"Please sort these voice recordings into 2 identities": Effects of task instructions on performance in voice sorting studies

Nadine Lavan[1,2], Siobhan E. Merriman[2], Paayal Ladwa[2], Luke F.K. Burston[2], Sarah Knight[1] and Carolyn McGettigan[1]

[1] *Department of Speech, Hearing and Phonetic Sciences, University College London*

[2] *Department of Psychology, Royal Holloway, University of London*

Correspondence to:

Nadine Lavan, Department of Speech, Hearing and Phonetic Sciences, University College London, 2 Wakefield Street, London WC1N 1PF, United Kingdom. E-mail: n.lavan@ucl.ac.uk

or

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

Acknowledgements: This work was supported by a Research Leadership Award from the Leverhulme Trust (RL-2016-013) awarded to Carolyn McGettigan

Abstract

We investigated the effects of two types of task instructions on performance of a voice sorting task by listeners who were either familiar or unfamiliar with the voices. Listeners were asked to sort 15 naturally varying stimuli from two voice identities into perceived identities. Half of the listeners were to sort the recordings freely into as many identities as they perceived; the other half were forced to sort stimuli into two identities only. As reported in previous studies, unfamiliar listeners formed more clusters than familiar listeners. These listeners perceived different naturally varying stimuli from the same identity as coming from different identities, while being highly accurate at telling apart the stimuli from different voices. We show that a change in task instructions – forcing listeners to sort stimuli into two identities only – helped unfamiliar listeners to overcome this selective failure at "telling people together". This improvement, however, came at the cost of an increase in errors in telling people apart. For familiar listeners, similar non-significant trends were apparent. Therefore, even when informed about correct number of identities, listeners may fail to accurately perceive identity further highlighting that voice identity perception in the context of natural within-person variability is a challenging task. We discuss our results in terms of similarities and differences to findings in the face perception literature and their importance in applied settings, such as forensic voice identification.

Keywords: person perception, voice identity, voice sorting, within-person variability

Introduction

Voice identity perception in the context of variable stimuli has been shown to be a challenging task for listeners (see Lavan, Burton, Scott & McGettigan, 2018 for a review) because talkers constantly adjust what they are saying, and how they are saying it, for different surroundings and audiences. These adjustments introduce within-person

variability in the sound of the voice, such that the voice can vary dramatically between different contexts – for example, what we sound like when we shout over background noise compared to when we mutter to ourselves while trying to solve a problem. In order to perceive voice identity successfully, listeners therefore need to be able not only to perceive different voices as indeed coming from different people: they also need to generalise across this within-person variability in the sound of the voice to link different sounding samples of a single voice to the same person. This is especially challenging when dealing with voices that are unfamiliar to listeners. Studies have shown that listeners are less accurate in discriminating speakers when presented with variable sounding examples of unfamiliar voices. Speaker discrimination accuracy has been shown to decrease when making judgements across disguised and undisguised voices (Reich and Duke, 1979, see also Eriksson, 2010; Hollien, Majewski & Doherty, 1982; Schlichting & Sullivan, 1997), sung and spoken words (Peynircioğlu, Rabinovitz, & Repice, 2017), different vocalisations (Lavan, Scott, & McGettigan, 2016), and recordings produced in different languages (Wester, 2012). Although familiar voice identity perception is generally more reliable, listeners can nonetheless be affected by within-person vocal variability. Lavan et al. (2016) have shown that, whereas listeners who are familiar with the voices (here: “familiar listeners”) systematically outperform listeners who are unfamiliar with the voices (here: “unfamiliar listeners”) on a speaker discrimination task, performance nonetheless decreases in the presence of extensive within-person variability (laughter vs. vowels sounds). Similarly, Wagner and Köster (1999) report that familiar voices are essentially unrecognisable when produced using falsetto voice.

Theoretical accounts model familiar and unfamiliar voice identity perception as at least partially distinct processes. Maguinness, Roswandowitz and Von Kriegstein (2018) highlight how familiar and unfamiliar voice identity perception may dissociate on a neural

and behavioural level. Kreiman and Sidtis (2011) propose that unfamiliar voices are processed – at least in experimental tasks – based on their acoustic and perceptual features in a stimulus driven way. According to this view, in a situation that would require listeners to discriminate explicitly between two stimuli, this may take the shape of close comparisons between perceptually salient features of the stimuli until a decision is reached as to whether the two stimuli are likely to be produced by the same person or two different people. In contrast to this, familiar voices are thought to be processed in a holistic way, based on the detection of perceptual features that are diagnostic for a familiar identity. These features may take the form of a particular quality of voice or characteristic speaking style that allow us to detect voices as being familiar – notably, the nature of these features may vary from voice to voice, and from listener to listener (Kreiman & Sidtis, 2011). Through the detection of such diagnostic features of a familiar voice, vocal signals produced by familiar speakers are thought to be matched to representations of the specific speaker's vocal inventory that are stored in long-term memory.

More empirical evidence for dissociations between familiar and unfamiliar voice identity perception has recently been reported for studies using voice sorting paradigms (Lavan, Burston & Garrido, 2018; Lavan, Burston, Ladwa, Merriman, Knight & McGettigan, 2019). In such voice sorting paradigms, listeners are exposed to naturally varying recordings of two voice identities – specifically, recordings from dialogues extracted across different scenes and episodes of TV shows. Each voice identity is represented by a number of stimuli. Participants, who are either familiar with the TV show or not, are then asked to sort the recordings into voice identities. Although familiar listeners sort the sounds into a relatively small number of voice identities, unfamiliar listeners tend to perceive a much larger number. An analysis of the composition of the perceived voice identities (i.e. examining the patterns of stimuli within them) reveals that both groups are highly accurate

at “telling people apart” – that is, listeners are able to distinguish between different identities despite the within-person variability included in the stimuli. This is evidenced by the very small number of mixing errors, which arise when listeners sort stimuli from different identities into one perceived identity. Unfamiliar listeners, however, fail to “tell people together”, because they split naturally varying stimuli produced by one person into a larger number of perceived identities. Although familiar listeners also make some “telling people together” errors, they tend to be able to perceive variable stimuli produced by the same voice as belonging to one identity only, despite obvious acoustic and perceptual discrepancies between the stimuli. It has been proposed that this failure in “telling people together” is due to unfamiliar listeners having no person-specific representation of the voice that would allow them to know *how* a novel voice varies and thus recognise within-person variability appropriately. The subjective experience for listeners when exposed to naturally varying stimuli of unknown voices is therefore by necessity mostly driven by the perceptual differences between stimuli, with within-person variability perceived as between-person variability.

This selective failure of “telling people together” when participants are unfamiliar with a person is not limited to the voice identity perception literature, but has also been reported in the face perception literature (Jenkins, White, Van Montfort & Burton, 2011). Although this finding seems to be highly consistent, Andrews, Jenkins, Cursiter, and Burton (2015) have shown that participants who are unfamiliar with the faces used in the study can readily overcome this issue. The authors changed the task instructions for their face sorting tasks from allowing participants to sort the stimuli into any number of perceived identities without any restrictions, to only allowing participants to create two perceived identities. If participants were still struggling with accurate identity perception in the context of natural within-person variability, many “mixing errors” could be expected, where

participants group stimuli from 2 different identities into the same perceived identity. The authors, however, report almost perfect performance, with most participants arriving at the correct solution. The change in task instructions thus allowed participants to perceive within-person variability in stimuli of the same person as such.

In the current study, we explored how task instructions (free versus restricted sorting) may affect performance of familiar and unfamiliar listeners in a voice sorting task (see Andrews, Jenkins, Cursiter & Burton, 2015 for faces). First, in a free sorting task, we instructed participants to sort 30 brief recordings (2 identities x 15 stimuli each) into identities, without giving any guidance as to how many identities may be present (for this, some of the data collected for Lavan et al. 2019 were reused; please see the Methods section for details). For the second type of task instruction, we ran a restricted sorting task: here, participants were told that only 2 identities were present and instructed to sort the recordings accordingly into 2 clusters. We predicted that, for the free sorting task, unfamiliar listeners would form more clusters than familiar listeners by failing to “tell people together”, that is, by perceiving different stimuli from the same voice identity as different voice identities due to the natural within-person variability included in the stimuli (Lavan, Burston & Garrido, 2018, Lavan et al., 2019, see also Andrews et al., 2015; Jenkins et al., 2011; Redfern & Benton, 2017; Zhou & Mondloch, 2016 for faces). Based on previous voice sorting studies, we furthermore predicted that there would be only few errors for “telling people apart”. For restricted sorting, we predicted that errors in “telling people apart” and “telling people together” should be reduced for both groups.

Methods

Participants

The present study re-used data from a previous experiment (N = 26 data sets; Lavan et al., 2019). We also tested an additional 84 new participants for the purpose of the current experiment. Participants were recruited via Prolific.ac, social media, the participant pool of the Department of Psychology at [institution removed], and survey exchange services, such as surveytandem.com. Those recruited via Prolific and via the participant pool at [institution removed] were compensated for their time. The study was approved by the local ethics committee. All listeners were native speakers of English, aged between 18 and 40 years old, and did not report any hearing difficulties. We did not collect any data on their specific accent background and our sample is thus likely to have included people with a range of accents. We recruited familiar and unfamiliar listeners: if participants reported having watched at least one season of *Breaking Bad*, they were classed as familiar listeners. Participants who reported not having seen any episodes of the TV show were assigned to the unfamiliar group. For the data specifically collected for this study, a number of participants were excluded based on the following criteria: 4 participants who had not seen a full season of *Breaking Bad* were excluded; 2 unfamiliar listeners were excluded because they reported having recognised one of the voices (Bryan Cranston) from other TV shows; 8 further listeners were excluded as they created more clusters in the restricted sorting condition than instructed; 4 listeners failed the attention check (see Materials; these listeners did not sort 2 stimuli of a computer-generated voice into a separate cluster); finally, 1 listener was excluded because they moved less than 80% of the sounds on the slide (i.e. did not appear to engage with the task, see Procedure).

In total, 80 participants were thus included in the final sample: 20 familiar listeners (mean age: 23.1 years, SD: 3.6 years, 14 female) and 20 unfamiliar listeners (mean age: 23.0 years, SD: 4.4 years, 13 female) for the free sorting manipulation, and a further 20 familiar listeners (mean age: 26.3 years, SD: 5.0 years, 7 female) and 20 unfamiliar listeners (mean age: 27.6 years, SD: 6.3 years, 14 female) for the restricted sorting manipulation. The 26 re-used data sets comprised 13 participants for each listener group in the free sorting manipulation: these are the participants who, in Lavan et al. (2019), were first presented with the sorting task used in the current experiment (as opposed to being presented with it after having already completed another sorting task), thus having an almost identical task experience to the newly recruited participants.

Materials

The materials used in this experiment are the same as the ones used for the “low expressiveness” condition in Lavan et al. (2019). Audio clips (duration between 1.2 and 4 seconds) from two of the prominent characters of the TV show *Breaking Bad* (Hank Schrader and Walter White) were extracted. The clips contained meaningful utterances with only minimal background noise and there was no interference from other voices. No iconic catchphrases or otherwise diagnostic linguistic information (e.g. referring to a character’s job, etc.) were present in the stimuli. Other than featuring emotionally neutral speech, the sound clips thus varied naturally in all features (e.g. speaking style, verbal content, speaking environment, and conversation partners). Stimuli were normalized for peak amplitude (to 0.400 Pa), long silences were cut, and all sounds low-pass filtered at 10kHz (using a Hann pass-band filter with upper and lower edges 0Hz and 10000Hz, smoothing 20Hz) using Praat (Boersma & Weenink, 2018) to account for systematic differences in the audio quality.

The 30 stimuli (2 identities [Hank, Walter] x 15 stimuli) were embedded into a Microsoft Powerpoint slide. Two identical stimuli spoken by a synthetic female voice, saying “Hello. My name is Sarah” were also included as attention checks to verify that participants were completing the task correctly (i.e. by forming a single identity cluster for the 2 female voice stimuli on each slide; see exclusion criteria). Each embedded sound was represented by a number on the screen. These numbers were evenly distributed across the slide, with no clusters being obvious from the outset.

Procedure

Participants downloaded the Powerpoint slide from the testing platform. Each participant downloaded the same original slide. Participants were told to sort the stimuli into identity clusters by dragging and rearranging them on the slide. After forming all of the clusters, participants were encouraged to outline each cluster for ease of interpretation by the experimenter. For the free sorting task, participants were instructed to sort the stimuli into clusters, so that each cluster included the stimuli produced by a single speaker, and thus represented a perceived speaker identity. Therefore, participants could form any number of clusters (up to the total number of stimuli presented). For the restricted sorting task, participants were told that there were three speakers (i.e. 2 voices from *Breaking Bad* plus the female voice) present in the task and that they should therefore sort the 32 stimuli into 3 clusters. There was no limit on how many times participants could play the sounds. Participants completed the task online via Qualtrics.

Results

Confirmatory analysis: Number of clusters in the free sorting task

PLEASE INSERT FIGURE 1 HERE

Initially, we counted the number of clusters formed by each participant by hand for both the free sorting task and the restricted sorting task (after removing the 'catch' stimuli). Figure 1 (right-hand plot) illustrates that all participants included in the restricted sorting task formed two clusters as per the instructions. For the free sorting task, data were not normally distributed in most cases, as indicated by Shapiro-Wilk tests. We therefore used non-parametric tests for the following analyses, which were performed in the R environment using the *coin* package (Hothorn, Hornik, van de Wiel & Zeileis, 2008). Familiar listeners perceived significantly fewer clusters than unfamiliar listeners (Familiar: Mode = 2, Median = 4, Range = 2-9; Unfamiliar: Mode = 10, Median = 10, Range = 4-20; Mann-Whitney *U* test: $Z = 4.17$, $p < .001$). This finding is in line with what has been reported in other face and voice sorting studies: participants who are familiar with the faces or voices used in the studies perceive a lower number of identities than participants who are not familiar with the faces or voices. These analyses are partially based on re-used data that have already been reported elsewhere (Lavan et al., 2019), therefore do not form an independent replication of previous findings. We do, however, note that the additional data collected specifically for the free sorting task in this experiment (data from 7 familiar and 7 unfamiliar listeners) show the same pattern of results (number of clusters for familiar listeners: Median = 2, Range = 2-7; for unfamiliar listeners: Median = 10, Range = 4-20). Due to the relatively small sample size, we have refrained from a statistical analysis of this subset of data.

Confirmatory analysis: The effects of familiarity and instruction type on performance for "telling people apart" and "telling people together"

PLEASE INSERT FIGURE 2 HERE

To assess *how* listeners formed clusters, we created participant-wise response matrices. In these matrices, for each of the 15 stimuli within a single identity, we coded whether each possible pairing between those stimuli was sorted into the same vocal identity (coded as 1) or into a different vocal identity (coded as 0). We then repeated this for the second identity. The group-averaged response matrices are shown in Figure 2a. These matrices are symmetrical across the diagonal and can be conceptually divided into within-identity submatrices, indexing listeners' performance for "telling people together" and across-identity submatrices, indexing listeners' performance for "telling people apart" (see Figure 2b).

To explore the effects of familiarity (familiar vs. unfamiliar) and instruction type (free vs. restricted sorting) on listeners' performance for "telling people together" and "telling people apart", we computed the participant-wise averages of the within-identity and across-identity submatrices respectively (see Figure 2b). Specifically, the average of a single participant's within-identity submatrices comprised that participant's telling together score, while the average of their across-identity submatrices provided their telling apart score. Perfect performance (i.e. forming two clusters of 15 stimuli, with correct assignment of all stimuli to their corresponding identity) would result in an average of 1 for the within-identity submatrices and an average of 0 for across-identity submatrix (for additional details of the analyses, see Lavan, Burston & Garrido, 2018).

PLEASE INSERT FIGURE 3 HERE

We first investigated how instruction type affected performance for each listener group. For familiar listeners, Wilcoxon signed-rank tests indicated that task performance for “telling together” and “telling apart” was not significantly affected by instruction type (telling together: $Z = 1.38$, $p = .168$; telling apart: $Z = 1.92$, $p = .055$). Numerical trends, however, indicate that, although performance improved for “telling people together”, listeners made more errors for “telling people apart” during the restricted sorting task (Figure 3a). For unfamiliar listeners, performance for “telling people together” was dramatically better for the restricted sorting task ($Z = 5.20$, $p < .001$). Conversely, however, errors for “telling people apart” also significantly increased ($Z = 4.63$, $p < .001$).

We then compared task performance for the two listener groups. Unfamiliar listeners’ performance was consistently worse compared to the performance of familiar listeners in the forced sorting task, for both telling people apart ($Z = 2.69$, $p = .007$) and “telling people together” ($Z = 2.90$, $p = .003$). In the free sorting task, unfamiliar listeners performed worse for “telling people together” ($Z = 4.95$, $p < .001$) but comparably to familiar listeners for “telling people apart” ($Z = .62$, $p = .522$). Please note again that the analyses for the free sorting task include re-used data from Lavan et al., (2019). However, we note that the additional data collected specifically for the free sorting task in this experiment (7 datasets of familiar listeners and 7 data sets for unfamiliar listeners) show a similar pattern of results: Telling people apart probabilities were low for familiar listeners (Median = 0.4, Range = 0-0.12) and unfamiliar listeners (Median = 0.03, Range = 0-0.47), indicating good performance in both groups. Similarly, “telling people together” probabilities were high for familiar listeners (Median = 0.77, Range = 0.43-1), indicating good performance, and low for unfamiliar listeners (Median = 0.23, Range = 0.06-0.68), indicating poor performance.”

Exploratory analysis: Cumulative error rates

We conducted exploratory analyses to explore further how the instruction type affected the cumulative error rates across both "telling people together" and "telling people apart". To compute the cumulative error rates, we summed the individual error rates for "telling people together" (1 minus the mean probability of the lower triangle of each within-identity matrix for each participant) and the error rates for "telling people apart" (original telling apart measure) per participant, separately for the restricted and free sorting tasks (maximum cumulative error rate = 1). Within the listener groups, Mann-Whitney *U* tests between the sorting conditions showed that, although cumulative error rates were dramatically lower for unfamiliar listeners for the restricted sorting task ($Z = 4.03$, $p < .001$, see Figure 3b), error rates remained stable for familiar listeners ($Z = .01$, $p = .989$).

Exploratory analysis: Similar patterns of errors across listener groups and instruction types

Figure 2a indicates differences in how confusable the voices were. For example, bands of darker colours in the within-person submatrices indicate that listener groups consistently struggled to assign these particular stimuli to the correct identity. We therefore ran further item analyses to quantify the observation that geometries of responses may be similar across different listener groups and instructions types. Significant correlations would indicate consistencies in the relative difficulty of individual stimuli across listeners groups and task instructions. For this purpose, we computed Kendall's τ_a correlation coefficients for average "telling together" and "telling apart" matrices across instruction type and across listener groups. If Kendall's τ_a was higher than 95% of the chance predictions ($p < 0.05$) obtained by shuffling the values within the comparisons of interest (5000 permutations), we rejected the null hypothesis. These analyses showed that matrices were significantly correlated across instruction types for both "telling apart" and "telling together", for both familiar listeners ("telling together": $\tau_a = .56$, $p < .001$; "telling apart": $\tau_a = .30$, $p < .001$) and unfamiliar listeners ("telling together": $\tau_a = .30$, $p < .001$; "telling apart": $\tau_a = .35$, $p < .001$).

Similarly, matrices for both "telling apart" and "telling together" were significantly correlated when comparing familiar listeners and unfamiliar listeners on the free sorting task ("telling together": τ_a = .26, p < .001; "telling apart": τ_a = .12, p = .001) and the restricted sorting task ("telling together": τ_a = .18, p < .001; "telling apart": τ_a = .48, p < .001). Thus, while manipulating the instruction changes overall performance, especially in the case of unfamiliar listeners, the patterns of responses and errors appear to remain partially stable across instruction type. Strikingly, these patterns also seem to be stable across listener groups, indicating that some of the underlying processes leading to errors may be shared despite the differences in familiarity.

Discussion

In the current study, we investigated how changes in the instructions for a voice sorting task affect familiar and unfamiliar listeners' sorting behaviour. Half of the participants received instructions previously used in other voice sorting tasks: here, listeners were asked to sort voice clips into as many identities as they perceived, without being given any information on how many voices were actually present. Unfamiliar listeners perceived more identities than familiar listeners (unfamiliar listeners: median = 10 identities, mode = 10 identities; familiar listeners median: 4 identities, mode = 2 identities; see also Lavan, Burston & Garrido, 2018, Lavan et al., 2019; Jenkins et al., 2011 for faces). Familiar and unfamiliar listeners were similarly accurate in telling people apart, with only relatively few mixing errors (i.e. assigning stimuli from two speakers to the same perceived identity) occurring for both groups. Unfamiliar listeners, however, selectively failed to "tell people together"; that is, they failed to generalise across the within-person variability and thus misperceived within-person variability as between-person variability. Although performance for familiar listeners was not perfect (cf. Lavan et al., 2018b; Lavan et al.,

2019), errors were significantly less frequent for this listener group compared to unfamiliar listeners.

With a different set of task instructions, the pattern of results changed. For the other half of the participant group, we implemented a restricted sorting task, where we asked listeners to sort the 30 stimuli into 2 identities (plus the identity included as an attention check) and thus forced listeners to “tell people together”. For these task instructions, the selective failure in “telling people together” largely disappeared in a separate group of unfamiliar listeners. The probability that unfamiliar listeners sorted stimuli from the same identity into the correct cluster rose sharply from 0.20 when given the free sorting instructions to 0.70 for the restricted sorting instructions. This drastic improvement in accuracy is, however, likely to be underpinned by different perceptual processes compared to the ones underpinning good performance (in the free sorting task) for familiar listeners. Familiar listeners can access voice identity representations that are likely to include information on how voices vary, thus allowing them to tell people together. Unfamiliar listeners, however, do not have such representations. In the absence of such person-specific representations, we speculate that unfamiliar listeners may have to rely on comparing the properties of a given example to the clusters already made. If a listener identifies that a specific stimulus is less likely to belong to one of the clusters, they are forced to sort this stimulus into the second cluster to follow the “restricted sorting” task instructions. Although our data show that unfamiliar listeners perform dramatically better in the restricted sorting task, these listeners are unlikely to perceive all the stimuli within each cluster as coming from the same identity. A similar improvement to what was observed for unfamiliar listeners was also apparent for familiar listeners – this improvement was, however, statistically non-significant. Although familiar listeners were already performing well on this aspect of the

task for free sorting, the probability of correctly "telling together" the different stimuli increased from 0.73 to 0.83 with the restricted sorting instruction[1].

Crucially, these improvements for both groups came at a cost: unfamiliar listeners made significantly more mixing errors for the "restricted sorting" instructions compared to the "free sorting" instructions. The same pattern was also apparent for familiar listeners, although it was again not statistically significant. These findings align broadly with what has been reported in the face perception literature. For face sorting however, Andrews et al. (2015) do not report a significant increase in mixing errors (i.e. errors in "telling people apart") between free sorting and restricted sorting groups. Finding a significant increase in "telling apart" errors for voices is thus an interesting example of how face and voice identity processing may differ from one another. These differences between modalities may arise due to different processes being at work during face and voice identity processing, reflecting differences in the nature of the stimuli (visual vs. auditory; static vs. dynamic). Because voice identity processing appears to be generally less robust compared to face identity processing (Barsics, 2014; Stevenage, Hugill, Lewis, 2012), these differences could, however, also reflect that mixing errors only occurred very rarely in the face sorting study.

In the current study, exploratory analyses highlighted that, while the "restricted sorting" instructions reduced the overall number of errors made by unfamiliar listeners due to the dramatic decrease in "telling people together" errors, cumulative error rates remained stable for familiar listeners across the two instruction types. For familiar listeners, the

[1] Although familiar listeners generally perform well for "telling people together" in voice (and face) sorting tasks, we note that due to the nature of the tasks, performance may be somewhat inflated for "free sorting": in these tasks, listeners who arrived at the correct solution of forming 2 clusters with one identity in each (N = 7 in the current data set) may have indeed recognised every single stimulus as belonging to the relevant identity. Alternatively, they may have gleaned that only two identities are likely to be present, and thus assigned ambiguous sounds to the more likely cluster in the absence of genuine recognition. Thus, "perfect performance" in the task may reflect a combination of successful voice identity processing and task-specific decision-making strategies.

reduction of errors made in “telling people together” was balanced by the increase in errors for “telling people apart”. An item analysis of the geometries of the response patterns comparing groups and instruction types shed further light on these errors. Here, significant correlations across listener groups and task instructions indicated that patterns of errors are shared for the voice stimuli used in this study. Thus, despite overall differences in accuracy between groups and task instructions, stimuli that were relatively harder to tell together/apart in the free sorting task were also the ones that were difficult to tell together/apart in the restricted sorting task – for unfamiliar and familiar listeners alike. The correlations of geometries therefore reflect the presence of stimulus-specific effects. Within the set of naturally varying stimuli, some are likely to be a relatively bad likeness of a speaker’s voice due to specific acoustic and perceptual features. These stimuli are therefore more difficult for familiar listeners to recognise explicitly (or, alternatively, these stimuli may be more susceptible to misrecognition; see Lavan et al., 2019). In the absence of explicit recognition, familiar and unfamiliar listeners may thus process identity based on the acoustic and perceptual features of a signal within this kind of task (see Kreiman & Sidtis, 2011). These features may, for example make a stimulus of Walter White’s voice stand out as sounding equally dissimilar to the remaining Walter and Hank stimuli. If this is the case, listeners will perceive this stimulus as a single-stimulus identity in the free sorting task and, for the restricted sorting task, will assign the stimulus to “Walter”/Cluster A and to “Hank”/Cluster B with equal probability. Such stimuli are therefore likely to exhibit higher errors rates independent of task instructions and listeners groups, giving rise to similarities in the geometries of the response patterns.

Our study highlights the challenges of voice identity perception in the context of within-person variability. Within-person variability has previously been shown to lead to failures in “telling people together”, mostly for unfamiliar voices. We show that, even when provided

with the correct number of identities, within-person variability can be sufficiently substantial for listeners to assign a stimulus to the wrong identity when forced to make a decision. This is most strongly exemplified by the significant increase in errors for “telling people apart” for unfamiliar listeners given the correct number of identities. This is particularly striking given that the stimuli included rich segmental information (full utterances, including different phonemes), which has in the past been shown to be beneficial for identity judgements (Bricker & Pruzansky, 1966; Schweinberger, Herholz & Sommer, 1997). However, the within-person variability as it has been included in the current stimuli is unlikely to cover each speaker’s full vocal inventory. All stimuli were extracted from relatively neutral utterances and did not, for example, include any whispered speech, laughter or screaming (see e.g. Lavan et al., 2016 for Lavan et al., 2019 for effects of such expressive voices on identity perception).

In the current study, substantial individual differences are apparent within the familiar and unfamiliar listener groups (see Figure 1 and Figure 3). These differences may hint at familiarity not being the only listener characteristic that affects performance. While individual differences in voice identity processing have been previously reported (Aglieri, Watson, Pernet, Latinus, Garrido & Belin, 2018), we have little insight into what may underpin the differences in performance in the current study. It could be imagined that, for example, listeners with more experience in the relevant accent (here: American English) may more readily perceive fine grained idiosyncrasies in the pronunciation used by the two actors, leading to more accurate identity perception (Stevenage, Clarke & McNeill, 2012). Similarly, studies have shown that a listener’s voice identity ability depends on their broader language ability. For example, it has been reported that dyslexics are impaired in voice identity perception tasks due to the impoverished phonological processing associated with dyslexia (Parrachione, Del Tufo & Gabrieli, 2011).

To date, it is, however, not clear what cues listeners may use to perform either our task or voice identity perception more broadly. This question may be particularly complex to address for unfamiliar listeners. As noted above, listeners unfamiliar with a voice lack a person-specific representation, and thus have limited access to a “ground truth”, as is the case when familiar listeners correctly recognise a voice. When perceiving identity (in the context of variability), unfamiliar listeners therefore need to make use dynamically of any meaningful percept they can extract from a given voice. To tell different voices apart, previous research suggests that unfamiliar listeners use relatively low-level acoustic cues, such as the fundamental frequency (related to pitch perception) or the vocal tract length (related to general percepts of voice quality; Baumann & Belin, 2010). Similarly, unfamiliar listeners may make (additional) use of broader perceived person characteristics, such as age, accent or speaking style (Kreiman & Sidtis, 2011). For “telling people together”, unfamiliar listeners appear to struggle with this process if not guided by task instructions – at least when dealing with brief, context-free recordings of voices. There are no comprehensive studies of which cues unfamiliar listeners use in the context of within-person variability. The selective failure to “tell people together” that can be partially remedied through restricting the number of possible identities, however, indicates that cues can be perceptually available to unfamiliar listeners. How listeners transition from being unfamiliar - and thus unable to cope with within-person variability in a signal - to being familiar - and able to access a stable person-specific representation *despite* the variability - remains unclear.

Although our findings are situated within the person-perception literature in the field of cognitive psychology, they have relevance for more applied, forensic contexts. Earwitness testimony has been studied for decades, documenting that auditory memory can be

unreliable in many cases (Clifford, 1980). Despite these conclusions, earwitness testimony and voice line-ups are used in some countries (reviews with best-practice recommendations: de John-Lendle, Nolan, McDougall & Hudson, 2015; McGorrery & McMahon, 2017). For example, our study highlights the importance of task instructions. The eyewitness and earwitness literatures have already demonstrated the importance of unbiased task instructions, such as informing witnesses that the target may be present or absent in a line up (for Broeders & Val Amerlvoort, 2011; Malpasse & Devine, 1981). The current study provides further evidence that if forced to classify recordings as one of two given categories, listeners – especially those unfamiliar with the voices - will make increased “mixing errors”, assigning the same identity to recordings of two separate individuals. Our findings also highlight that identity perception for unfamiliar voices can be unreliable in the presence of within-person variability. Additionally, our findings stress how unreliable identity perception for unfamiliar voices can be in the presence of within-person variability: Even small changes in intonation, voice quality, or speaking style seem to lead listeners to label two recordings of the same voice as coming from separate individuals. In voice line-ups, this kind of variability is almost guaranteed to be present and could therefore lead to a false negative in instances when a target or perpetrator is present (or even false positives, when an unknown voice is selected as familiar).

Overall, voice-sorting tasks, which are readily understandable and intuitive to participants, are useful tools for quantifying voice-identity perception. In prior investigations of perceiving identity from voices, familiarity has often been conflated with task design, such that familiar listeners are tested on recognition and unfamiliar listeners on voice-matching or discrimination tasks. These traditional tasks load differentially on “telling together” and “telling apart”, whereas sorting tasks allow us to explore both simultaneously. In the current study, we have added new insights into the effects of task instruction on

performance, showing that informing listeners of the number of underlying voice identities drastically improves the ability of unfamiliar listeners to “tell people together" – however, these improvements are not sufficient for familiar-like performance.

References

Aglieri, V., Watson, R., Pernet, C., Latinus, M., Garrido, L. and Belin, P. (2017) The Glasgow Voice Memory Test: assessing the ability to memorize and recognize unfamiliar voices. *Behavior Research Methods*, 49(1), 97-110. doi: 10.3758/s13428-015-0689-6

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *The Quarterly Journal of Experimental Psychology*, *68*(10), 2041-2050. doi: 10.1080/17470218.2014.1003949

Barsics, C. G. (2014). Person recognition is easier from faces than from voices. *Psychologica Belgica*, *54*(3), 244-254. doi: 10.5334/pb.ap

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological Research*, *74*(1), 110-120.

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. doi: 10.1007/s00426-008-0185-z

Broeders, A., & van Amelsvoort, A. (2001). A practical approach to forensic earwitness identification: constructing a voice line-up. *Problems of Forensic Sciences*, *47*, 237-245.

Clifford, B. R. (1980). Voice identification by human listeners: On earwitness reliability. *Law and Human Behavior*, *4*(4), 373. doi: 10.1007/BF01040628

de Jong-Lendle, G., Nolan, F., McDougall, K., & Hudson, T. (2015). Voice lineups: a practical guide. *18th Proceedings of the International Congress of Phonetic Sciences. Glasgow, Scotland*, 10-14.

Eriksson, A. (2010) The disguised voice: imitating accents or speech styles and impersonating individuals. In C. Llamas and D. Watt (eds) Language and Identities 86–96. Edinburgh: Edinburgh University Press.

Hollien, H., Majewski, W., & Doherty, E. T. (1982). Perceptual identification of voices under normal, stress, and disguise speaking conditions. Journal of Phonetics, 10, 139-148.

Hothorn T, Hornik K, van de Wiel MA, Zeileis A (2008). “Implementing a class of permutation tests: The coin package.” *Journal of Statistical Software*, 28(8), 1–23. doi: 10.18637/jss.v028.i08

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313-323. doi: 10.1016/j.cognition.2011.08.001.

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons.

Lavan, N., Burston, L. F., & Garrido, L. (2018). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology*. Epub ahead of print. doi: 10.1111/bjop.12348

Lavan, N., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019). Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology*. Epub ahead of print. doi: 10.1177/1747021819836890

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2018). Flexible voices: identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, 26 (1), 90-102.

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. doi: 10.3758/s13423-018-1497-7.

Maguinness, C., Roswandowitz, C., & von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*, *116*, 179-193. doi: 10.1016/j.neuropsychologia.2018.03.039

Malpass, R. S., & Devine, P. G. (1981). Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of applied Psychology*, *66*(4), 482-489. doi: 10.1037/0021-9010.66.4.482

McGorrery, P. G., & McMahon, M. (2017). A fair 'hearing' Earwitness identifications and voice identification parades. *The International Journal of Evidence & Proof*, *21*(3), 262-286. doi: 10.1177/1365712717690753

Perrachione, T. K., Del Tufo, S. N., & Gabrieli, J. D. (2011). Human voice recognition depends on language ability. *Science*, *333*(6042), 595-595. doi: 10.1126/science.1207327.

Peynircioğlu, Z. F., Rabinovitz, B. E., & Repice, J. (2017). Matching Speaking to Singing Voices and the Influence of Content. *Journal of Voice*, *31*(2), 256-e13. doi: 10.1016/j.jvoice.2016.06.004.

Redfern, A. S., & Benton, C. P. (2017). Expressive faces confuse identity. *i-Perception*, *8*(5), 2041669517731115. doi: 10.1177/2041669517731115

Reich, A. R., & Duke, J. E. (1979). Effects of selected vocal disguises upon speaker identification by listening. *The Journal of the Acoustical Society of America*, *66*(4), 1023-1028.

Schlichting, F., & Sullivan, K. P. H. (1997). The imitated voice - a problem for voice line-ups? Forensic Linguistics, 4(1), 148-165.

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. doi: 10.1044/jslhr.4002.453

Stevenage, S. V., Hugill, A. R., & Lewis, H. G. (2012). Integrating voice recognition into models of person perception. *Journal of Cognitive Psychology*, *24*(4), 409-419.

Stevenage, S. V., Clarke, G., & McNeill, A. (2012). The “other-accent” effect in voice recognition. *Journal of Cognitive Psychology*, *24*(6), 647-653. doi: 10.1080/20445911.2012.675321

Wagner, I., & Köster, O. (1999). Perceptual recognition of familiar voices using falsetto as a type of voice disguise. In *Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco* (pp. 1381-1385).

Wester, M. (2012). Talker discrimination across languages. *Speech Communication*, *54*(6), 781-790.

Zhou, X., & Mondloch, C. J. (2016). Recognizing “Bella Swan” and “Hermione Granger”: No own-race advantage in recognizing photos of famous faces. *Perception*, *45*(12), 1426-1429. doi: 10.1177/0301006616662046

Figures

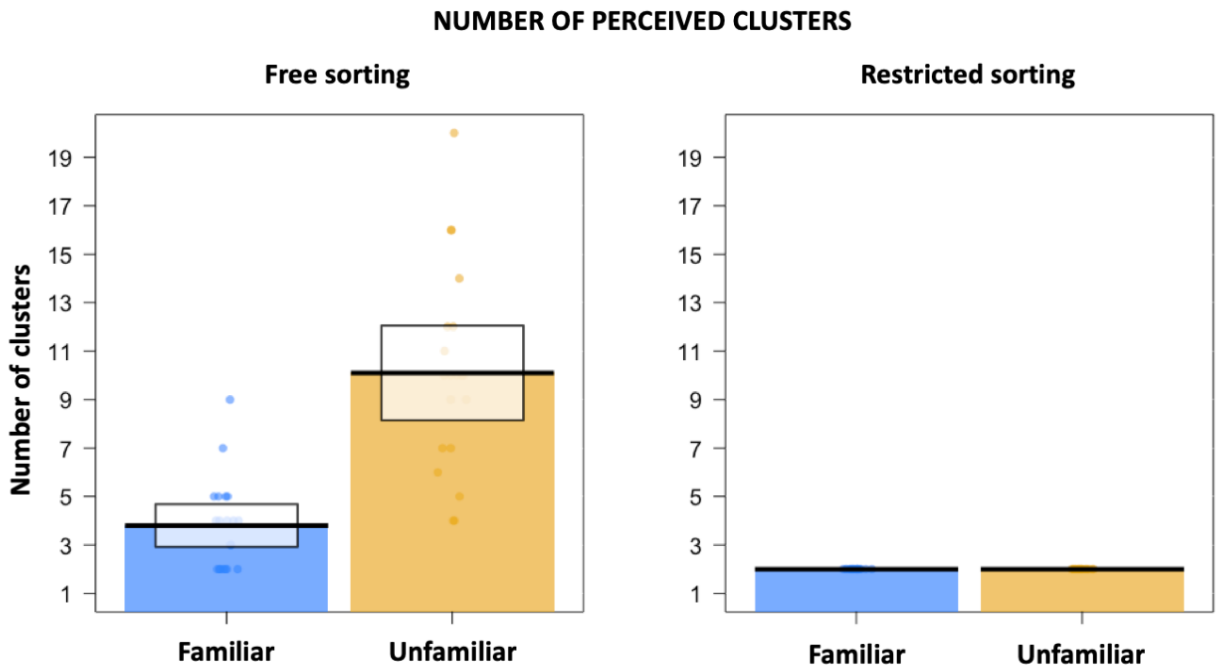

Figure 1 Number of perceived identities for familiar and unfamiliar listeners by instruction type. Bars show the means across participants, and each dot shows one participant. Boxes show the 95% confidence intervals for the means (only shown for the free sorting task, because all listeners followed the instructions for the restricted sorting task and formed two clusters only). Note: these groupings represented the number made after excluding the cluster of the ‘catch’ identities.

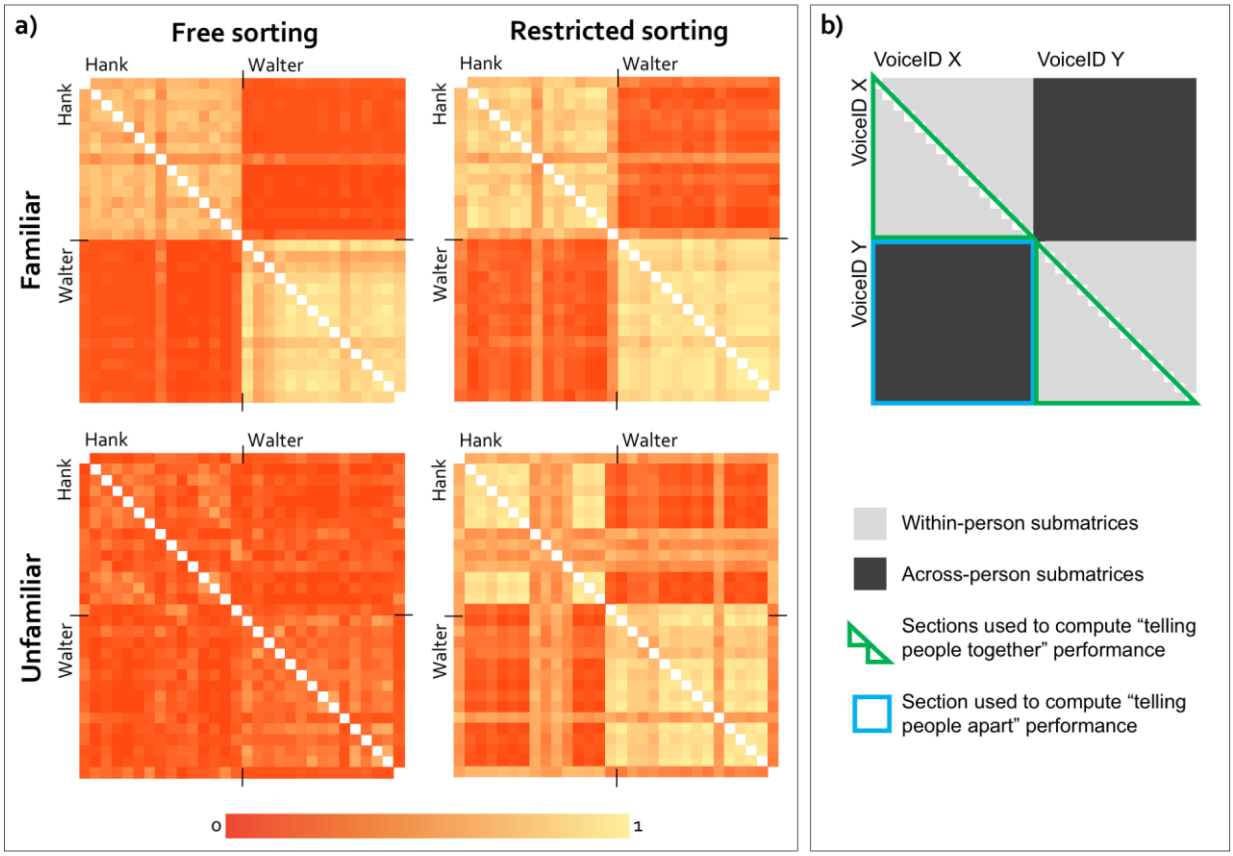

Figure 2 a) Matrices of averaged listeners' responses for the voice sorting task for familiar and unfamiliar listeners. Within these 30 x 30 matrices (15 sound files x 2 identities), each cell shows the probability that two stimuli were grouped within the same perceived identity: cells with a value of 1 indicate that the respective stimuli were always clustered together, cells with a value of 0 indicate that these sounds were never in the same clusters. b) Illustration of the different sections of the per-participant matrices that were analysed below.

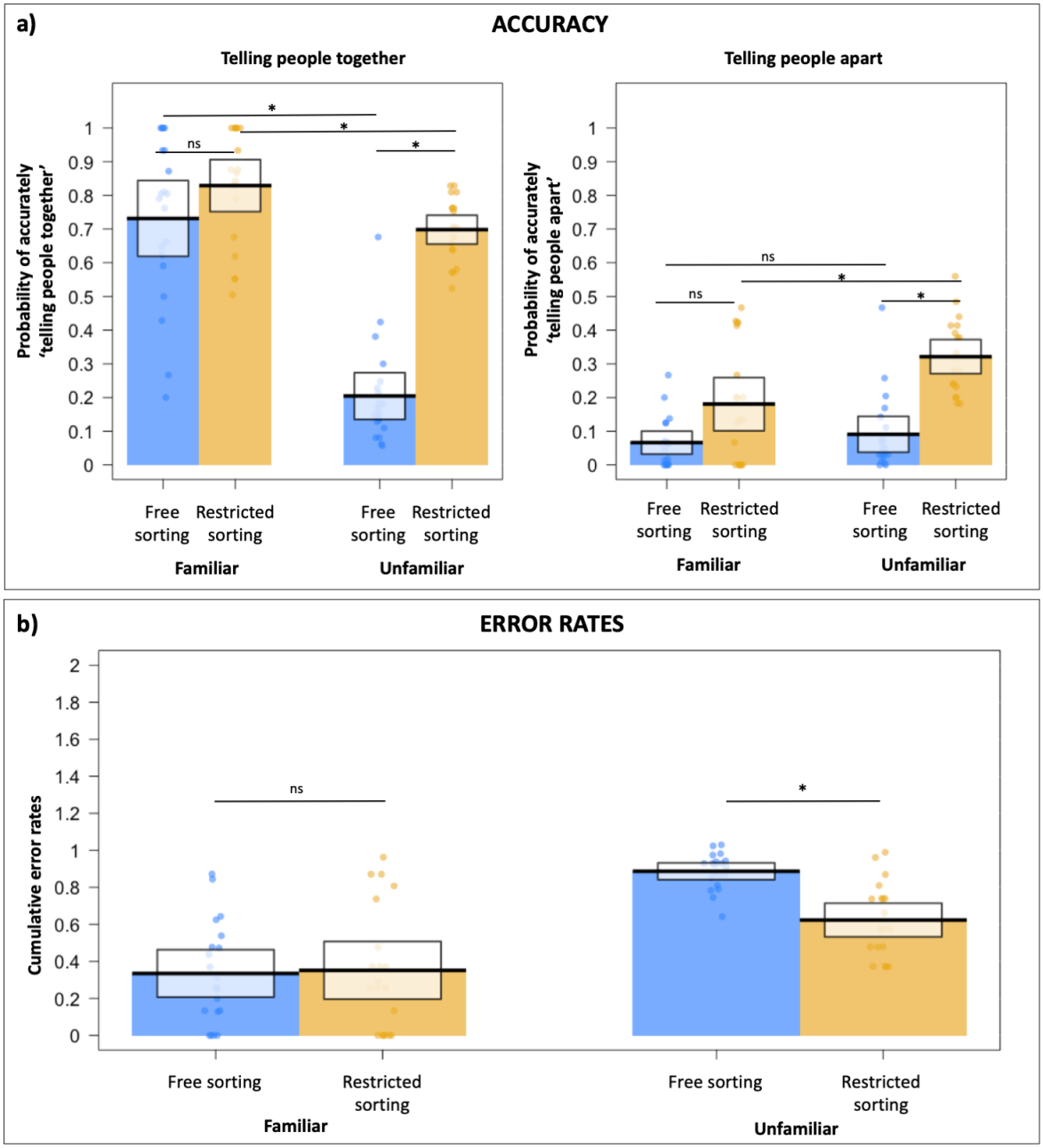

Figure 3 a) Plots of the participants' accuracy for "telling people together" and "telling people apart" by group and task instruction. B) Plots of the cumulative error rates of "telling people apart" and "telling people together" plotted by group and task instruction. Boxes on all plots show the 95% confidence intervals for the means.