How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices.

Nadine Lavan, *Department of Psychology, Royal Holloway, University of London; Division of Psychology, Brunel University;*

Luke F. K. Burston, *Department of Psychology, Royal Holloway, University of London*

Lúcia Garrido, *Division of Psychology, Brunel University*

Address correspondence to: Nadine Lavan, Department of Psychology, Royal Holloway University of London, Egham, Surrey, TW20 0EX, United Kingdom. E-mail: nadine.lavan@rhul.ac.uk

Acknowledgements: This work was supported by a research project grant from the Leverhulme Trust (RPG-2014-392) awarded to Lúcia Garrido. We would like to thank Ibtisam Abdi and Saira Mahmood Khan for help with the data entry and Matthew Longo for comments on a draft of this manuscript.

Abstract

Within-person variability is a striking feature of human voices: our voices sound different depending on the context (laughing vs. talking to a child vs. giving a speech). When perceiving speaker identities, listeners therefore need to not only "tell people apart" (perceiving exemplars from two different speakers as separate identities) but also "tell people together" (perceiving different exemplars from the same speaker as a single identity). In the current study, we investigated how such natural within-person variability affects voice identity perception. Using voices from a popular TV show, listeners, who were either familiar or unfamiliar with the show, sorted naturally-varying voice clips from 2 speakers into clusters to represent perceived identities. Across three independent participant samples, unfamiliar listeners perceived more identities than familiar listeners and frequently mistook exemplars from the same speaker to be different identities. These findings point towards a selective failure in "telling people together". Our study highlights within-person variability as a key feature of voices that has striking effects on (unfamiliar) voice identity perception. Our findings not only open up a new line of enquiry in the field of voice perception but also call for a re-evaluation of theoretical models to account for natural variability during identity perception.

Keywords: voice; identity; variability; familiarity; recognition

Introduction

Voices are highly variable. The same person sounds very different depending on the speaking context: for example, we modulate pitch, speech rate and speaking style depending on whether we are giving a public lecture, talking to a friend, or singing (see Kreiman, Park, Keating & Alwan, 2015; Lavan, Burton, Scott & McGettigan, 2017). Such within-person variability makes identity perception from vocal signals a challenging task: listeners do not only have to tell different voices apart, they also need to generalise percepts of identity across substantial within-person variability to maintain a level of constancy in identity perception (i.e. "telling people together"; see Burton, 2013 for faces). Arguably, being able to "tell people together" can only be reliably achieved for familiar voices (e.g. Jenkins, White, Montford & Burton, 2011 for faces) – we may need to have learned how a specific voice varies to not mistake the substantial inherent within-person variability as between-person variability.

Traditionally, studies of how we recognise people from their voices have explicitly controlled for and thus minimised within-person variability: the experimental stimuli used tend to be carefully selected recordings of vowels, words or short sentences produced in neutral intonation, mostly recorded in a single recording session. This approach has allowed us to gain insights into how we tell people apart via the distinguishing features of individual voices. It has, however, also restricted our understanding of voice identity perception to this particular set of contexts, neglecting the study of the perceptual mechanisms that we use to compute stable and consistent representations of familiar voices despite high within-person variability (Lavan et al., 2017; Lavan, Scott & McGettigan, 2016). Similar issues have recently been highlighted for face identity processing (e.g. Burton, 2013, Burton, Kramer, Ritchie & Jenkins, 2015), opening up a fruitful new line of enquiry in this field.

Below, we first review the few studies that have started investigating how within-person variability and familiarity affect voice identity processing. We then summarise findings from the face perception literature showing striking interactions between familiarity and within-person variability, and we propose that adapting a paradigm from this field (Jenkins et al., 2011) to voices will allow us to shed new light on mechanisms of voice identity processing.

Effects of within-person variability in voice identity processing

Listeners are less accurate when making identity judgements in the presence of within-person variability. For example, listeners are less accurate at correctly matching speakers when they produce pairs of sentences in different languages than in the same language (Wester, 2012; Zarate, Tian, Wood & Poeppel, 2015). Linguistic (dis)similarity of stimuli has also been reported to affect speaker discrimination performance in a top-down fashion: identities can be more accurately discriminated from pairs of stimuli that are semantically or phonetically related, such as 'day-dream' or 'day-bay', than from linguistically unrelated stimuli, such as 'day-bee' (Narayan, Mak & Bialystok, 2016). Similarly, listeners fail to reliably discriminate between unfamiliar speakers when making judgements based on pairs of disguised and undisguised voices (e.g. hypernasal voice vs. neutral voice; Reich & Duke, 1979), across different vocalisations (e.g. vowels vs. laughter; Lavan et al., 2016), and across sung versus spoken words (Peynircioğlu, Rabinovitz, & Repice, 2017). In forensic contexts, studies of the reliability of earwitness' judgements report that listeners' ability to identify a voice from a line up decreases when vocal variability (for example through changes in emotional state) is introduced between study and test (Read & Craik, 1995; Saslove & Yarmey, 1980). Even when listeners are familiar with a voice, they are unable to accurately recognise known

individuals speaking in falsetto voice versus modal ("normal") voice (Wagner & Köster, 1999).

Despite this growing body of literature, current models of voice processing do not explicitly account for within-person variability: Prototype models are often as a theoretical basis to map out how different identities are encoded and how they may relate to each other (Latinus & Belin, 2010; Latinus, McAleer, Bestelmeyer & Belin, 2013; Lavner, Rosenhouse & Gath, 2001; Papcun, Kreiman & Davies, 1989; see also Maguiness, Roswandowitz & Von Kriegstein, 2018). These prototype models however solely focus on *between*-speaker variability, with each identity being conceptualised as a single point in space, neglecting to account for the substantial within-person variability. The reviewed above studies thus reinforce the importance of studying the effects of within-person variability: variability is a key feature of human voices and there is some evidence that it makes voice identity perception less reliable. More empirical evidence is needed, using novel stimuli and tasks,

Effects of familiarity with a speaker on voice perception

Familiarity with a speaker has a profound effect on voice perception. Some authors have even proposed that familiar and unfamiliar voice processing differ fundamentally from each other: In their model of voice identity processing, Kreiman and Sidtis (2011) propose that unfamiliar voice perception relies on the comparison and discrimination of (acoustic) features in a voice (see also Van Lancker & Kreiman, 1987). In contrast to this, familiar voice perception is thought to rely on abstracted processing of identity of a voice's acoustic features, which can be achieved without explicit discrimination. Surprisingly, however, few studies have directly contrasted differences in identity processing for familiar and unfamiliar voices within the same task. To date, a strong

association exists between task type and listener characteristics is present in the existing literature: Studies have either tested voice recognition/identification in the context of familiar voices (for an overview, see Kreiman & Sidtis, 2011) or explored voice discrimination in the context of unfamiliar voices (e.g. Reich & Duke, 1969; Wester, 2012; Zarate et al., 2015),

When directly comparing listener groups who are either familiar or unfamiliar with a set of test voices on a speaker discrimination task, a clear familiarity advantage has been shown (Lavan et al., 2016). Complementary findings have also been reported for speech comprehension: listeners are consistently better at understanding the speech of familiar voices compared to unfamiliar voices (Johnsrude, Casey & Carlyon, 2014; Johnsrude, Mackey, Hakyemez, Alexander, Trang & Carlyon, 2013). Taken together, we can see general processing differences for familiar and unfamiliar voices, with advantages being apparent for extracting information from familiar voices.

Interactions of familiarity and within-person variability: insights from face perception

An issue that has not been extensively explored in the voice perception literature to date is the interaction of familiarity with within-person variability (but see Lavan et al., 2016). For face identity perception, stark differences in the processing of within-person variability for unfamiliar faces compared to familiar faces have been reported. We are able to reliably recognise familiar individuals even under challenging viewing conditions, for example, when images are degraded or include substantial within-person variability (e.g. Yip & Sinha, 2002, Hole, George, Eaves, & Rasek, 2002, Jenkins et al., 2011). With decreasing familiarity, our ability to tolerate such within-person variability also decreases (Burton, Wilson, Cowan and Bruce, 1999; Bruce, Henderson, Newman, & Burton, 2001). For unfamiliar faces, variability across images, such as

changes in viewpoint, expression or lighting, or type of camera results in poor face identity matching and recognition (Bruce et al., 1999; Henderson et al., 2001; Hill & Bruce, 1996; Kemp, Towell & Pike, 1997).

These differences in how we cope with within-person variability in familiar and unfamiliar faces has been attributed to the nature of different representations available for familiar and unfamiliar people (Hancock, Bruce & Burton, 2000; Burton et al., 2015). While viewers have built up a relatively stable and representation of a familiar face that is robust to changes in image properties, no such person-specific representations exist for unfamiliar faces. For unfamiliar faces, viewers are therefore thought to rely more in the visual properties of the specific unfamiliar face. These visual properties vary from image to image, resulting in unreliable perception of identity from unfamiliar faces.

A striking demonstration of the differences in the processing of identity in familiar and unfamiliar participants was provided by Jenkins et al. (2011) using a face identity sorting task. Two groups of participants - one from the UK, the other from the Netherlands - sorted 40 images of two Dutch celebrities (20 images per identity) into piles of the same identity. While participants from the Netherlands, who were familiar with these individuals, sorted the images most frequently into two piles (median = 2), participants from the UK, who were unfamiliar with the individuals, sorted the images most frequently into 9 piles (median = 7.5). Despite perceiving more identities than were actually present, unfamiliar participants only rarely sorted pictures of two different identities into the same pile. Unfamiliar participants were therefore able to successfully "tell people apart", while they struggled to "tell people together" and perceived the highly variable images from a single identity to belong to a number of different identities.

This finding has since been replicated and extended: for example, the marked differences between familiar and unfamiliar viewer's behaviour have been shown to disappear when participants know how many identities to expect (Andrews, Jenkins, Cursiter & Burton, 2015). In these cases, both viewer groups sorted the pictures into two piles with high accuracy and with few identity confusions. Redfern and Benton (2017) manipulated the expressiveness of unfamiliar faces, contrasting highly expressive versus less expressive (closer to neutral) faces in a sorting task. When faces were highly expressive, participants were more likely to sort two pictures from different identities into the same pile, making more errors when "telling people apart". Furthermore, Zhou and Mondloch (2016) showed an other-race effect in a face sorting task, where viewers sorted unfamiliar other-race faces into more perceived identities than unfamiliar own-race faces. This effect, however, was not present for familiar faces, where participants were highly accurate in both conditions. These face sorting studies show compelling interactions between familiarity and within-person variability where familiar listeners appear to be able to generalise across the variability, while unfamiliar listeners fail to do so in many cases.

The current study

Within-person variability a key feature of human voices that to date been largely neglected in the study of voice perception – despite there being evidence that it affects voice identity perception. The face perception literature has shown that sorting tasks have been shown a powerful tool for investigating different aspects of identity processing in the context of within-person variability: both can participants' ability to

"tell people apart" and "tell them together" be assessed within a single task, while also being able to contrast performance for familiar versus unfamiliar voices.

In the current study, we therefore investigate how within-person variability affects voice identity perception for familiar and unfamiliar voices using a voice sorting task. We selected voices from a popular TV show (*Orange is the New Black*) and asked participants who had watched the show and participants who had not watched the show to sort 30 voice samples (2 voices, 15 exemplars per voice) into perceived identities. Crucially, our voice samples included natural within-person variability in the voice, having been extracted from across different speaking situations, environments and interlocutors. We ran this voice sorting task in three independent participant samples, each using different stimulus sets to assess the replicability of effects. We predicted that unfamiliar listeners will perceive more identities than familiar listeners: in the absence of stable mental representation of a voice identity, natural within-person variability can be mistaken for between-person variability and can thus have a detrimental effect on accuracy. We also predicted that unfamiliar listeners would be biased to mistaking within-person variability as between-person variability, thus selectively failing to "tell people together" while being mostly able to "tell people apart" (see Jenkins et al., 2011; Andrews et al., 2015; Redfern & Benton, 2017; Zhou & Mondloch, 2016 for faces).

Methods

Participants

152 participants were recruited via social media (e.g. Twitter and Facebook) and the participant pool of the Division of Psychology at XXX University. Participants were

either entered into a prize draw or received course credit for their participation. The study was approved by the local ethics committee. The 152 participants were randomly allocated to the three versions of the task (Set 1-3; see below). Matching the sample size used by Jenkins et al. (2011), we aimed to recruit at least 20 participants for both our familiar and unfamiliar listener groups per set. Familiarity was assessed via self-report: if participants reported to have watched more than one season of Orange Is the New Black, they were assigned to the familiar group. Participants who reported to have not seen any episodes of the TV show were assigned to the unfamiliar group. Participants who reported to have seen some episodes but not a full season were excluded from all analyses ($N = 7$). Participants, who reported that they had recognised or remembered more than three of the specific exemplars included in their set were also excluded ($N = 3$) as their responses may have been driven by the specific memory of the scene as opposed to direct voice identity recognition. Additionally, we excluded participants who moved less than 80% of the exemplars (i.e. 24 exemplars out of 30; see below for information on the task) from their original position on the slide ($N = 1$) or whose performance (indexed by number of perceived identities; see below) differed by more than 3 standard deviations from the mean of their listener group and set ($N = 3$).

This resulted in a final data set of 68 familiar and 70 unfamiliar participants in total: 25 familiar (21 female, mean age: 18.68 years, SD: 1.15 years) and 22 unfamiliar participants (19 female, mean age: 18.70 years, SD: 1.72 years) for Set 1, 22 familiar (15 female, 1 other, mean age: 24.36 years, SD: 7.78 years) and 22 unfamiliar participants (16 female, 1 other, mean age: 28.91 years, SD: 10.90 years) for Set 2 and 21 familiar (18 female, mean age: 26.48 years, SD: 4.12 years) and 26 unfamiliar participants (22 female, 1 other, mean age: 26.28 years, SD: 10.32 years) for Set 3.

Materials

In the current study, we used exemplars of voices of three female characters with significant speaking roles from the TV show "Orange Is the New Black" (VoiceID 1: Nicky Nichols, VoiceID 2: Alex Vause and VoiceID 3: Piper Chapman). The show was selected as it features a large number of characters with significant speaking roles, providing a large pool of possible voices that could be presented in the experiment.

15 exemplars per identity were extracted (mean duration: 3.12 seconds; *SD*: .32 seconds): These exemplars included full utterances with as little background noise as possible, avoiding catch phrases and other diagnostic verbal information (example stimulus: "*and that she is on her way out of town*"). Each exemplar was extracted from a different scene to sample substantial natural within-person variability (see Supplementary Materials 2 and Supplementary Figure 1 for plots of affective and acoustic properties of the stimuli). Only recordings from the first three seasons of the TV show were included (released between two and four years before testing started) to decrease the likelihood that participants had recently heard the stimuli and would therefore remember the scenes in which they occurred. Exemplars were normed for intensity using PRAAT (Boersma and Weenink, 2017).

Procedure

There were three versions of the task (referred to as sets throughout the paper), including all possible pairs of the three different voices (Set 1: Nicky Nichols and Alex Vause, Set 2: Piper Chapman and Alex Vause, Set 3: Piper Chapman and Nicky Nichols) to assess the replicability of effects. Participants completed the experiment using the online testing platform Qualtrics (reference for Qualtrics?), where they downloaded a Microsoft Powerpoint slide that included 30 embedded sound files (2 identities x 15

exemplars). Each of these exemplars was represented by a number (see the bottom panel of Figure 1 for examples of listeners' completed solutions). The numbers were distributed evenly across the slide, with no clusters being obvious from the outset. In line with the methods used in Jenkins et al. (2011), participants were asked to sort the 30 exemplars into clusters, which each cluster including the exemplars produced by a single speaker, thus representing a perceived speaker identity. This was done via dragging and dropping the exemplars on the slide. Participants could replay the exemplars as many times as they wanted, and there was no time limit on completing the task.

Results

Data were analysed both in terms of the number of perceived identities as well as how the exemplars were grouped, contrasting "telling people apart" versus "telling people together". For all analyses, we reported the effects for each set separately, given that the stimuli were different in each set, which also allowed us to assess consistency of effects across different stimuli. Shapiro-Wilk tests showed that the data was not normally distributed in most conditions (i.e. for each set and familiarity condition), and therefore we used non-parametric tests throughout.

How many identities did familiar and unfamiliar listeners perceive?

For this analysis, we counted the number of clusters (i.e. how many identities did listeners perceive) per participant. In the two cases (1 familiar listener, 1 unfamiliar listener, both from Set2) where it was not apparent whether exemplars were intended to be clustered together or not, the maximum number of separate clusters was assumed. Familiar listeners perceived fewer clusters than unfamiliar listeners on

average for all sets (see Figure 1, top panel. Familiar: Set 1 Median = 3, Mode = 2, Range = 2-8, Set 2 Median = 3, Mode = 2, Range = 2-8, Set 3 Median = 3 , Mode = 2, Range = 2-12; Unfamiliar: Set Median = 4, Mode = 4, Range = 2-10, Set 2 Median = 7, Mode = 5, Range = 4-17, Set 3 Median = 9, Mode = 11, Range = 3-15). Wilcoxon rank-sum tests showed that familiar listeners perceived significantly fewer identities than unfamiliar listeners for two out of the three individual sets (Set1: $Z = 1.888$, $p = .030$; Set2: $Z = 4.452$, $p < .001$; Set3: $Z = 4.211$, $p < .001$; α was Bonferroni-Holm-corrected for three comparisons).

In addition to differences in the number of perceived identities, we observed patterns of responses that were qualitatively different for familiar and unfamiliar listeners: familiar listeners tended to create at least one – often two – large clusters (14+ items per cluster) plus a number of single-exemplar clusters, resulting in a bimodal distribution. Unfamiliar listeners, however, tended to create a number of smaller clusters (2-6 items, see Figure 1, middle panel). After collapsing all raw cluster counts across the three sets, a Chi Square test of independence confirmed that the distributions of frequencies of cluster sizes for familiar and unfamiliar listeners are independent ($\chi^2[14] = 188.43$, $p < .001$).

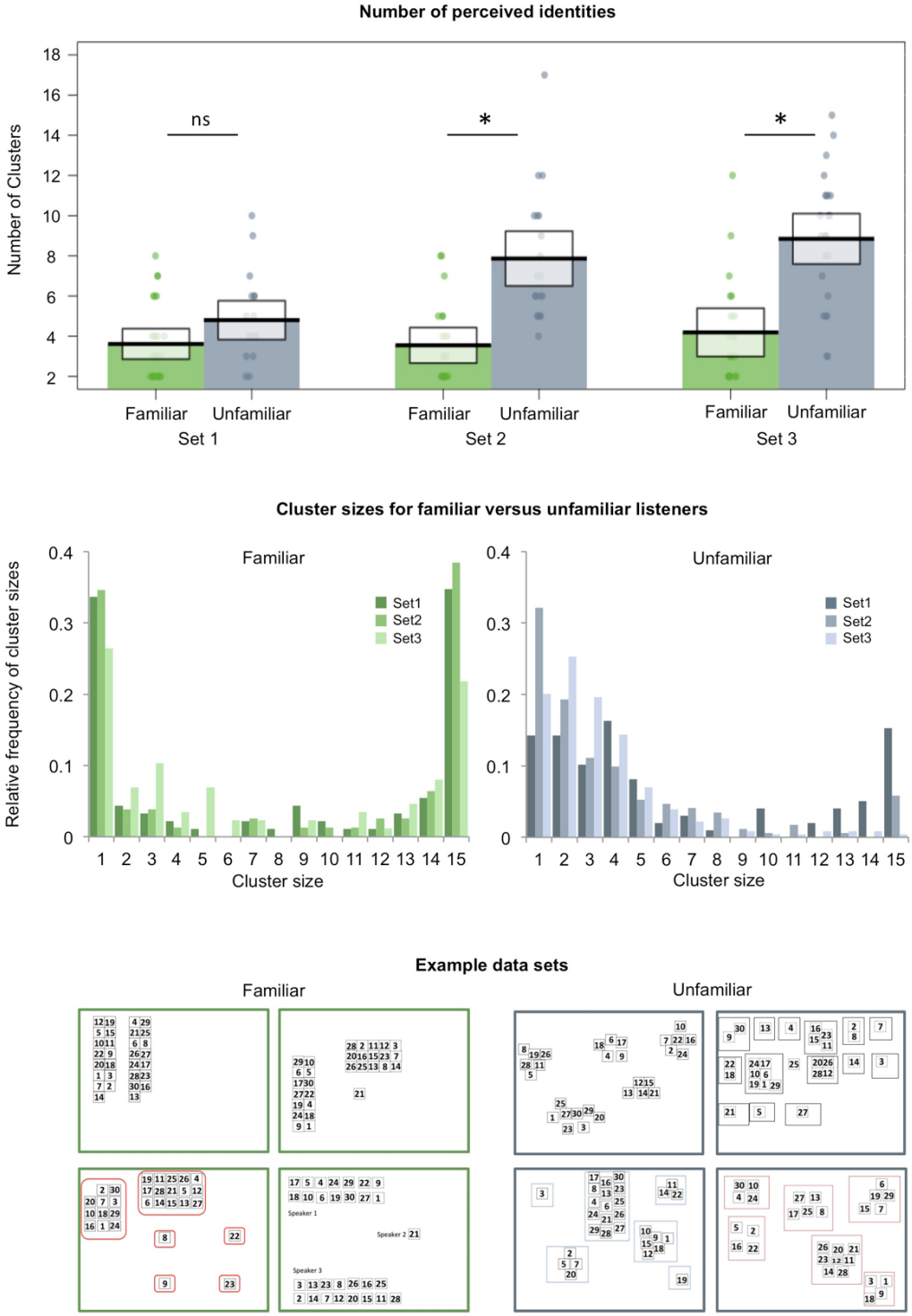
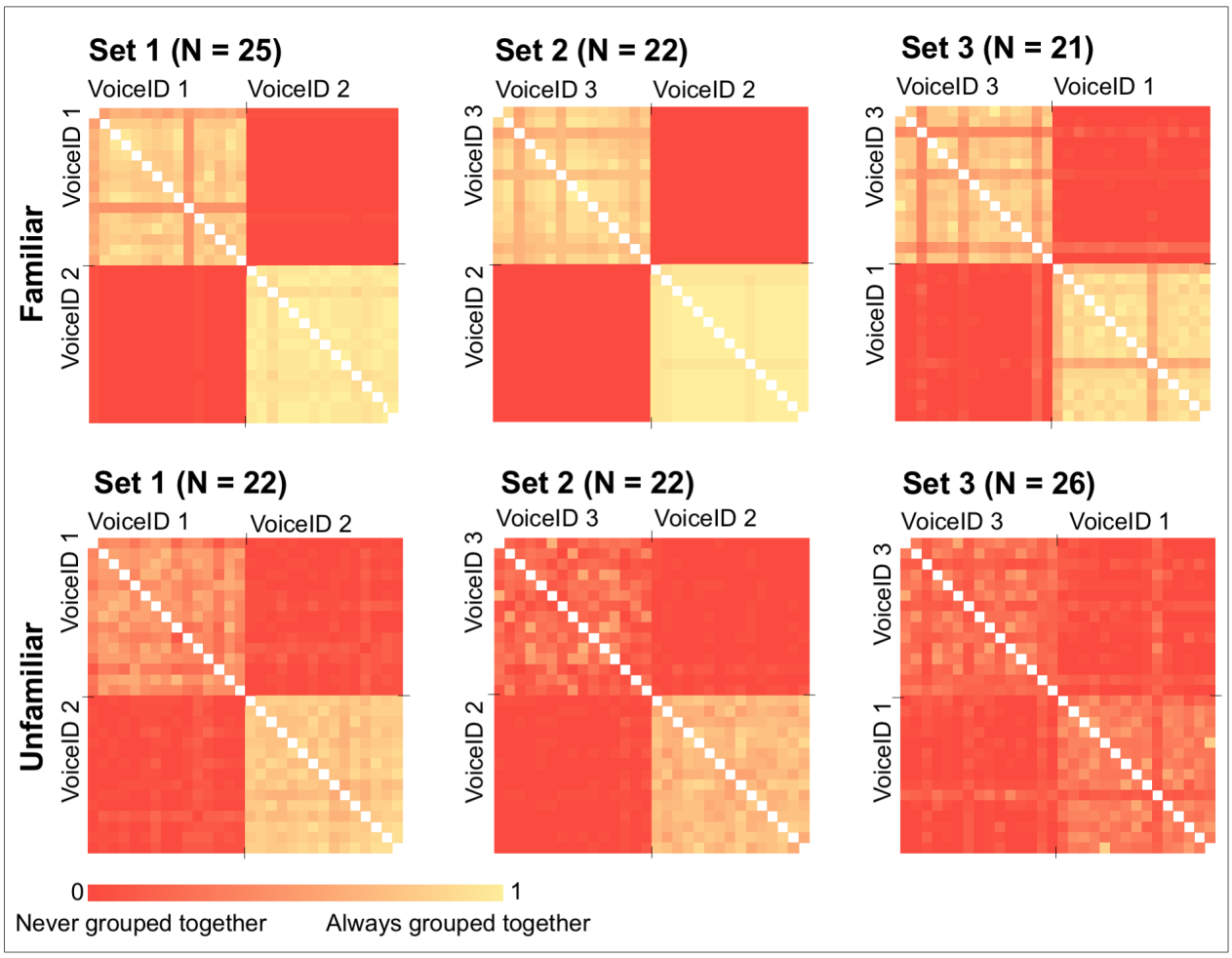


Figure 1 Top panel: Number of perceived identities by set and averaged across all sets for familiar and unfamiliar listeners. Bars show the means across participants, and each dot shows one participant. Boxes show the 95% confidence intervals for the means of each. Stars show significant differences between familiar and unfamiliar listeners. Middle panel: Plots of the relative frequency of cluster sizes (count per cluster size divided by the total number of clusters within each set) for familiar and unfamiliar listeners. Bottom panel: Representative example data sets as an illustration of familiar and unfamiliar participants' response patterns.

"Telling people apart" versus "telling people together"



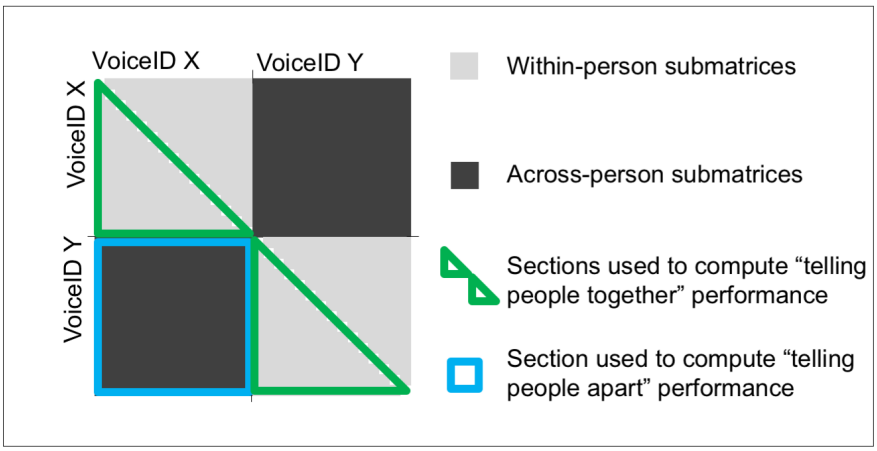


Figure 2 Top panel: Matrices of averaged listeners' responses for the three version of the task for familiar and unfamiliar listeners. Within these 30 x 30 matrices (15 sounds files x 2 identities), each cell shows the probability that two exemplars were grouped within the same perceived identity: cells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

To assess the differences of familiar and unfamiliar listeners' ability to "tell people apart" and conversely "tell people together", we created 30x30 response matrices for each participant (15 sounds files x 2 identities; each cell shows the probability that two exemplars were grouped in the same cluster: cells coded as 1 indicate that the two respective exemplars were always grouped together; cells coded as 0 indicate that the two exemplars were never grouped together — Figure 2 bottom panel). These per-participant response matrices thus provide a detailed representation of *how* listeners grouped the different sounds into perceived identities. We used these matrices to characterise errors in "telling people apart" and "telling people together". Figure 2 (top panel) shows the group-averaged response matrices – these matrices are symmetric across the diagonal. Conceptually, these matrices are divided into within-identity and across-identity submatrices (see Figure 2, bottom panel). Within-identity submatrices index listeners' ability to "tell people together": for the ideal solution (creating the 2 correct clusters), each cell within these submatrices would be 1 as all pairs of exemplars from the same identity were put into the same cluster. The across-identity submatrix indexes the ability to "tell people apart": an ideal solution here would result in all cells within this submatrix to be 0 as no pairs of exemplars from different identities were ever put into the same cluster (see Figure 2, bottom panel).

To quantify whether familiar and unfamiliar listeners' performance for "telling people together" and "telling people apart" differed, we computed the mean probability of how often two exemplars from the same identity were grouped together by taking the mean of the values in the upper triangle of each symmetrical within-identity submatrix (excluding the diagonal which is by definition always 1 and therefore not meaningful; 2 x 105 cells, see Figure 2 bottom panel). For familiar listeners, median "telling people together" probabilities were relatively high for familiar listeners (Set 1 =

.88, Set 2 = .93, Set 3 = .81) and lower for unfamiliar listeners (Set 1 = .64, Set 2 = .46, Set 3 = .18). Wilcoxon rank-sum tests confirmed that familiar listeners are significantly more likely to group exemplars from the same identity together than unfamiliar listeners for all three sets (all Zs > 3.412, all ps < .001; α was Bonferroni-Holm-corrected for three comparisons).

A comparable analysis was run for "telling exemplars apart" submatrices. We computed the mean of the values in the across-identity matrix (225 cells, see Figure 2 bottom panel). The median value for "telling people apart" was very low (or 0) for familiar listeners (Set 1 = 0, Set 2 = 0, Set 3 = .02) as well as unfamiliar listeners (Set 1 = 0, Set 2 = 0, Set 3 = .04). Wilcoxon rank-sum test showed that unfamiliar listeners indeed made significantly more errors than familiar listeners in one of the three sets (Set1: Z = 2.670, p = .004; Set 2: Z = 1.770, p = .038; Set 3: Z = 1.483, p = .069; α was Bonferroni-Holm-corrected for three comparisons).

Misclassifications or errors in "telling people apart" were relatively rare for both familiar and unfamiliar listeners, while errors in "telling people together" were relatively more frequent. To explicitly compare these error rates, we computed 1 minus the mean probability of the lower triangle of each within-identity matrix for each participant. The error rates for "telling people apart" were the mean probability of the across-identity matrices. Wilcoxon signed-rank tests showed that there was indeed a significant difference in error rates in "telling people apart" versus "telling people together" for familiar and unfamiliar listeners for all sets (familiar: all Zs > 2.947, all ps < .002; unfamiliar: all Zs > 3.944, all ps < .001; α was Bonferroni-Holm-corrected for three comparisons). "Telling people together" can thus be considered to be a more challenging or error-prone process.

Thus striking differences in the behaviour of familiar and unfamiliar listeners are apparent: unfamiliar listeners were less likely to group exemplars from the same identity together compared to familiar listeners. In contrast, both listener groups can be considered to have largely succeeded at telling the two different identities apart, given the very low error rates. For additional analyses of the consistency of matrices across participants, (the lack of) effects of acoustic measures, perceived likeness and affective ratings on listeners' responses and voice-specific effects see the Supplementary Materials.

Discussion

The current study for the first time explored how *natural* within-person variability affects voice identity processing in familiar and unfamiliar listeners within the same paradigm. When asked to group 30 sound clips from a popular TV show (2 voices, 15 exemplars each) into perceived identities, familiar listeners thought that on average between 3 and 4 speakers were included in the set. In contrast, unfamiliar listeners perceived more speakers (on average between 4 to 9 speaker). Unfamiliar listeners frequently perceived exemplars from the same speaker as different identities pointing to selective difficulties to "tell people together", failing to generalise identity information across variable signals. Both listener groups only made a relatively small number of errors in "telling people apart" by grouping exemplars from two identities into the same cluster. These findings are thus a first direct demonstration of unfamiliar listeners' failure to "tell people together" in the context of naturally varying voice recordings (for comparable findings for faces, see Jenkins et al., 2011) and highlight the need to further study how within-person variability, a feature central to human voices, impacts of voice identity perception.

While current models of voice processing do not explicitly account for within-person variability and only little empirical evidence probing this issue is available, the findings of our study can be integrated into and advance current models of voice processing. The model of voice identity processing proposed by Sidtis and Kreiman (2011, 2012) focuses on the distinction of familiar and unfamiliar voice processing during identity perception. Here, familiar voice recognition and unfamiliar voice discrimination are considered to be mechanistically distinct (featural comparison versus pattern recognition), are dissociable from one another and thus predict differences in the behaviour of familiar and unfamiliar listeners. In our study, we indeed found striking differences between familiar and unfamiliar listeners' performance, using the same task for both listener groups.

Prototype models of voice processing may offer some insights into the nature of the different representations of familiar and unfamiliar voices. Such prototype models propose that listeners encode and process voice identity information in relation to a prototype, which is an (context-dependent) average voice (Latinus & Belin, 2010; Latinus et al., 2013; Lavner, Rosenhouse & Gath, 2001; Papcun et al., 1989; see also Maguiness et al., 2018). While empirical studies show some support for these models, studies have to our knowledge only explored prototype models with a focus on *between*-speaker variability by using different voice identities. The mechanisms assumed for prototype models can, however, be readily extended and applied to the processing of *within*-person variability: for a familiar voice, listeners can access a specific prototype or representation of a particular voice. These representations of familiar voices are likely to include the characteristics of how a specific voice varies. Due to this, they can thus still relatively reliably process voice identity in the face of within-person variability. For unfamiliar voices, neither specific representation is available nor

have the characteristics of how a specific voice varies been encoded. The lack of specific information may thus result in the processing of identity being less reliable.

Our studies' findings closely resemble the results reported for faces (Jenkins et al., 2011). Many parallels have in the past been described between face and voice processing (e.g. Campanella & Belin, 2007; Kuhn et al., 2017; Yovel & Belin, 2013) – the similarities of results apparent in auditory and visual identity sorting studies are nonetheless remarkable. The materials used in face and voice sorting studies are, however, very different from each other: not only do the materials derive from two different modalities, they also provide participants with in the case of faces with static information and dynamic information in the case of voices. Given these profound differences in the signals, the nature of the within-person variability present in both sets will also differ accordingly, with no clear one to one correspondence between the sources of variability: how does, for example, variability in the lighting of images relate to variability introduced by background noise? Neither is there a direct equivalent for differences in viewpoint in the auditory domain, nor can we adequately describe a regional accent in the (static) visual domain. In short: salient features for identity processing and sources of variability seem to be modality-specific.

Aside from differences in materials, the task of identity sorting allows participants to choose their own strategy to complete the tasks with no explicit instructions guiding them: strategies may differ between faces and voice versions of the task – an aspect that cannot be further analysed with the available data. Despite these factors, patterns of results for face and voice sorting tasks are comparable: such parallels may suggest that sorting tasks tap into stages of identity processing in familiar as well as unfamiliar participants, that may either rely on abstracted amodal processes or possibly modality-bound processes that are mirrored in the auditory and visual

domain (see e.g. Yovel & Belin, 2013). Mapping out in which contexts the processing of face and voice identities is comparable and in which circumstances the two modalities differ remains a largely open question and warrants further work.

One of the novel aspects of this study is its use of the relatively uncontrolled exemplars that include substantial, natural within-person variability (see 'ambient images' for faces, Jenkins et al. 2011): exemplars varied in extrinsic features, such as the overall quality of the recording, type and amount of background noise among any number of other factors. Furthermore, the exemplars differed in their linguistic/verbal content (different utterances), verbal register, type of utterance, vocal effort (quiet conversation versus shouting) as well as their perceived affective properties, such as valence and arousal, perceived 'likeness', among any number of features (see supplementary materials). While the current study shows how uncontrolled natural within-person variability from a range of sources can affect speaker identity perception, other studies have shown how specific sources of variability can affect perception (e.g. language spoken [Zarate et al., 2015], linguistic content [Naranyan et al., 2016], vocalisations type [Lavan et al., 2016], distinctiveness [Papcun, Kreiman & Davis, 1989] and duration of the exemplars [Schweinberger, Herholz & Sommer, 1997]). How these different types of variability relate to each other and interact in the context of identity perception is largely unexplored. Similarly, we do not know whether different types of variability might be more disruptive to perception than others or whether their effects are comparable to each other. Further studies are therefore needed to better characterise the nature of within-person variability and its effects on identity perception.

Within-person variability has until recently been neglected by studies of identity perception in the visual modality (Burton, 2013) and has yet to be included in studies of

identity perception in the auditory modality (Lavan et al., 2017). The present study demonstrates that within-person variability poses challenges for the reliable processing of identity from voices – especially for unfamiliar listeners. Within-person variability may, however, not always be a challenge that listeners need to overcome as recent intriguing findings from the face identity perception literature suggest. Burton et al. (2016) showed that within-person variability is specific to an individual's face, i.e. *how* the face of one person varies is different from that of another. Variability may therefore encode diagnostic information about a person's identity, as opposed to merely being noise. There is also some evidence that within-person variability may indeed be instrumental to building up robust representations of a person, given that participants are more successful at learning a novel identity from training with variable sets of face stimuli compared to when trained on less variable sets (Murphy, Ipser, Gaigg & Cook, 2015; Ritchie & Burton, 2017). Given the striking parallels between the findings of the current study and reports from face sorting tasks, it is possible that the processing proposed for identity learning from variable faces may also extend to how voices are learnt. Future work will therefore not only need to map out how listeners' judgements are affected by within-person variability, but also to explore whether and how within-person variability may be an essential part of voice identity learning.

References

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.

Balas, B., & Pearson, H. (2017). Intra-and extra-personal variability in person recognition. *Visual Cognition*, 1-14.

Boersma, P. & Weenink, D. (2017). Praat: doing phonetics by computer [Computer program].

Bruce, V., Henderson, Z., Newman, C., & Burton, A. M. (2001). Matching identities of familiar and unfamiliar faces caught on CCTV images. *Journal of Experimental Psychology: Applied*, *7*(3), 207.

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *The Quarterly Journal of Experimental Psychology*, *66*(8), 1467-1485.

Burton, A. M., Kramer, R. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*(1), 202-223.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999). Face recognition in poor-quality video: Evidence from security surveillance. *Psychological Science*, *10*(3), 243-248.

Campanella, S., & Belin, P. (2007). Integrating face and voice in person perception. *Trends in Cognitive Sciences*, *11*(12), 535-543.

Cumming, G. (2012). *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. New York: Routledge.

Hancock, P. J., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in cognitive sciences*, *4*(9), 330-337.

Hill, H., & Bruce, V. (1996). Effects of lighting on the perception of facial surfaces. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(4), 986-1004.

Hole, G. J., George, P. A., Eaves, K., & Rasek, A. (2002). Effects of geometric distortions on face-recognition performance. *Perception*, *31*(10), 1221-1240.

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

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.

Johnsrude, I., Casey, E., & Carlyon, R. P. (2014). Listen to your mother: Highly familiar voices facilitate perceptual segregation. *The Journal of the Acoustical Society of America*, *135*(4), 2423-2423.

Kemp, R., Towell, N., & Pike, G. (1997). When seeing should not be believing: Photographs, credit cards and fraud. *Applied Cognitive Psychology*, *11*(3), 211-222.

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

Kreiman, J., Keating, P. A., Park, S. J., Rastifar, S., & Alwan, A. (2015). Within-and between-talker variability in voice quality in normal speaking situations. *The Journal of the Acoustical Society of America*, *137*(4), 2418-2418.

Kuhn, L. K., Wydell, T., Lavan, N., McGettigan, C., & Garrido, L. (2017). Similar representations of emotions across faces and voices. *Emotion,* 17(6), 912-937.

Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, *23*(12), 1075-1080.

Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, *2*, 175.

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2017). Flexible voices: identity perception from variable vocal signals. [https://psyarxiv.com/pczvm]

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., Rosenhouse, J., & Gath, I. (2001). The prototype model in speaker identification by human listeners. *International Journal of Speech Technology*, *4*(1), 63-74.

Maguinness, C., Roswandowitz, C., & Von Kriegstein, K. (in press). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia.*

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 577 - 581.

Narayan, C. R., Mak, L., & Bialystok, E. (2017). Words get in the way: Linguistic effects on talker discrimination. *Cognitive Science*, *41*(5), 1361-1376.

Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *The Journal of the Acoustical Society of America*, *85*(2), 913-925.

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.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Read, D., & Craik, F. I. (1995). Earwitness identification: Some influences on voice recognition. *Journal of Experimental Psychology: Applied*, *1*(1), 6 - 18.

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

Ritchie, K. L., & Burton, A. M. (2017). Learning faces from variability. *The Quarterly Journal of Experimental Psychology*, *70*(5), 897-905.

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology*, *65*(1), 111 - 116.

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.

Sidtis, D., & Kreiman, J. (2012). In the beginning was the familiar voice: Personally familiar voices in the evolutionary and contemporary biology of communication. *Integrative Psychological and Behavioral Science*, *46*(2), 146-159.

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, *25*(5), 829-834.

Wagner, I., & Köster, O. (1999). Perceptual recognition of familiar voices using falsetto as a type of voice disguise. Proceedings of the 14th International Congress of Phonetic Sciences, 2, 1381-1384.

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

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499. [http://arxiv.org/pdf/1308.5499.pdf]

Yip, A. W., & Sinha, P. (2002). Contribution of color to face recognition. *Perception*, *31*(8), 995-1003.

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, *17*(6), 263-271.

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.

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.

SUPPLEMENTARY ANALYSES

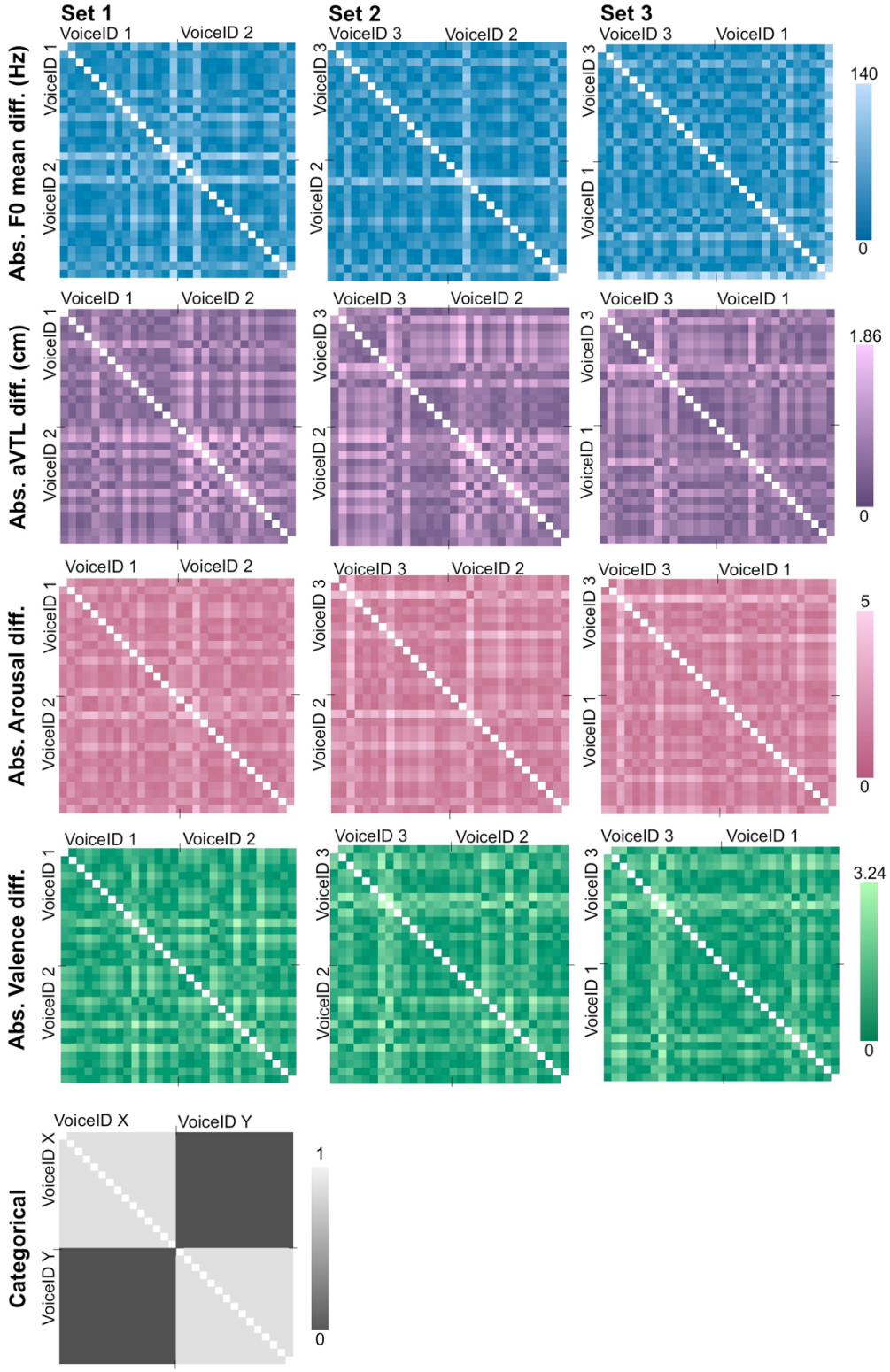
1. How similar are individual response matrices to each other?

In the current study, familiar and unfamiliar listeners responses differed from one another on a number of levels. Whether the response patterns of individual listeners within a group differ from each other or are highly similar was explored in the following analyses. Each participant's 30x30 response matrix was correlated with every other participant's matrices within their set and listener group using Kendall's τ_a. We obtained a mean correlation per participant and then computed the mean across participants.

These analyses showed that the matrices for all 3 sets and both listeners groups were correlated (Familiar Mean Kendall's τ_a; Set 1 = .359, Set 2 =.386, Set 3 = .269; Unfamiliar Mean Kendall's τ_a; Set 1 = .178, Set 2 =.148, Set 3 = .032). Wilcoxon's signed rank tests comparing mean correlations for all participants within each set against 0, familiar: all *Z*s > 3.832, all *p*s < .001; unfamiliar: all *Z*s > 3.893, all *p*s < .001; α was Bonferroni-corrected for three comparisons). However, mean correlations were significantly stronger among familiar listeners compared to unfamiliar listeners for all sets (Wilcoxon's rank sum tests, all *Z*s > 5.664, all *p*s < .001; α was Bonferroni-corrected for three comparisons). This may indicate that, due to better task performance (i.e. the number of perceived identities was closer to the veridical number of identities present), familiar listeners arrived at more similar solutions compared to unfamiliar listeners. While some consistency is present in the ratings of the unfamiliar listeners, participants

seem to have arrived at quite dissimilar solutions, which may indicate that unfamiliar listeners may have used a number of different strategies to complete the task.

2. What affects listeners' response patterns: Acoustic, affective and categorical candidate models

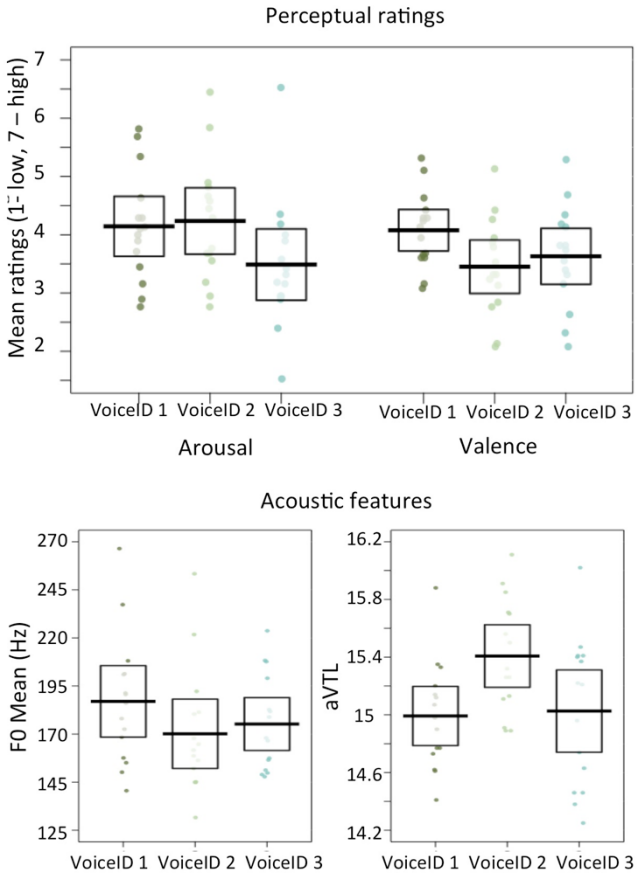


Supplementary Figure 1 Matrices for candidate models for the three version of the task. In these matrices, cells code for the absolute difference of acoustic measures (Fo mean, apparent vocal tract

length [aVTL]) and affective properties (arousal, valence). In the categorical model cells that include exemplars from the same identity are coded as 1 and cells that include two different identities are coded as 0.

To explore whether acoustic properties can be linked to average response matrices of the voice sorting task (see Figure 2 in the main text), we created candidate models for a number of acoustic and perceptual properties of the stimuli. Previous studies have shown that the acoustic features of a vocal signal, especially source characteristics (such as F0 mean) and filter characteristics (such as formant measures or vocal tract length) can be linked to listeners' judgements of speaker identity (Baumann & Belin, 2010; Lavner, Gath & Rosenhaus, 2000; see also Kreiman & Sidtis, 2011 for an overview).

We therefore extracted measures of F0 mean and apparent vocal tract length for each stimulus (aVTL; see Cartei, Cowles & Reby, 2012 for details on how aVTL was computed). We then computed the absolute pairwise difference on each acoustic measure between each pair of stimulus for each set. This resulted in mean f0 and aVTL models for each set, consisting of 30 x 30 matrices (2 identities x 15 exemplars) (Supplementary Figure 1). These matrices show substantial within-speaker variability both in F0 mean as well as aVTL (which is affected in the case of our exemplars by the different vowels included in each set): The absolute differences in these measures appear as similar for exemplars from the same speaker as they are for exemplars from two different speakers, with no clear separation into speaker identities. Similarly, the within-person variability is also illustrated in Supplementary Figure 2, which shows the mean and distribution of acoustic properties for each voice identity.



Supplementary Figure 2 Plots of arousal and valence ratings and acoustic features. Boxes show the 95% confidence intervals for the means of each group (plots were created using the *yarrr* package in the R environment).

To examine whether the acoustic properties could explain participants' behaviour, the lower triangle of each of these candidate models (see Supplementary Figure 1) was correlated with the lower-triangle of the participants' mean voice identity matrices (Figure 2, main paper) using Kendall's τ_a (for similar analyses see, for example, Kuhn, Wydell, Lavan, McGettigan & Garrido, 2017). We computed these correlations per set and per listener group. To assess whether these correlations were significant, we compared these observed correlations to a distribution of values predicted by chance using random permutation tests for each comparison. For each of 5,000 permutations, we correlated the lower triangle of each mean voice identity matrix with

a random permutation of the candidate model. An observed correlation that is higher than 95% of the chance predictions ($p < 0.05$) allows us to reject the null hypothesis. None of the Fo mean or VTL candidate models were significantly correlated with the average matrices per set and per listener group (all Kendall's $\tau_a < .066$, all $p > .019$, α was Bonferroni-corrected for three comparisons), with one exception: the VTL candidate model negatively correlated with the average matrix of familiar listeners' solutions for Set2 (Kendall's $\tau_a = .072$, $p = .01$).

In addition to these two acoustic candidate models, we created models of the affective properties of exemplars, using post-hoc ratings of arousal and valence. A group of 38 listeners, comprised of 15 listeners who were familiar with the show (2 male; mean age = 19.06 years, SD = .77 years) and 22 listeners who were not familiar with the show (2 male; mean age = 20.68 years, SD = 5.37 years) completed this task for course credit at XXX, none of whom had previously taken part in the main experiment. Participants were presented with the 45 exemplars (15 exemplars x 3 identities) in randomized order and provided ratings of arousal (1 = the person sounded very drowsy or sleepy, 7 = the person sounded highly energetic and alert) and valence (1 = very negative, 7 = very positive) for all stimuli. The order of stimuli and of scales was randomized across participants. For all scales, participants were asked to rate the quality of the voice and disregard the verbal content of the stimuli. From these ratings, average ratings of arousal and valence were computed per item. The responses from familiar and unfamiliar listeners were combined. No participants were excluded.

From these ratings, matrices coding for the absolute differences in the respective ratings between pairs of sounds were computed. These model matrices for arousal and valence (see Supplementary Figure 1) were correlated with the group-averaged matrices for the voice sorting task (see Figure 2 in main text). These ratings

also show significant within-person variability (as illustrated in Supplementary Figures 1 and 2), which is similar to the between-speaker variability observed. None of the arousal or valence candidate models was significantly correlated with the average matrices per set and per listener group (all Kendall's τ_a < .056, all p > .033, α was Bonferroni-corrected for three comparisons).

Finally, a categorical candidate model where cells code pairs of exemplars from the same identity as 1 and cells that include exemplars from different identities as 0. Here, the candidate model was highly correlated with the mean response matrices for both familiar and unfamiliar listeners for all sets (all Kendall's τ_as > .450, all p < .001). Thus, a categorical candidate model seems to be the best representation the listeners' responses for all sets. These results suggest that listeners did not consistently use the selected acoustic measures or affective features to make their judgements in the voice sorting task.

Not finding a clear relationship between acoustic measures for familiar listener fits with theoretical models of familiar voice processing: according to the models, familiar identity perception is not closely tied to acoustic features (Kreiman & Sidtis, 2011). It is more surprising, that acoustic candidate models could not be linked to unfamiliar listeners' data (while categorical candidate modes correlate with their data), as models predict identity processing based on the physical features of the voices. Here, it may be the case due to large within-person variability, acoustic properties (see the lack of a clear separation of identities in the acoustic candidate models in Supplementary Figure 1), such as the ones probed here, were rendered non-diagnostic (while they were highly diagnostic in studies that minimized within-person variability). Unfamiliar listeners' may therefore have relied on other unknown acoustic features to

arrive at non-random solutions (see the correlation of averaged listeners' responses with the categorical models).

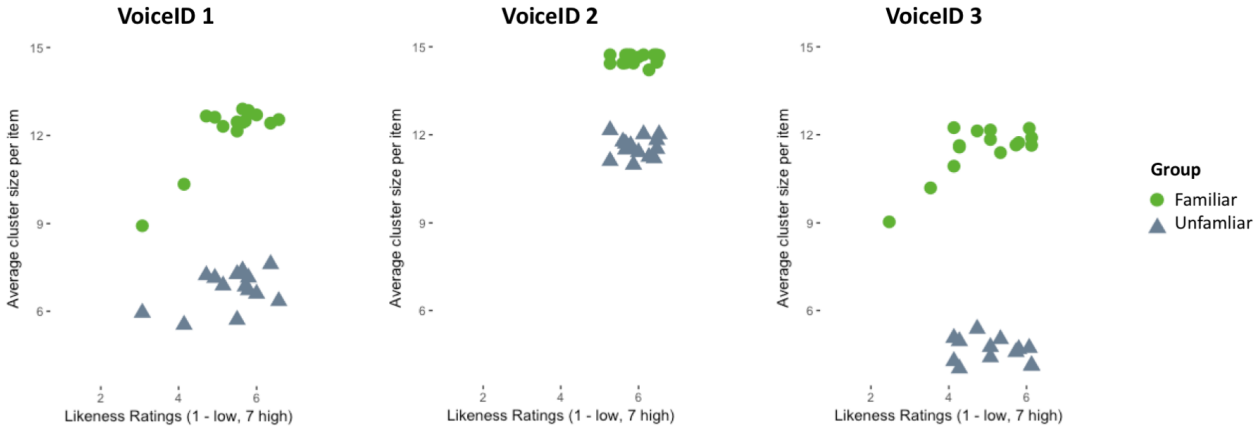
3. Effects of different speaker identities: not all voices are alike

The group-averaged voice identity matrices show that there were identity-specific effects, with exemplars for some identities being easier to "tell together" than for other identities. We tested whether the probabilities with which listeners grouped the exemplars of an identity together differed for the three voices, using within-set comparisons (e.g. comparing VoiceID1 from Set 1 versus VoiceID2 from Set 1; see main text for methods). Wilcoxon signed rank tests showed that the probability of "telling together" for the same identity differed across most voices, in both familiar and unfamiliar listeners (familiar: [VoiceID1 versus VoiceID2; VoiceID2 versus VoiceID3]: both Zs > 2.992; ps < .002; unfamiliar: [VoiceID1 versus VoiceID2; VoiceID2 versus VoiceID3]: both Zs > 2.576; both ps < .006); α was Bonferroni-corrected for three comparisons). Only the comparison of VoiceID1 versus VoiceID3 for familiar listeners did not reach significance (Z = .397; p = .665). Therefore, there were differences in how readily listeners were able to "tell together" the exemplars of the three identities for unfamiliar listeners and familiar listeners: for example, familiar and unfamiliar listeners alike were more successful at telling the exemplars of VoiceID2 together compared to the other two identities, marking this particular voice as being distinctive or less inherently variable than the other voices (see also Supplementary Figure 3 in the main text).

We also examined effects of context by testing whether the probability with which listeners' grouped different exemplars of an identity together differed depending on the other identity that was presented (e.g. VoiceID1 from Set 1 versus VoiceID1 from

Set3; see above for Methods). Interestingly, the probabilities did not differ from each other for any of the three identities used in neither for familiar nor for unfamiliar listeners (Wilcoxon rank-sum tests; familiar: all *Z*s < 1.101, all *p*s > .135; unfamiliar: all *Z*s < 1.307, all *p*s > .096; α was Bonferroni-corrected for three comparisons). This result shows that the nature of the other identity included in the sets did not significantly change how difficult it was to tell exemplars of the same identity together, speaking against effects of context.

4. *Not all exemplars are alike: effects of perceived likeness*



Supplementary Figure 3 Scatterplots of the exemplar-wise mean cluster size and likeness ratings per identity for familiar and unfamiliar listeners. Cluster size was averaged across the two samples in different sets.

Not all exemplars may be alike in the context of within-person variability: some exemplars can sound more like a familiar person than others (e.g. Ritchie, Kramer & Burton, 2017 for faces). To investigate whether perceived likeness has an effect on how identity information is processed, we computed a measure of cluster size for each exemplar, averaged across the two instances each exemplar occurred across the 3 sets. Our previous analyses have shown that familiar and unfamiliar listeners generally

succeed at "telling identities apart" but unfamiliar listeners struggle to "tell identities together", resulting in a larger number of perceived identities. In this context, cluster size per exemplar can thus serve as an index of how difficult a listener found it to associate a particular exemplar with the other exemplars of this identity.

We collected perceptual ratings of perceived likeness from an independent group of 15 listeners who were familiar with the TV show (13 female; mean age = 19.06 years, *SD* = .77 years) at XXX. These participants overlapped with the participant sample reported in the Section 2 of the Supplementary Materials. They received course credit for their participation. The study was approved by the local ethics committee. Participants were presented with the 45 exemplars (15 x 3 identities), blocked by speaker identity. The order of identity blocks and order of stimuli within each block were randomised. Participants provided ratings of perceived likeness on a scale from 1-7 ("*How much does this sound like [character name]?*"; 1 = not at all, 7 = very much). They were asked to rate the quality of the voice and disregard the verbal content of the stimuli. From these ratings, mean ratings of likeness were computed per exemplar. Likeness ratings for VoiceID1 were lost from one participant due to a technical error. No participants were excluded. Figure 3 illustrates the substantial variability in perceived likeness for different items of the same speaker.

To explore the relationship between average cluster size and perceived likeness, we ran correlation analyses: Kendall's τ_a was computed for average likeness ratings per item and average cluster size per item, separately for unfamiliar listeners and for familiar listeners and for each voice. Significance was determined through random permutation tests (5000 iterations). If the observed value of Kendall's τ_a was higher than 95% of the chance predictions ($p < 0.05$), we rejected the null hypothesis.

For familiar listeners, correlations were not significant after correcting for multiple comparisons for VoiceID1 (Kendall's τ_a = .352, p = .035) and VoiceID 3 (Kendall's τ_a = .324, p = .040). The items with lowest ratings of likeness were, however, also clearly the items with the smallest average cluster size (see Figure 3). The correlation for VoiceID2 was non-significant (Kendall's τ_a = .095, p = .298) due to a ceiling effect (see Figure 3 in the main text). The correlations for unfamiliar listeners were not significant (Kendall's τ_a < .115, p > .277). Despite the lack of significant results, this analysis nonetheless shows interesting trends pointing towards a relationship between perceived likeness and the ability to group items together that future studies can explore.

References

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

Cartei, V., Cowles, H. W., & Reby, D. (2012). Spontaneous voice gender imitation abilities in adult speakers. *PloS one*, *7*(2), e31353.

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

Kuhn, L. K., Wydell, T., Lavan, N., McGettigan, C., & Garrido, L. (2017). Similar representations of emotions across faces and voices. *Emotion*, 17(6), 912-937.

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.

Ritchie, K. L., Kramer, R. S., & Burton, A. M. (2018). What makes a face photo a 'good likeness'?. *Cognition*, *170*, 1-8.