Breaking voice identity perception: Expressive voices are more confusable for listeners

Journal:	*Quarterly Journal of Experimental Psychology*
Manuscript ID	QJE-STD-18-292.R1
Manuscript Type:	Standard Article
Date Submitted by the Author:	05-Dec-2018
Complete List of Authors:	Lavan, Nadine; Royal Holloway University of London, Psychology Burston, Luke; Royal Holloway University of London, Psychology Ladwa, Paayal; Royal Holloway University of London, Psychology Merriman, Siobhan; Royal Holloway University of London, Psychology Knight, Sarah; Royal Holloway University of London, Psychology McGettigan, Carolyn; Royal Holloway, University of London, Psychology
Keywords:	within-person variability, expressiveness, sorting task, voice identity

DOI: 10.1177/1747021819836890

Breaking voice identity perception: Expressive voices are more confusable for listeners

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

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

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

Number of words: 4907

This paper has been posted as a preprint of PsyArXiv: https://psyarxiv.com/mq587

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

DOI: 10.1177/1747021819836890

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

The human voice is a highly flexible instrument for self-expression, yet voice identity perception is largely studied using controlled speech recordings. Using two voice sorting tasks with naturally-varying stimuli, we compared the performance of listeners who were familiar and unfamiliar with the TV show *Breaking Bad*. Listeners organized audio clips of speech with 1) low and 2) high expressiveness into perceived identities. We predicted that increased expressiveness (e.g. shouting, strained voice) would significantly impair performance. Overall, while unfamiliar

DOI: 10.1177/1747021819836890

listeners were less able to generalise identity across exemplars, the two groups performed equivalently well when telling voices apart. However, high vocal expressiveness significantly impaired telling apart in both groups: this led to increased *misidentifications*, where sounds from one character were assigned to the other. Our data suggest that vocal flexibility has powerful effects on identity perception, where changes in the acoustic properties of vocal signals introduced by expressiveness lead to effects apparent in familiar and unfamiliar listeners alike. At the same time, expressiveness appears to have affected other aspects of voice identity processing selectively in one listener group but not the other, thus revealing complex interactions of stimulus properties and listener characteristics (i.e. familiarity) in identity processing.

Keywords: within-person variability, voice identity, sorting task, expressiveness

DOI: 10.1177/1747021819836890

Introduction

We find it intuitive that we should be able to recognise people from their voices alone: before technology could identify a caller from their number, we expected our friends to identify us from “Hello, it’s me” on the phone. The voice as a signal is, however, highly variable, meaning that the same person can sound very different depending on the context in which they are speaking (e.g. conversational speech vs. shouting vs. singing; see also Lavan, Burton, Scott & McGettigan, 2018a). Conversely, two voices that may not sound alike in one context, may suddenly be hard to distinguish in another – voice impersonators generate these confusions professionally. Such within-person variability has important consequences for vocal identity processing: listeners are not only faced with the challenge of telling different voices apart, but they also need to generalise percepts of identity across highly variable vocal signals to maintain perceptual constancy (i.e. “telling people together”; see Burton, 2013 for faces).

Previous studies have shown that while listeners can readily discriminate between unfamiliar voices and recognise familiar(ised) voices under some conditions (see

DOI: 10.1177/1747021819836890

Kreiman & Sidtis, 2011 and Mathias & Von Kriegstein, 2013 for reviews), many factors can affect voice identity processing, rendering it at times highly unreliable. For example, identity processing has been shown to be less accurate for some vocalisations compared to others for both familiar and unfamiliar listeners: speaker discrimination and recognition is less accurate for whispered speech compared to voiced speech (Bartle & Dellwo, 2015; Yarmey, Yarmey, Yarmey & Parliament, 2001). Similarly, it has been shown that speaker discrimination is less reliable from spontaneous vocalisations compared to volitional sounds (for laughter, see Lavan, Scott & McGettigan, 2016; Lavan, Short, Wilding & McGettigan, 2018b). Finally, speaker identification and discrimination are more challenging in speech in an unfamiliar language compared to a familiar language (Zarate, Tian, Woods & Poeppel, 2015; Winters, Levi & Pisoni, 2008). Unfamiliar listeners have furthermore been shown to struggle to accurately generalise identity information across signals that include within-person variability: performance drops when listeners made judgements across different languages (Wester, 2012), divergent linguistic content (Narayan, Mak & Bialystok, 2017), different types of vocalisation (Lavan et al., 2016), disguised and undisguised voices (Reich & Duke, 1979) and across sung vs. spoken

DOI: 10.1177/1747021819836890

speech (Peynircioğlu, Rabinovitz, & Repice, 2017). While voice identity processing for familiar voices is usually more robust to disruption introduced by within-person variability, there are nonetheless striking examples of when familiar voice processing fails: familiar individuals are not well recognised when speaking in a falsetto voice (Wagner & Köster, 1999) or when listeners make speaker discrimination judgements across different types of vocalisation (Lavan et al., 2016).

Recently, a voice sorting task has reported striking differences between familiar and unfamiliar voice identity processing in the context of natural-within person variability within the same task (Lavan, Burston & Garrido, 2018c). Listeners were asked to sort 30 exemplars of two voices into clusters according to perceived identity. These stimuli crucially varied naturally, that is, stimuli were sampled from scenes in a popular TV show and thus featured different speaking styles and environments (see 'ambient images' for faces, e.g. Jenkins, White, Van Montfort & Burton, 2011). Listeners who were unfamiliar with the TV show formed significantly more clusters than familiar listeners (i.e. they perceived more identities). While familiar and unfamiliar listeners' performance for "telling people apart" was comparable, the

DOI: 10.1177/1747021819836890

differences in the number of clusters formed could be linked to a selective failure in "telling people together" for unfamiliar listeners (i.e. failing to perceive different exemplars of the same voice as belonging to the same identity). This study thus replicated previous findings from face sorting tasks (Jenkins et al., 2011, Zhou & Mondloch, 2016). Sorting tasks provide a powerful method to explore identity processing for naturally-varying voices, while also allowing for comparisons of familiar and unfamiliar participants' behaviour within the same task. For face perception, sorting tasks have recently been used to probe more nuanced aspects of identity processing: Zhou and Mondloch (2016) report an other-race effect in a face sorting task for unfamiliar but not familiar participants. Redfern and Benton (2017) used a sorting task to investigate the role of facial expressiveness on identity perception using naturally-varying pictures of individuals unknown to the participants: when contrasting high-expressiveness with low-expressiveness faces in two sorting tasks, viewers made significantly more errors for "telling people apart" when sorting highly expressive faces, by mixing pictures of different people into a single perceived identity. There was no effect on the overall number of clusters made.

DOI: 10.1177/1747021819836890

The current study is a novel exploration of the role of expressiveness in voice identity perception, building on and extending Redfern and Benton's (2017) face sorting study. We contrasted speech that was either low-expressiveness (neutral/conversational speech) or high-expressiveness (speech that deviates from neutral/conversational speech) using voice sorting tasks. When voices become expressive, their acoustic and perceptual properties change dramatically compared to neutral, conversational speech (e.g. Juslin & Laukka, 2003; Banse & Scherer, 1996 for emotional speech). For example, angry shouting may raise the average pitch of speech, increase loudness and introduce 'roughness' (Arnal, Flinker, Kleinschmidt, Giraud & Poeppel, 2015). For a fearful whisper, on the other hand, no (or few) canonically voiced speech segments are present, but the speech rate may increase compared to neutral speech (Ito, Takeda & Itakura, 2005). Aside from such acoustic and perceptual differences, low-expressiveness and high-expressiveness speech differ in their prevalence in everyday life: highly expressive speech is likely to occur less frequently than low-expressiveness speech, possibly leading to impoverished representations of this type of speech (e.g. Lavan et al. 2016, Lavan et al., 2018a for discussions).

DOI: 10.1177/1747021819836890

In the current study, familiar and unfamiliar listeners completed two voice sorting tasks: in each, we asked listeners to sort 30 exemplars of either high-expressiveness or low-expressiveness speech from two voices (15 exemplars per voice) into clusters, according to perceived identity. We predicted that familiar listeners would form fewer clusters than unfamiliar listeners, and that unfamiliar listeners would selectively fail to accurately "tell people together" (Lavan et al., 2018c; Jenkins et al., 2011). We furthermore predicted that while expressiveness would not affect the total number of clusters formed, unfamiliar listeners in particular would make more errors in "telling people apart", by mixing identities within clusters (see Redfern & Benton, 2017). Making judgements across different types of vocalisation has been shown to affect familiar and unfamiliar listeners alike (Lavan et al., 2016): we therefore finally predicted that familiar listeners would also be affected by high expressiveness, a relatively less frequent type of speech.

Methods

Participants

DOI: 10.1177/1747021819836890

68 participants completed the study. Sample size was determined to match Lavan et al. (2018). Participants were recruited via social media and the participant pool of the Department of Psychology at Royal Holloway, University of London. Participants were either entered into a prize draw, received course credit or were paid £5 for their participation. The study was approved by the local ethics committee. We recruited familiar and unfamiliar listeners: if participants reported to have watched at least one season of *Breaking Bad*, they were assigned to the familiar group: these participants had watched 4.6 seasons on average, with last viewing times ranging from a recently as the day of testing to around 5 years ago. Participants who reported to have not seen any episodes of the TV show were assigned to the unfamiliar group. A number of participants were excluded based on the following criteria: familiar participants were excluded if they reported that they had recognised or remembered more than 3 of the specific exemplars included in the sorting tasks ($N = 3$). The average number of exemplars remembered after exclusions was matched across sorting tasks, with listeners remembering on average 0.69 exemplars for the low expressiveness task and 0.62 exemplars for the high expressiveness task, and is thus unlikely to bias our data with regard to the main contrast of high vs low expressiveness. Unfamiliar

DOI: 10.1177/1747021819836890

participants were excluded if they recognised the voice of one of the actors from another TV show/movie (N = 5 – all of them recognised the actor Bryan Cranston's voice). Additionally, we excluded participants whose performance (indexed by number of perceived identities; see below) differed by more than 3 standard deviations from the mean of their listener group (N = 2) and/or who failed the attention checks (see below) in either task (N = 2). This resulted in a final data set of 29 familiar (21 female, mean age: 22.52 years, SD: 6.64 years) and 27 unfamiliar participants (15 female, mean age: 20.4 years, SD: 2.26 years).

Materials

Short audio clips, containing speech of low and high expressiveness from two of the prominent characters of TV show *Breaking Bad* (Hank Schrader and Walter White), were used in this experiment. To create an initial set of stimuli with high-expressiveness versus low-expressiveness speech, we extracted sound clips that ranged between 1.2 and 4 seconds in duration and contained meaningful utterances, with only minimal background noise and no interference from other voices. Exemplars did not include iconic catchphrases or otherwise diagnostic linguistic

DOI: 10.1177/1747021819836890

information (e.g. referring to a character's job, etc). Exemplars were normalized for peak amplitude (to 0.400 Pa), and 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). Long silences were cut.

Pilot ratings: High versus low expressiveness

40 low-expressive speech exemplars (19 exemplars of Hank Schrader, 21 exemplars of Walter White) and 50 high-expressiveness speech exemplars (30 exemplars of Hank Schrader, 20 exemplars of Walter White) were included in a ratings experiment. 21 listeners (2 male; mean age: 22.2 years) rated all exemplars for their perceived arousal (*'How emotionally aroused was the speaker?*: 1 = very drowsy and sleepy to 7 = very alert and energetic), valence ('*How positive was the clip you heard?*: 1 = very negative to 7 = very positive) and expressiveness ('*How much did the voice sound different from normal speech? For example: shouting, laughing and whispering would be more expressive than conversational speech.*': 1 = normal to 7 = very expressive) using the online platform Qualtrics. Ten "catch" trials were also included: a set of additional sound clips was generated using the

DOI: 10.1177/1747021819836890

online text-to-speech app (https://www.naturalreaders.com/online/), where a synthetic voice asked listeners to give a specific rating for the current trial. One participant was excluded from further analyses as they did not follow the spoken instructions on any of the 10 "catch" trials.

Based on these ratings, a final stimulus set was selected with 15 high-expressiveness and 15 low-expressiveness exemplars per identity. Independent samples t- tests confirmed that the two identities were matched for arousal, expressiveness, and valence within the low-expressiveness and high-expressiveness stimulus sets (all *p*s > .311). We furthermore ensured that high-expressiveness and low-expressiveness stimulus sets are maximally different from each other in perceived expressiveness and arousal (both *p*s < .001). Total duration was additionally matched across high-expressiveness and low-expressiveness stimulus sets (*p*s > .211). To minimise systematic differences in the overall variability between low-expressiveness and high expressiveness sets, we took care to match standard deviations across conditions. We furthermore primarily chose negative- to neutral-valence items for the high variability condition (all rated between 1 and 4, one

DOI: 10.1177/1747021819836890

item: 4.9) to broadly match the range of ratings to the low expressiveness condition. Overall, the high expressiveness exemplars thus mainly consisted of shouting or strained speech. All exemplars had significant voiced portions with the exception of one fully whispered exemplar[1]. The properties of these exemplars are reported in Table 1.

--- Insert Table 1 about here ---

Procedure

The selected exemplars (2 identities [Hank, Walter] x 2 expressiveness [high, low] x 15 exemplars) were then embedded into two Microsoft Powerpoint slides, one including the 30 high-expressiveness exemplars, the other including the 30 low-expressiveness exemplars. Additionally, each slide included 2 identical exemplars spoken by a synthetic female voice (created via the natural reader text-to-speech

[1] This exemplar did not stand out as being particularly difficult to process for familiar or unfamiliar listeners: for unfamiliar listeners, the item's "telling apart" probability (see methods) was .14 (grand average = .13, SD = .05, range = .04 - .25) and its "telling together" probability was .16 (grand average = .19, SD = .05, range = .10 - .27). For familiar listeners, the whispered item's "telling apart" probability was .13 (grand average = .15, SD = .11, range = .04 - .60) and its "telling together" probability was .80 (grand average = .61, SD = .18, range = .17 - .81).

DOI: 10.1177/1747021819836890

synthesis app, see above), saying either “Hello. My name is Laura” or “Hello. My name is Sarah”. These items were 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 exemplars on each slide; see exclusion criteria). On the two slides, 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 (see also Lavan et al., 2018c).

Participants completed this task online via Qualtrics, where they were asked to download the Powerpoint slides described above. Participants were then asked to sort the exemplars into clusters, so that each cluster included the exemplars produced by a single speaker, thus representing a perceived speaker identity. Clusters were formed by dragging and dropping exemplars on the slide. There was no limit on how many times participants could play the sounds, nor was there a time limit on completing the task. The ordering of the tasks was counterbalanced across participants. Please see the supplementary materials for plots showing that there we no meaningful order or learning effects. After completing each ask, listeners then re-

DOI: 10.1177/1747021819836890

uploaded the now sorted Powerpoint slides onto Qualtrics and completed a number of debrief questions (see exclusion criteria).

Results

Number of perceived identities

--- Insert Figure 1 about here ---

The number of clusters formed by each participant on each of the two sorting tasks was analysed (after removing the "catch" items). Shapiro-Wilk tests indicated that data were not normally distributed in most cases. We therefore used non-parametric tests for the following analyses in the R environment using the *coin* package. Familiar listeners perceived significantly fewer identities than unfamiliar listeners for both sorting tasks (High expressiveness. Familiar: Mode = 2, Median = 3, Range = 2-9; Unfamiliar: Mode = 9, Median = 8, Range = 4-15. Low expressiveness. Familiar: Mode = 3, Median = 3, Range = 2-9; Unfamiliar: Mode = 6, Median = 9, Range = 3-16). Mann-Whitney *U* tests confirmed that these differences were significant (High

DOI: 10.1177/1747021819836890

expressiveness: Z = 5.27, p < .001; low expressiveness: Z = 5.27, p < .001). However, there was no difference between the number of clusters formed for high versus low expressiveness, in either familiar or unfamiliar listeners (Familiar: Z = .73, p = .768; Unfamiliar: Z = -.06, p = .476, see Figure 1).

"Telling people apart" versus "telling people together"

To further investigate *how* listeners formed clusters, we created 30x30 item-wise response matrices for each participant, sorted by identity (catch items were excluded). In these participant-wise response matrices, each cell codes for whether the relevant pair of exemplars was placed within the same cluster (coded as 1) or placed in two separate clusters (coded as 0). These matrices are symmetrical across the diagonal and can be conceptually divided into within-person submatrices indexing listeners' performance for "telling people together" and across-person submatrices, indexing listeners' performance for "telling people apart" (see Figure 2b).

DOI: 10.1177/1747021819836890

The group-averaged response matrices are shown in Figure 2a. To explore the effects of familiarity and expressiveness on listeners' performance for "telling people together" and "telling people apart", we computed the participant-wise averages of the within-person and across-person submatrices respectively (see Figure 2b). Perfect performance (i.e. forming two clusters of 15 exemplars, with correct assignment of all exemplars to their corresponding identity) would thus result in an average of 1 for the within-person submatrices and an average of 0 for across-person submatrix (for a detailed description of the analyses, see Lavan et al., 2018c). Shapiro-Wilk tests again indicated that data were not normally distributed in most cases. We therefore used non-parametric tests.

--- Insert Figure 2 about here ---

First, we probed the effect of familiarity on task performance. In line with the analyses of the number of clusters, familiar listeners were better than unfamiliar listeners at "telling exemplars together" for both high and low expressiveness speech, with higher values indexing better performance (Low expressiveness,

DOI: 10.1177/1747021819836890

Familiar: Median = .75; Unfamiliar: Median = .18; High expressiveness: Familiar: Median = .64; Unfamiliar: Median = .19). These differences were significant as confirmed by Mann-Whitney *U* tests (Low expressiveness: $Z = 5.74$, $p < .001$; High expressiveness: $Z = 5.32$, $p < .001$). No obvious differences were apparent for "telling exemplars apart" (Low expressiveness: Familiar = .06; Unfamiliar = .08; High expressiveness: Familiar = .13; Unfamiliar = .11): Mann-Whitney *U* tests confirmed that these differences were not significant (Low expressiveness: $Z = .98$, $p = .163$; High expressiveness: $Z = .14$, $p = .445$). Familiar listeners are thus better at "telling people together" than unfamiliar listeners, for both high- and low-expressiveness exemplars, while performance was comparable between groups for "telling people apart".

We then investigated the effect of expressiveness on task performance within each listener group, by comparing differences in the within- and across-person matrices between the two tasks. For "telling exemplars together", familiar listeners were significantly worse for high-expressiveness speech (Wilcoxon's signed rank: $Z = 3.09$, $p = .001$), while performance for unfamiliar listeners was comparable across

DOI: 10.1177/1747021819836890

the two tasks (Wilcoxon's signed rank: Z = .913, p = .819). For "telling exemplars apart", performance was significantly worse for high-expressiveness speech, for both the familiar and unfamiliar listener groups (Wilcoxon's signed rank, Familiar: Z = 2.95, p = .002, Unfamiliar: Z = 2.37, p = .009). High expressiveness therefore has a detrimental effect on how accurately familiar and unfamiliar listeners can "tell people apart", and appears to negatively impact how well familiar listeners can "tell people together". We note that this lack of an effect for unfamiliar listeners may be due to a "floor" effect in performance: performance for "telling people together" for unfamiliar listeners aligns well with the lowest performance reported in a previous voice sorting study (.19 and .18 respectively and this study, .18 in Lavan et al., 2018c for Set 3).

Misperceptions of identity

--- Insert Figure 3 about here ---

DOI: 10.1177/1747021819836890

Participants appeared to systematically misperceive a number of exemplars as the second voice identity (e.g. Hank perceived as Walter; see Figure 2a for lines that are darker in the within-person submatrices and lighter in the across-person submatrices). The most striking examples of this can be found for familiar listeners in the high-expressiveness task. Here, familiar listeners clustered these exemplars more frequently with the exemplars of the other identity than with the exemplars of the correct identity. To quantify this observation, a misperception index was computed for an exploratory analysis: for each participant, we computed an item-wise average for the across-person cells (see Figure 2b, dark grey areas) and subtracted these from the corresponding item-wise average from the within-person cells (see Figure 2b, light grey areas). This resulted in an index ranging between possible endpoints of -1 to 1 per item, per participant: 1 indicates that a particular item was consistently grouped with all the items from the same voice. In contrast, -1 indicates that a particular item was consistently grouped with all items of the other voice (i.e it was consistently misperceived as the other voice; see Figure 3). A comparison of mean misperception scores per participant for high-expressiveness

DOI: 10.1177/1747021819836890

and low-expressiveness conditions confirmed that identities were overall more confusable for highly expressive speech, in both familiar and unfamiliar listeners (Wilcoxon's signed rank test; familiar listeners: $Z = 6.52$, $p < .001$, unfamiliar listeners: $Z = 6.29$, $p < .001$).

Discussion

Using naturally varying clips from a popular TV show, we investigated how voice identity perception is affected by expressiveness in speech. For both high- and low-expressiveness speech, familiar listeners perceived fewer identities in a voice sorting task than unfamiliar listeners: familiar listeners most frequently perceived the veridical number of two identities for highly expressive speech, and three identities for low-expressiveness speech, while unfamiliar listeners most frequently perceived 9 identities in the highly expressive speech compared to 6 identities for less expressive speech. This study replicates previous findings highlighting that unfamiliar identity perception is highly susceptible to the effects of within-person variability, while familiar voice/face processing remains relatively unaffected (Lavan et al., 2018c; Jenkins et al., 2011). This advantage for familiar listeners was linked to

DOI: 10.1177/1747021819836890

being better able to "tell together" different exemplars from the same voice. Unfamiliar listeners, on the other hand, frequently split the exemplars of a single voice identity into different clusters. In contrast, both listener groups performed equally well for "telling people apart". This pattern of results may indicate a bias in unfamiliar listeners, who in the absence of a person-specific representation of a voice are likely to assess any acoustic differences between exemplars as cues to dealing with separate identities, thus frequently perceiving within-person variability as between-person variability. In contrast, familiar listeners have access to a person-specific representation that is likely to include information on how this voice varies (see Burton, Kramer, Ritchie & Jenkins, 2014 for faces). By accessing a person-specific representation, familiar listeners are thus able to largely overcome this perceptual bias, allowing them to "tell people together", leading to more accurate perception of identity.

While expressiveness did not have an effect on the total number of identities perceived, more detailed analyses of listeners' responses revealed that for highly expressive speech both familiar and unfamiliar listeners more frequently failed to "tell

DOI: 10.1177/1747021819836890

people apart" – that is, listeners more frequently mixed exemplars from the two different voices within a cluster. Familiar listeners' performance for "telling people together" furthermore decreased for highly expressive speech, while there was no significant change in unfamiliar listeners' performance for "telling together". These results align well with previous findings in the face perception literature (Redfern & Benton, 2017), where unfamiliar participants more frequently mixed clusters for high-expressive faces, thus making more errors in "telling people apart". Here, we extend these findings to the auditory modality, and to familiar listeners, which Redfern and Benton (2017) did not include in their study. We therefore show that highly expressive speech also detrimentally affects performance even for listeners who are familiar with the voices.

Which properties of the high-expressiveness clips affected listeners' behaviour? It could be argued that high-expressive and low-expressiveness sets may differ in exemplar variability: despite attempting to match standard deviations and ranges of perceptual properties (see methods section), low-expressiveness exemplars in the current study broadly contained only one speaking style (neutral/conversational

DOI: 10.1177/1747021819836890

speech) with subtler differences in tone (e.g. dismissive, patronising) while the high-expressiveness exemplars included a number of broad styles (e.g. shouting, growling or strained voice). If multiple speaking styles are present, it could be predicted that within-talker variability should be higher and it should thus be harder for listeners to generalise identity information across such variable exemplars (see Lavan et al., 2016). Along this line of reasoning, "telling people together" should therefore be *harder* for high-expressiveness speech due to the increased within-person variability between exemplars: listeners should perceive this within-person variability as between-person variability. Crucially, however, this would also predict that "telling people apart" should be *easier*, as individual exemplars are likely to be more acoustically distinct from one another. However, this explanation was not fully supported by our data: familiar listeners' performance for "telling people together" was indeed negatively affected, but so was "telling people apart". For unfamiliar listeners, performance did not change for "telling people together", while more errors occurred for "telling people apart". An alternative explanation for the current patterns of results is that listeners have greater exposure in everyday life (and also within *Breaking Bad*) to low-expressiveness speech, while high-expressiveness speech

DOI: 10.1177/1747021819836890

(particularly the negatively-valenced speech used here) is relatively less frequent. This interpretation makes predictions in line with the current findings: identity representations are less well-formed for highly expressive speech, resulting in worse performance for all aspects of identity processing (here: “telling people apart” and “telling people together”) compared to what can be achieved from the relatively more exposed conversational, neutral speech. Variability and exposure are, however, notoriously difficult to adequately describe and quantify in naturally-varying stimuli. Furthermore, both are complex concepts in themselves: there are many different types of variability and exposure, some potentially more informative and helpful during identity learning and perception than others. Not much is known to date about these aspects of identity processing and more work is needed to be able to better explain the mechanisms behind effects as the ones reported here.

Intriguingly, we observed exemplars for which familiar listeners consistently misidentified Hank as Walter (and vice versa). None of the exemplars was deliberately selected to mislead in this way, nor are the actors likely to have intended

DOI: 10.1177/1747021819836890

to sound like one another. Yet these were striking examples of systematic failure in familiar voice perception. Human voices can be extremely variable, such that two exemplars of the same voice can differ dramatically from each other. Conversely this flexibility also means that within-person voice spaces are extensive, and may partially overlap across different voice identities (Lavan et al., 2018a). Thus, a given vocal signal produced by one person may both match a listener's mental representation of the corresponding voice identity, but might also be a sufficiently good fit for another person's voice space (for a mechanistic account of voice identity processing, see Maguinness, Roswandowitz & Von Kriegstein, 2018). Whether this effect is driven by the properties of the stimuli or the listener (or both) remains unclear. It is possible that voices are acoustically more similar to one another when highly expressive: Expressiveness may erase or change idiosyncratic properties of voices at the production stage. Similarly, highly expressive voices are less frequently encountered in everyday life or, if encountered, not primarily processed with regard to the identity (Goggin, Thompson, Strube & Simental, 1991; see Stevenage & Neil, 2014 for a review). Listeners may therefore be less expert at decoding identity from such signals, being less able to perceive the diagnostic differences in the acoustic

DOI: 10.1177/1747021819836890

properties of highly expressive voices compared to the over-exposed less expressive voices. Another factor that may increase errors across all aspects of the task is a lack of informative contextual cues that could help to disambiguate identity in more naturalistic settings (e.g. visual identity cues, preceding speech context). There may also be exemplar- or voice pair-specific effects at play: the voice identities used in the study were relatively similar (middle-aged males, with similar accent) and thus more vulnerable to confusion.

The current study demonstrates the power of natural variation in the voice to significantly disrupt identity perception, even in listeners experienced with the dramatic variations of TV characters' speech. Much is still to be learned about how the physiology and acoustics of the voice are shaped by communicative contexts, and the limits of our capacity to generalise across these. Future studies should further explore how familiarity with a voice or certain vocal signals interacts with stimuli properties (e.g. variability or frequency of occurrence): can a listener reach a level of familiarity with a voice that would lead to perfect performance, no matter

DOI: 10.1177/1747021819836890

what the stimulus? Similarly, it is unclear what drives the sizeable individual differences in performance observed for familiar (and unfamiliar) listeners: any number of factors could be at play, ranging from the recency of exposure, type of exposure (binge-watching vs watching the show over a number of months or years), overall engagement with the TV show or simply individual differences voice identity processing (Aglieri, Watson, Pernet, Latinus, Garrido & Belin, 2017). Conversely, we need to better establish how much (or little) variability unfamiliar listeners can cope with before making the substantial "telling people together" errors observed here and in previous studies. Overall, our findings again highlight the pressing need for within-person variability to be incorporated in theoretical accounts of how voice identities are represented in the human brain (Lavan et al., 2018a).

Supplementary Material

The Supplementary Material is available at: qjep.sagepub.com

References

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

DOI: 10.1177/1747021819836890

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology*, *25*(15), 2051-2056.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, *70*(3), 614-636.

Bartle, A., & Dellwo, V. (2015). Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech. *International Journal of Speech, Language & the Law*, *22*(2), 229-248.

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

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *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.

Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of language familiarity in voice identification. *Memory & cognition*, *19*(5), 448-458.

Ito, T., Takeda, K., & Itakura, F. (2005). Analysis and recognition of whispered speech. *Speech Communication*, *45*(2), 139-152.Boersma, Paul & Weenink, David (2018). Praat: doing phonetics by computer [Computer program].

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

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression and music performance: Different channels, same code?. *Psychological bulletin*, *129*(5), 770-814.

DOI: 10.1177/1747021819836890

Lavan, N., Burston, L., & Garrido, L. (2018c). How many voices did you hear? Natural variability disrupts identity perception in unfamiliar listeners. *British Journal of Psychology*.

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

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.

Lavan, N., Short, B., Wilding, A., & McGettigan, C. (2018b). Impoverished encoding of speaker identity in spontaneous laughter. *Evolution and Human Behavior*, *39*(1), 139-145.

Maguinness, C., Roswandowitz, C., & Von Kriegstein, K. (2018). Understanding the mechanisms of familiar voice-identity recognition in the human brain. *Neuropsychologia*. [e-pub ahead of print].

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

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.

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.

Stevenage, S. V., & Neil, G. J. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica*, *54*(3), 266-281.

DOI: 10.1177/1747021819836890

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

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

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

Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, *15*(3), 283-299.

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.

DOI: 10.1177/1747021819836890

Figure Captions

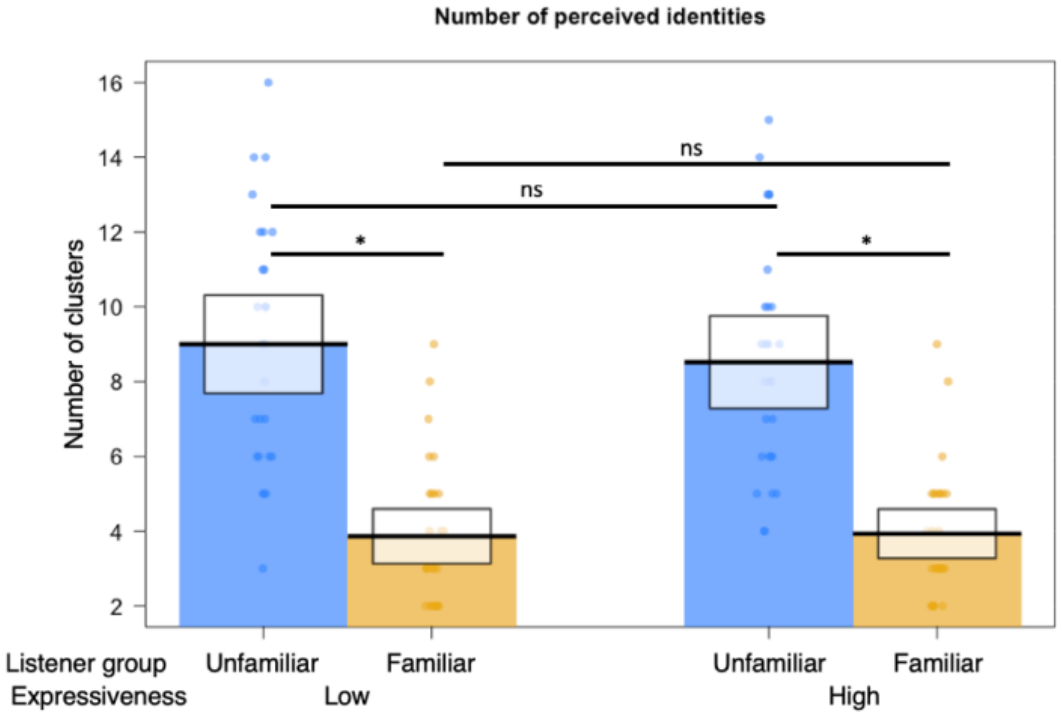
Figure 1. Number of perceived identities per task for familiar and unfamiliar listeners. Bars show the means across participants, and each dot shows the data for one participant. Boxes show the 95% confidence intervals for the means. Stars show significant differences between familiar and unfamiliar listeners (alpha = 0.0125 after correcting for multiple comparisons).

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 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. b) Illustration of the different sections of the per-participant matrices used to compute "telling people together" and "telling people apart" scores.

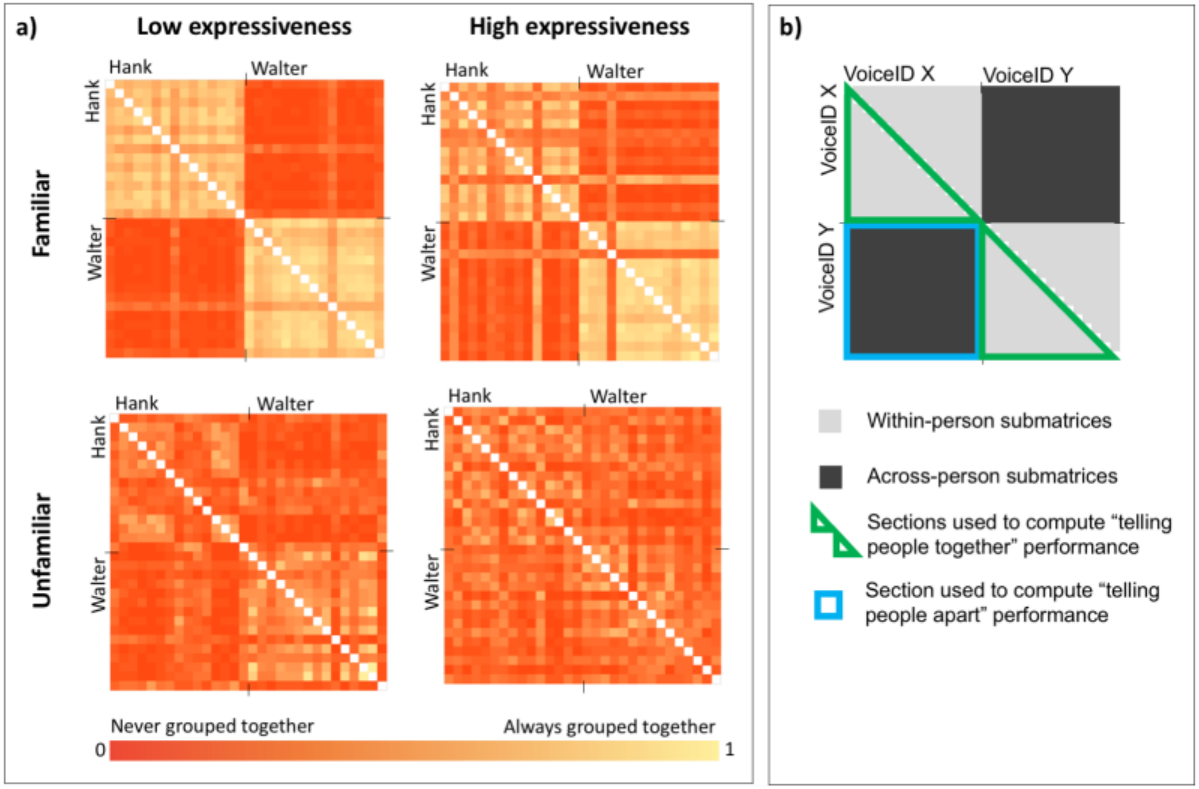
DOI: 10.1177/1747021819836890

Figure 3. Illustration of the misperception index averaged across participants for each of the 30 exemplars for familiar and unfamiliar listeners.

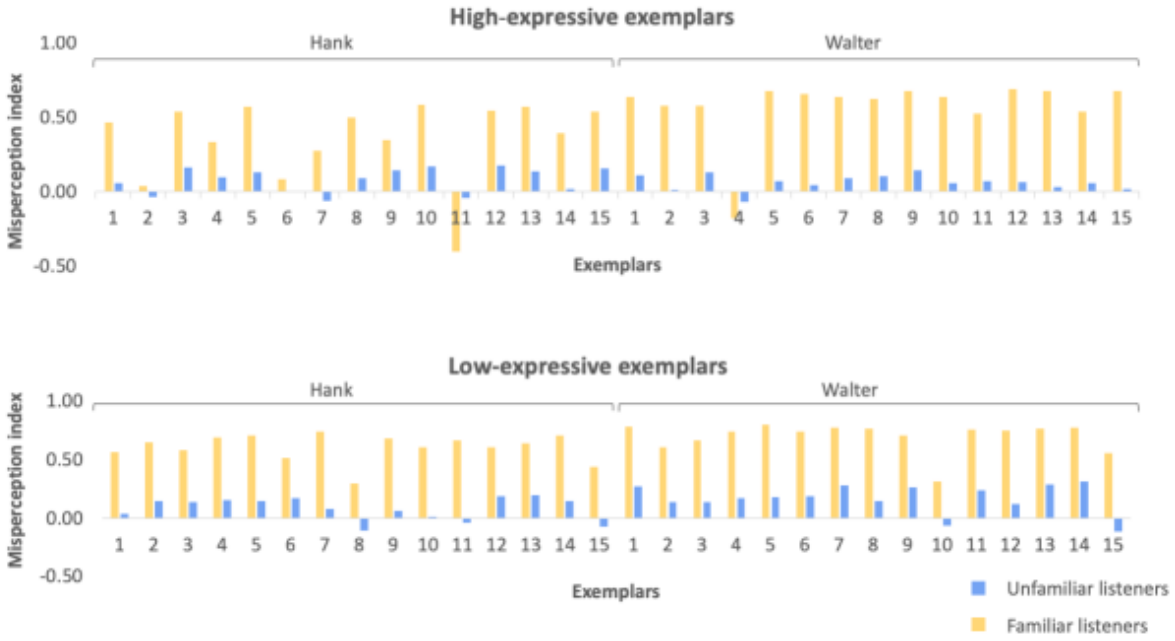
DOI: 10.1177/1747021819836890



DOI: 10.1177/1747021819836890



DOI: 10.1177/1747021819836890



DOI: 10.1177/1747021819836890

Table 1 Breakdown of the means and standard deviations of arousal, expressiveness and valence ratings as well a mean total duration per speaker and per conditions for the final set of exemplars.

	Arousal		Expressiveness		Valence		Duration (secs)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
High-expressiveness	5.6	0.9	5.1	0.7	2.4	0.9	1.9	0.6
Hank	5.7	0.7	5.1	0.6	2.4	0.8	1.8	0.6
Walter	5.4	1.1	5.1	0.8	2.4	0.9	2	0.5
Low-expressiveness	3.6	0.7	2.6	0.6	3.7	0.6	1.7	0.5
Hank	3.7	0.7	2.5	0.6	3.7	0.7	1.7	0.5
Walter	3.5	0.8	2.7	0.6	3.6	0.5	1.6	0.4

DOI: 10.1177/1747021819836890