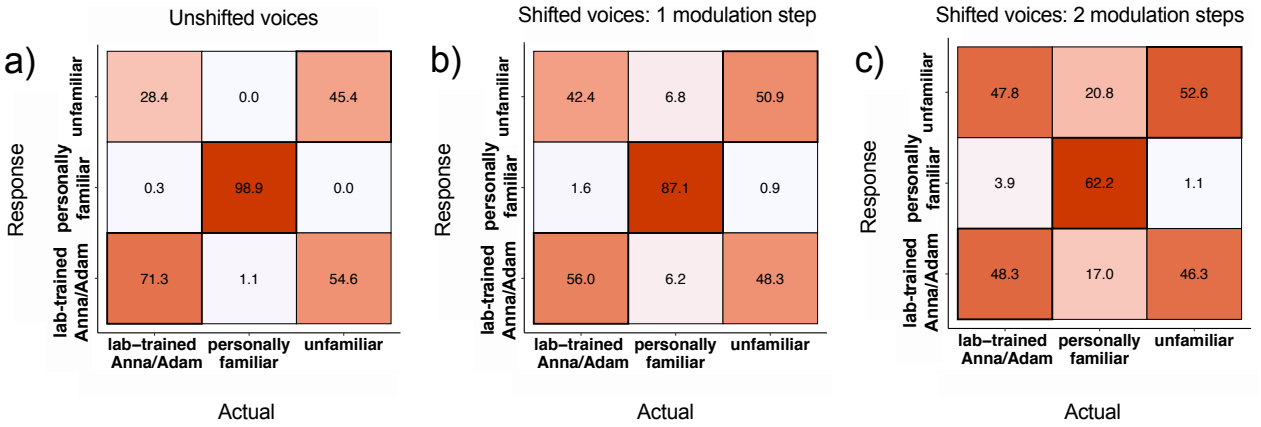

Figure 4. Confusion matrices displaying the couples group's responses in the modulation task (Task 2). Matrices are shown for each modulation step: a) Unshifted condition: participants' raw responses to the speaker's "original" voices; b) 1 modulation step: displays hits, misses, and false alarms for the three identities when these voices had been modulated by one step (collapsed across direction of acoustic modulation); c) 2 modulation steps: displays hits, misses, and false alarms for the three identities modulated by 2 steps (collapsed across direction of acoustic modulation). See the Supplementary Materials for the corresponding control group confusion matrix. Please see the online version of this article for colour figures.

Confusion matrices displaying the group averages of raw responses for each trial were constructed to examine the types of categorisation errors made by listeners (see Figure 4) – this shows that increasing distance from the original voice led to decreases in hits (i.e. labelling the partner as the partner) and increases in misses (i.e. labelling the partner as another identity) while false alarms (i.e. labelling another identity as the partner) remained very low and stable across conditions.

Controls

Averaging across all modulation steps per speaker identity showed that the mean overall performance for the lab-trained "Beth"/"Ben" voice (corresponding to the romantic partners of the couples group) was 55.8% (SD = 20.5%), with mean scores on the individual modulation steps ranging from 44.3% - 65.2%. Mean overall performance for the lab-trained "Anna"/"Adam" voice was 47.3% (SD =11.6%) with mean scores on individual steps ranging

from 25.9% - 64.7%. Lastly, for the unfamiliar voice, mean overall performance was 37.8% (SD = 13%), ranging from 27% - 52.9% across the individual modulation steps.

To assess the effect of acoustic modulation on recognition of the three identities, we analysed the interaction between modulation step and familiarity using LMMs as described for the couples group above. Comparing the full model to a reduced model that did not contain an interaction, we found a significant interaction between modulation step and familiarity ($\chi^2(8) = 32.49$, $p < .0001$). As in the couples group, we assessed both the effect of modulation step on performance within each identity, and differences between familiarity conditions (lab-trained voices and unfamiliar) within each modulation step. Post-hoc pairwise comparisons (using *emmeans*) were first run comparing performance between successive modulation steps (e.g. -2 steps vs. -1 step, -1 step vs. unshifted condition) for each identity separately, and Bonferroni corrected for 4 comparisons per identity (adjusted alpha = .0125). The results showed that modulation step did not have an effect on performance for all three identities (ps > .0125), except for one comparison: for lab-trained "Anna"/"Adam", a shift of one step in the negative direction resulted in significantly lower performance than performance for the original unshifted "Anna"/"Adam" voice (p = .0021).

Next, we compared performance for the three identities (lab-trained "Anna"/"Adam", lab-trained "Beth"/"Ben", unfamiliar), using three Bonferroni-corrected pairwise comparisons at each modulation step (adjusted alpha = .0167). Significantly better performance was observed for lab-trained "Beth"/"Ben" compared to lab-trained "Anna"/"Adam" for voice tokens shifted by 1 and 2 steps in the negative direction (see Figure 3c). Performance was also significantly better for lab-trained "Beth"/"Ben" compared to the unfamiliar voice (p < .0167) at all modulation steps. Performance for lab-trained "Anna"/"Adam" voice was better than the unfamiliar voice for the unshifted condition, and for tokens shifted in the positive direction (all ps < .0167). A table detailing all post-hoc tests for controls is included in the Supplementary Materials (see Supplementary Table 4).

Discussion

In this task, we sought to examine whether acoustic modulation would affect perception of a personally-familiar voice differently to a lab-trained voice. We found that in the couples group, listeners were able to recognise their partner's voice with a high level of accuracy, and significantly better than the two other talkers, even with acoustic manipulations. We propose that these results can be explained by greater experience with a romantic partner's vocal inventory, which facilitates better overall recognition of the personally-familiar voice compared to the lab-trained voice. Specifically, when perceptually-salient voice features such as GPR and VTL are unavailable or altered, listeners may be better able to use other available cues to vocal identity such as accent information or speech rate when the voice is personally familiar (Maguinness, Roswandowitz, & von Kriegstein, 2018).

Acoustic manipulations nonetheless affected listeners' performance for the personally-familiar voices: we observed a symmetrical drop in accuracy with increasing acoustic modulation, and a consequently sharp "tuning function" in response to shifts in GPR and VTL. Although a similar pattern was also present for the lab-trained voice, it was far less marked and the resulting "tuning function" was substantially flatter. Thus, despite having access to a wider range of diagnostic vocal cues to their partner's identity, listeners are at the same time quite sensitive to deviations from its expected acoustic register.

Inspection of the patterns of responses in this task confirmed that there were very low rates of false alarms for the partner's voice – that is, the lab-trained and unfamiliar voices were rarely identified as the partner, across all conditions. In fact, the lower recognition performance in response to greater modulation was associated with increasing *rejection* of the partner's voice, such that it was misrecognised as the lab-trained or the unfamiliar identity more frequently as acoustic deviance increased. In contrast, mutual confusions of the newly-learned and unfamiliar voices were frequent, and increased with the degree of acoustic modulation. We interpret this finding in line with our prediction that listeners may be more attuned to their romantic partner's voice and the dynamics of their vocal system, owing to a more robust representation (Lavan et al., 2016; Fontaine, Love, & Latinus, 2017). Hence, if voice acoustics are altered sufficiently, listeners hearing an acoustically deviant version of a personally-familiar voice may be less willing to accept these voice tokens because they are not compatible with their stored representations of that voice identity. A similar pattern of results has been reported in the face perception literature, where face morphing studies showed that in order for a morph of a personally-familiar and an unfamiliar face to be perceived as familiar, the morphed face needed to include 60% or more of the personally-familiar face (Chauhan & Gobbini, 2018). Those authors concluded that personal familiarity in a variety of contexts sharpened tuning to the distinct features that represent the familiar identity. This in turn meant that their participants were able to be more conservative in rejecting images that violated the representations of personally-familiar individuals. We argue that a similar mechanism is observed in the current study. Here, fine-tuned representations built via personal familiarity allowed listeners to be more conservative, rejecting tokens that no longer matched the stored representation of their partner (due to artificial modulation) while also preserving the ability to accurately reject tokens from other speakers.

Task 3: Speech perception from personally-familiar voices

In the final task, we examined the effects of familiarity with a voice beyond identity perception, by testing speech perception in noise. Understanding the content of speech is an important part of voice perception, yet achieving accurate speech recognition is not always easy: for instance, in environments where there are multiple speakers, other background noise, or if the listener has a hearing impairment. However, familiar voices have been reported to be more intelligible than unfamiliar voices in studies comparing performance of familiar and unfamiliar listeners in various speech in noise perception tasks (Holmes, Domingo, & Johnsrude, 2018; Johnsrude et al., 2013; Kreitewolf, Mathias, & von Kriegstein, 2017; Newman & Evers, 2007; Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994). In such tasks, familiarity advantages have been found for both lab-trained voices (e.g. Nygaard et al., 1994; Nygaard & Pisoni, 1998; Kreitewolf, et al., 2017), as well as personally-familiar ones (e.g. Souza, Gehani, Wright, and McCloy, 2013; Holmes et al., 2018; Johnsrude et al., 2013; Holmes & Johnsrude, 2020).

In the current study, we measured speech perception in noise (multi-talker babble) to examine whether listeners would also show a familiarity advantage. Based on prior studies finding these familiar voice benefits, we predicted a higher percentage of correctly-reported sentences for stimuli spoken by the personally-familiar voice, compared to accuracy rates for sentences spoken by an unfamiliar speaker.

Methods

Stimuli

In this task, we tested speech perception from the personally-familiar voice and a novel unfamiliar voice. The unfamiliar voice was distinct from the unfamiliar voice ("Someone else") used in the familiarisation, and from the unfamiliar voice ("Clara/Charlie") used in Tasks 1 and 2. All recorded CRM sentences were first RMS normalised. Four-talker babble (multi-talker babble is background noise made up of multiple talkers, in this case four talkers) was then added to each of the sentences from the personally-familiar and unfamiliar voices at a signal-to-noise ratio (SNR) of -6dB. Sample stimuli used in this task are publicly available via the OSF, and can be accessed via the following link: https://osf.io/g2jk6/. The babble noise was created from recordings in the EUROM database of English speech (Rosen, Souza, Ekelund, & Majeed, 2013; Chan et al., 1995), and comprised speakers of the same sex as the to-be-masked speaker – hence, male voices in our experiment were masked with male babble, and female voices masked with female babble. Eighty sentence-in-noise stimuli (40 from each voice) were selected for use in the task.

Procedure

This task was always completed last in the testing session. Here, participants were instructed to listen to the CRM sentences produced by the target speakers (partner [lab-trained Beth/Ben for controls], unfamiliar), whilst ignoring the background noise (four-talker babble). Once each stimulus had played, participants were presented with a grid comprising four rows: each row contained the numbers 1-8 in one of the four colour options (red, green, blue, white). Participants were instructed to select the colour and number combination they had perceived from the target sentence. For example, for the sentence stimulus "Ready Baron, go to blue three now", the participant should select the blue three from the grid. The 80 stimuli (40 sentences per voice) were presented in a fully randomised order, and the task took around ten minutes to complete.

Data Analysis

Correct answers were defined as trials where participants correctly identified both the colour and number in the target sentence. We did not inspect partially correct answers (e.g. correct colour with incorrect number). The binary correct/incorrect sentence report scores per trial were analysed using generalised linear mixed models (GLMMs) via the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) in the *R* environment (R core team, 2013). For GLMMs, odds ratios and confidence intervals are reported. An odds ratio of 1 means that no effect is present. The further an odds ratio deviates from 1, the larger the size of the effect. Confidence intervals that do not cross 1 are significant.

Results

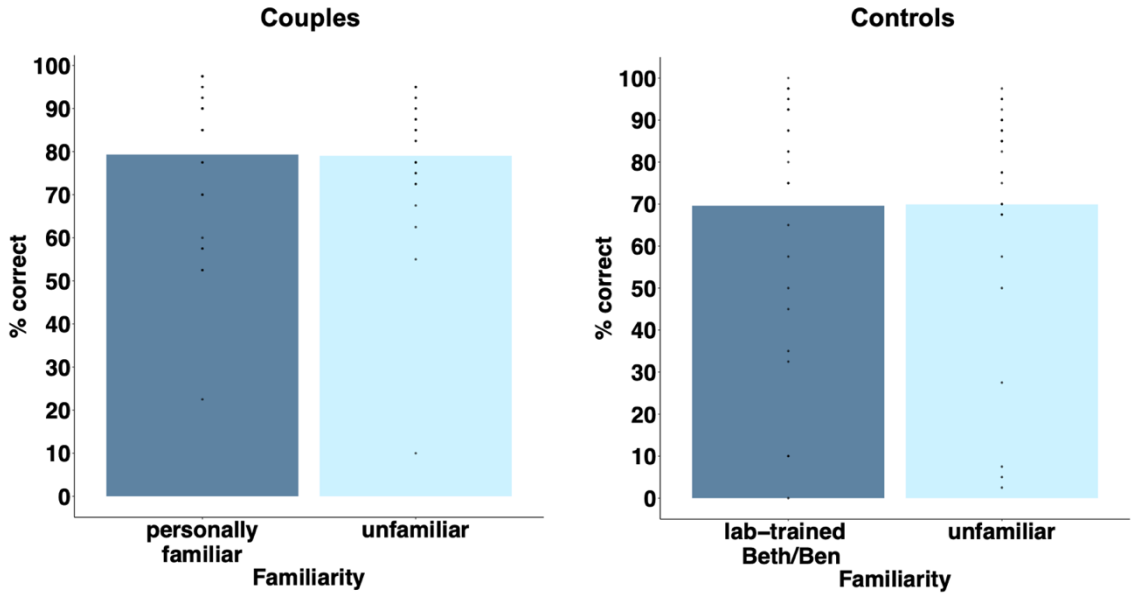

Figure 5. Bars display mean accuracy for the speech intelligibility task (Task 3) as a percentage for the personally-familiar and unfamiliar (couples) or lab-trained and unfamiliar (controls) identities. Points represent individual participants' scores for each identity. Please see the online version of this article for colour figures.

Data from five couples group participants (and the corresponding members of the control group) were removed for failing the attention checks (i.e. scoring less than 6/8 on vigilance trials). Thus, 26 participants per group were retained for the statistical analyses.

Couples

In order to investigate the effect of familiarity on speech perception accuracy, a binomial GLMM was constructed. In this confirmatory analysis, the outcome measure was the binary correct/incorrect sentence report score on each trial. Familiarity was defined as a fixed effect; participant and voice identity were entered as random effects. Statistical significance was established by comparing the full model that included the fixed and random effects, to a reduced model. The comparison of the full model to the reduced model was not significant ($\chi^2(1) = .085$, $p = .771$), indicating that accuracy was similar for personally-familiar (mean = 79.3%, SD = 18.9%) versus unfamiliar (mean = 79.0%, SD = 17.4%, OR= 0.86, CI = [-1.24, 0.90]) voices (see Figure 5). Against our prediction, we therefore did not find a familiarity benefit in our task.

Controls

If we assume that enhanced speech intelligibility in our study reflects relative familiarity with a voice, rather than variations in the acoustic clarity of some talkers, then any observed personal familiarity advantage for speech perception should be at least as large as that seen in the control group (for whom the familiar voice in this task is lab-trained).

We used a binomial GLMM to examine whether lab-trained familiarity (here, using the "Beth/Ben" voice only) had an effect on participants' accuracy for sentences in background four-talker babble. The full and reduced models were constructed in the same way as

described for the couples group. Statistical significance was established by comparing the full model that included the fixed and random effects, to a reduced model that did not contain familiarity. We found that the comparison of the full model to the reduced model was not significant ($\chi^2(1) = .007$, $p = .933$; see Figure 6). Thus, there was no speech perception benefit for the lab-trained identity (mean = 69.6%, SD = 30.3%) compared to the unfamiliar voice (mean = 69.9%, SD =28.5%, OR = 1.04, CI = [-0.91, 0.94).

Discussion

The third task in this study explored whether personal familiarity with a voice could provide advantages for recognising speech in background noise. Despite previous research finding familiar voice benefits for speech intelligibility, we saw no difference in accuracy when couples group participants heard their partner's voice against four-talker babble, compared to hearing an entirely novel speaker.

There are several possible explanations as to why we did not find a familiar voice benefit. One possibility is that our choice of the type of masker and the relative loudness of the target voice (i.e. the signal-to-noise ratio; SNR) may have affected our results. Although previous studies have used both similar maskers (i.e. multi-talker babble) and similar SNRs, it has been shown recently that the type of masker can affect the size of the familiarity benefit (Holmes & Johnsrude, 2020). Another possibility for this null result may be that the task in the current study was relatively easy for this particular sample, as on average, performance was ~80% for recognising both personally-familiar and unfamiliar speech in noise. In much of the previous research reporting familiar voice benefits, accuracy for unfamiliar targets tend to be moderate (i.e. ~40-65%; Johnsrude et al., 2013; Holmes, Domingo, & Johnsrude, 2018; Levi, Winters, & Pisoni, 2011). Furthermore, there is evidence that certain intelligibility-enhancing cues in stimuli themselves (e.g. dynamic video) are optimally effective for auditory speech at intermediate levels of background noise (e.g. Ma, Zhou, Ross, Foxe, & Parra, 2009; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007). That said, substantial familiar-voice benefits have been observed in the presence of high-accuracy speech recognition for the familiar voice (Nygaard, Sommers, & Pisoni, 1994) and while 80% accuracy is high, it is not perfect performance. Thus, while aspects of the experimental design and the stimuli may have affected our results, this null effect is nevertheless surprising, given the previous literature and the large effects of familiarity we observed in the other tasks reported here.

General Discussion

In the current study, we examined whether and how voice perception from personally-familiar voices may differ from lab-trained voices. We investigated this question across 3 challenging listening tasks. In the first task, we showed that listeners are able to recognise personally-familiar voices with much higher accuracy than lab-trained voices from vocalisations that included only minimal linguistic content ("uh", "uhm"). In the second task, we observed that personally-familiar voices can be recognised with much higher accuracy than lab-trained voices, even in the face of perceptually-disruptive acoustic manipulations to the voices. In the final task, we aimed to extend our findings of familiarity advantages from voice identity perception to speech perception. While we observed high accuracy on the speech perception task, we found no apparent familiarity advantage.

Our study is one of the first to directly compare a range of voice perception tasks, using both non-verbal and verbal stimuli, for voices of differential familiarity. We highlight that voice

identity perception from personally-familiar voices is indeed a robust and highly accurate process, even in the face of perceptual challenges that substantially decrease accuracy for lab-trained voices. The results further underline that 'being familiar' with a voice can have many different meanings. Listeners in our study were undoubtedly familiar with the lab-trained voices, as they were able to recognise and name the voices with high accuracy after familiarisation, and showed above-chance recognition in all conditions of the voice identity tasks (see Supplementary Materials). However, for the couples group, performance in both voice identity perception tasks was significantly lower for the lab-trained identity than for the personally-familiar voice. Thus, our study shows that this kind of lab-based familiarity may not be sufficient to support identity recognition in challenging listening situations that require listeners to generalise beyond what they have learned during training: Given the relatively low performance on both identity tasks, we would argue that our participants largely failed to achieve generalisation for lab-trained voices. This is perhaps unsurprising, and fits well with previous research showing that generalisation is challenging for lab-trained voices (e.g. judging identity across languages: Winters, Levi & Pisoni, 2008).

We therefore argue that the human ability to accurately process voice identity has indeed been partially underestimated in the literature by the use of lab-trained and celebrity voices as proxies for personal familiarity. We show conclusive evidence that personal familiarity affords substantial advantages over lab-trained familiarity: when hearing the romantic partner's voice, identity perception was virtually error-free for a challenging filler task and was robust to the interference of acoustic manipulations. We note here, however, that in our study these perceptual benefits did not translate to understanding speech in noise, despite this effect having been replicated several times in the existing literature. We suggest that future work might be able to resolve the task conditions under which familiarity advantages for speech intelligibility are greatest, particularly with regard to the combination of stimulus type, task (e.g. open- vs closed-set recognition), masker type (noise, one talker, multiple talkers), and masker level/SNR.

For identity recognition, we propose that the pattern of results we have observed can be explained by differences in participants' experience with the personally-familiar and lab-trained voices, and their resultant mental representations of those voices. For the lab-trained voices in our study, listeners had formed a mental representation that was relatively rigid - they had only encountered the voice in the context of the familiarisation training in this experiment, and from a particular set of stimuli. Generalisation from such a relatively under-specified representation is difficult (Lavan, Knight, Hazan, & McGettigan, 2019). In contrast, for the personally-familiar voices, listeners should have encountered the voice in many different contexts, and thus formed a well-rounded and robust representation. Based on this kind of varied exposure, listeners in Task 1 may well have remembered what their partner's conversational filler sounds are like from previous conversations, such that no generalisation was necessary for this task. Intriguingly, however, we also find evidence for a much better ability to generalise to truly unheard stimuli for personally-familiar voices: In Task 2, acoustic manipulations were applied that pushed both the lab-trained and personally-familiar voices outside their typical acoustic repertoire, at times going beyond anatomical constraints of the speaker's vocal tract (i.e. increasing/reducing the apparent vocal tract length). Nonetheless, listeners were able to perform this task well for the personally-familiar voice. Listeners may thus either be able to make use of residual diagnostic voice information in modulated stimuli (e.g. speech rate, pronunciation of certain phonemes), or to partially generalise from their robust mental representation of the personally-familiar voice to the acoustically manipulated - and truly novel - portrayal of it.

Our results align well with recent theoretical models outlining a mechanistic account of familiar vs unfamiliar voice identity perception (Maguinness et al., 2018, see also Lavner et al., 2001). These models propose that all voices are processed in relation to the mental representation of a (context-relevant) prototypical voice. Upon hearing a voice, listeners identify the differences between the perceived voice and the prototypical voice. If a voice is familiar, the extracted 'deviant' acoustic features are compared to an existing, stored reference pattern of 'deviant acoustic features' for a familiar voice (i.e. a mental representation of that familiar voice identity). If the deviant acoustic features of the perceived voice are sufficiently similar to the familiar voice's stored reference pattern of deviant acoustic features, the voice identity is recognised. In this model, novel voices are learned via iterative exposure and the consequent computation of the deviant acoustic features relative to the prototype. Eventually, the deviant acoustic features establish a stored reference pattern, opening up a route to familiar voice identity recognition. While initial reference patterns may be established after only brief exposure, they can be rendered more robust and, crucially, more flexible with each new exposure to the voice. In our study, we show that while our lab-trained voices were recognised with good accuracy after a limited amount of training, the reference patterns were not sufficiently developed to enable recognition from challenging, previously unheard vocalisations. For personally-familiar voices in our study, reference patterns were in contrast much better established, yielding high recognition accuracy in the face of considerable perceptual challenges. It is unclear how the potential for flexibility and robustness is encoded in these reference patterns: Although there is some evidence that mental representations of voices are based on the average of the acoustic input a listener has experienced (Lavan, Knight & McGettigan, 2019), this coding principle on its own cannot readily explain the flexibility of representations of highly familiar voices. Further work is therefore required to shed light on what kind of information is stored in mental representations (or reference patterns) of familiar voices, and how these are formed (e.g. Lavan, Burton et al., 2019).

There are also open questions about how much and what kind of exposure is needed for the establishment of a reference pattern that is as robust as we have observed it for personally-familiar voices. We acknowledge that the familiarisation phase in the current study was somewhat brief, leading to familiarity with the lab-trained voices that was relatively superficial (but, arguably, broadly representative of the degree of familiarity of lab-trained voices apparent in other studies). However, it is possible that listeners may be able to build much more robust representations of lab-trained voices through using different training protocols. Given sufficiently extensive training, it should be possible to approximate the familiarity acquired through personal experience with voices, thus blurring the distinctions between the two kinds of familiarity pitted against one another in the current study. Similarly, we also chose a particular type of personally-familiar voice - a romantic partner. Not all voices with which we are personally-familiar will be underpinned by highly robust representations: Just like lab-trained voices, the degree of familiarity with personally-familiar voices is bound by the context(s) the voices were experienced in, as well as other factors such as our feelings toward their owners. This is illustrated by a previous study of voice identity discrimination: Lavan and colleagues (2016) operationalised personal familiarity by using the voices of university lecturers as stimuli and measuring voice identity perception in students who had recently been taught by these lecturers. These students were able to make overall more accurate identity judgements compared to students who did not know the lecturers. However, when required to generalise their knowledge of the voices across different vocalisations, or to less familiar vocalisations (e.g. spontaneous laughter), their accuracy declined to the same degree as the accuracy of students who were unfamiliar with the voices. These findings can therefore be interpreted as showing the limits of this particular

kind of personal familiarity, where the voices were encountered largely in very specific contexts (i.e. an educational setting).

Given the potentially overlapping nature of different types of familiarity, we suggest that familiarity may be better conceptualised as a continuum, determined by the amount and variability of exposure. Beyond this, a wealth of other features of voices that may be specific to the listener (e.g. the emotional salience of a particular voice) or specific to the voice (e.g. vocal distinctiveness) may contribute to the particular make-up of how familiar voices are processed.

Overall, our findings contextualise the accuracy of voice identity perception, and frame it as a function of the relative familiarity with a voice: Accuracy for identity perception from voices that we have limited experience with is in general far from perfect, especially when comparing performance to face perception in broadly comparable tasks (Barsics, 2014). Since our experience with most voices is, in the end, bound up in a finite set of specific contexts, it therefore seems prudent to assume that voice identity perception in challenging listening situations is error-prone. Discussions around the use of voice identity judgements in forensic contexts have already frequently highlighted the limited validity of, for example, earwitness identifications (Smith et al., 2019; Cantone, 2010). Our study, however, highlights that, over and above individual differences in voice perception (Aglieri et al., 2017; Mühl, Sheil, Jarutytė, & Bestelmeyer, 2018; Jenkins et al., 2020), voice identity judgements for personally-familiar voices that we truly know can be remarkably reliable, with dramatically reduced errors and near-perfect accuracy being apparent.

References

Aglieri, V., Watson, R., Pernet, C., Latinus, M., Garrido, L., & Belin, P. (2017). The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices. *Behavior Research Methods*, *49*(1), 97-110.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*(1), 388–407.

Baker, R., & Hazan, V. (2011). DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, *43*(3), 761-770.

Barsics, C. G. (2014). Person recognition is easier from faces than from voices. *Psychologica Belgica, 54*(3), 244-254.

Bates, D., Maechler, M., Bolker, B. & Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. *J. Stat. Softw*. 67, 1–23.

Boersma, P. & Weenink, D. (2018). Praat: doing phonetics by computer [Computer program]. Retrieved from: http://www.fon.hum.uva.nl/praat/

Bolia, R. S., Nelson, W.T., Ericson, M.A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America, 107*(2), 1065-1066.

Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America*, *40*(6), 1441-1449.

Cantone, J. A. (2010). Do you hear what I hear: Empirical research on earwitness testimony. *Tex. Wesleyan L. Rev.*, *17*, 123-142.

Chan, D., Fourcin, A., Gibbon, D., Granstrom, B., Huckvale, M., Kokkinakis, G., Kvale, K., Lamel, L., Lindberg, B., Moreno, A., Mouropoulos, J., Senia, F., Trancoso, I., Veld, C., and Zeiliger, J. (1995). "EUROM—A spoken language resource for the EU," Eurospeech'95, in *Proceedings of the 4th European Conference on Speech Communication and Speech Technology, Vol. 1*, pp. 867–870.

Chauhan, V., & Gobbini, M. I. (2018). How familiarity warps representation in the face space. *bioRxiv*, 293225. doi**:** https://doi.org/10.1101/293225

Fontaine, M., Love, S. A., & Latinus, M. (2017). Familiarity and voice representation: From acoustic-based representation to voice averages. *Frontiers in Psychology, 8*(1180), 1-9.

Gaudrain, E. (2018). straight_process: A tool to process files with STRAIGHT, GitHub repository, https://github.com/egaudrain/straight_process

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. In *Proceedings of the 10th Annual Conference of the International Speech Communication Association,*

INTERSPEECH (pp. 152–155). Baixas, France: International Speech Communication Association

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

Holmes, E., Domingo, Y., & Johnsrude, I. S. (2018). Familiar voices are more intelligible, even if they are not recognised as familiar. *Psychological Science, 29*(10), 1575-1583.

Jenkins, R., Tsermentseli, S., Monks, C. P., Robertson, D. J., Stevenage, S. V., Symons, A. E., & Davis, J. P. (2020). Are super-face-recognisers also super-voice-recognisers? Evidence from cross-modal identification tasks. *PsyArXiv,* doi*:* https://doi.org/10.31234/osf.io/7xdp3

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.

Kawahara, H., & Irino, T. (2004). "Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation", in *Speech separation by humans and machines,* edited by P.L. Divenyi (Kluwer Academic, Massachusetts), pp. 167 – 180).

Kreitewolf, J., Mathias, S. R., & von Kriegstein, K. (2017). Implicit talker training improves comprehension of auditory speech in noise. *Frontiers in Psychology, 8*(1584), 1-8.

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology, 21*(4), R143-R145.

Lavan, N., Burston, L. F., & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology, 110*(3), 576-593.

Lavan, N., Burston, L.F.K., Merriman, S.E., Ladwa P., Knight, S., & McGettigan, C. (2019). Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology, 72*(9), 2240-2248.

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, *26*(1), 90-102.

Lavan, N., Knight, S., Hazan, V., & McGettigan, C. (2019). The effects of high variability training on voice identity learning. *Cognition*, *193*, 104026.

Lavan, N., Scott, S.K., & McGettigan, C. (2016). Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices. *Journal of Experimental Psychology: General, 145*(12), 1604-1614.

Lavner, Y., Gath, I., & Rosenhouse, J. (2000). The effects of acoustic modifications on the identification of familiar voices speaking isolated vowels. *Speech Communication*, *30*(1), 9-26.

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), 4053-4062.

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: a Bayesian explanation using high-dimensional feature space. *PLoS One*, *4*(3), 1-14.

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia, 116,* 179-193.

Mühl, C., Sheil, O., Jarutytė, L., & Bestelmeyer, P. E. (2018). The Bangor Voice Matching Test: A standardized test for the assessment of voice perception ability. *Behavior research methods*, *50*(6), 2184-2192.

Newman, R. S., & Evers, S. (2007). The effect of talker familiarity on stream segregation. *Journal of Phonetics*, *35*(1), 85-103.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics, 60*(3), 355-376.

Nygaard, L. C., Sommers, M. S., Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*(1), 42-46.

R Core Team. R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing, Vienna, Austria, 2013). http://www.Rproject.org/.

Rosen, S., Souza, P., Ekelund, C., & Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *The Journal of the Acoustical Society of America*, *133*(4), 2431-2443.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17*(5), 1147-1153.

Schweinberger, S. R., Herholz, A., & Sommer, W. (1997). Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language, and Hearing Research*, *40*(2), 453-463.

Smith, H. M., Bird, K., Roeser, J., Robson, J., Braber, N., Wright, D., & Stacey, P. C. (2019). Voice parade procedures: Optimising witness performance. *Memory,* 1-16.

Souza, P., Gehani, N., Wright, R., & McCloy, D. (2013). The advantage of knowing the talker. *Journal of the American Academy of Audiology, 24*(8), 689-700.

Stevenage, S. V., Symons, A. E., Fletcher, A., & Coen, C. (2020). Sorting through the impact of familiarity when processing vocal identity: Results from a voice sorting task. *Quarterly Journal of Experimental Psychology, 73*(4), 519-536.

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, *17*(1), 3-28.

Winters, S. J., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languages. *The Journal of the Acoustical Society of America*, *123*(6), 4524-4538.

Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*(7), 2064-2072.

Zarate, J. M., Tian, X., Woods, K. J., & Poeppel, D. (2015). Multiple levels of linguistic and paralinguistic features contribute to voice recognition. *Scientific Reports, 5*(11475), 1-9.