

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

Journal:	<i>Quarterly Journal of Experimental Psychology</i>
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

SCHOLARONE™
Manuscripts

1
2
3
4 **Breaking voice identity perception: Expressive voices are more**
5 **confusable for listeners**
6
7
8
9

10
11 Nadine Lavan^{1,2}, Luke F.K. Burston², Paayal Ladwa², Siobhan E. Merriman², Sarah
12
13
14
15 Knight¹ and Carolyn McGettigan^{1,2}
16
17

18
19 ¹ *Department of Speech, Hearing and Phonetic Sciences, University College London*
20

21
22 ² *Department of Psychology, Royal Holloway, University of London*
23
24
25
26
27
28

29 Number of words: 4907
30
31
32
33
34
35

36 This paper has been posted as a preprint of PsyArXiv: <https://psyarxiv.com/mq587>
37
38
39
40
41
42

43 Correspondence to:
44
45

46 Nadine Lavan, Department of Speech, Hearing and Phonetic Sciences, University
47
48

49 College London, 2 Wakefield Street, London WC1N 1PF, United Kingdom. E-mail:
50
51

52
53 n.lavan@ucl.ac.uk
54
55
56
57
58
59
60

1
2
3 Or
4
5
6
7
8
9

10 Carolyn McGettigan, Department of Speech, Hearing and Phonetic Sciences,
11
12
13
14 University College London, 2 Wakefield Street, London WC1N 1PF, United Kingdom.
15
16

17 E-mail: c.mcgettigan@ucl.ac.uk
18
19
20
21
22
23

24 Acknowledgements: This work was supported by a Research Leadership Award
25
26
27
28 from the Leverhulme Trust (RL-2016-013) awarded to Carolyn McGettigan
29
30
31
32
33

34 **Abstract**

35
36
37
38 The human voice is a highly flexible instrument for self-expression, yet voice identity
39
40
41
42 perception is largely studied using controlled speech recordings. Using two voice
43
44
45
46 sorting tasks with naturally-varying stimuli, we compared the performance of
47
48
49 listeners who were familiar and unfamiliar with the TV show *Breaking Bad*. Listeners
50
51
52 organized audio clips of speech with 1) low and 2) high expressiveness into
53
54
55
56 perceived identities. We predicted that increased expressiveness (e.g. shouting,
57
58
59 strained voice) would significantly impair performance. Overall, while unfamiliar
60

1
2
3
4 listeners were less able to generalise identity across exemplars, the two groups
5
6
7 performed equivalently well when telling voices apart. However, high vocal
8
9
10 expressiveness significantly impaired telling apart in both groups: this led to
11
12
13 increased *misidentifications*, where sounds from one character were assigned to the
14
15
16 other. Our data suggest that vocal flexibility has powerful effects on identity
17
18
19 perception, where changes in the acoustic properties of vocal signals introduced by
20
21
22 expressiveness lead to effects apparent in familiar and unfamiliar listeners alike. At
23
24
25 the same time, expressiveness appears to have affected other aspects of voice
26
27
28 identity processing selectively in one listener group but not the other, thus revealing
29
30
31 complex interactions of stimulus properties and listener characteristics (i.e.
32
33
34 familiarity) in identity processing.
35
36
37
38
39
40
41
42
43
44

45 **Keywords:** within-person variability, voice identity, sorting task, expressiveness
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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

1
2
3 Kreiman & Sidtis, 2011 and Mathias & Von Kriegstein, 2013 for reviews), many
4
5
6
7 factors can affect voice identity processing, rendering it at times highly unreliable.
8
9

10 For example, identity processing has been shown to be less accurate for some
11
12
13
14 vocalisations compared to others for both familiar and unfamiliar listeners: speaker
15
16
17
18 discrimination and recognition is less accurate for whispered speech compared to
19
20
21
22 voiced speech (Bartle & Dellwo, 2015; Yarmey, Yarmey, Yarmey & Parliament,
23
24
25 2001). Similarly, it has been shown that speaker discrimination is less reliable from
26
27
28
29 spontaneous vocalisations compared to volitional sounds (for laughter, see Lavan,
30
31
32 Scott & McGettigan, 2016; Lavan, Short, Wilding & McGettigan, 2018b). Finally,
33
34
35
36 speaker identification and discrimination are more challenging in speech in an
37
38
39
40 unfamiliar language compared to a familiar language (Zarate, Tian, Woods &
41
42
43 Poeppel, 2015; Winters, Levi & Pisoni, 2008). Unfamiliar listeners have furthermore
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60 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

1
2
3 speech (Peynircioğlu, Rabinovitz, & Repice, 2017). While voice identity processing
4
5
6
7 for familiar voices is usually more robust to disruption introduced by within-person
8
9
10 variability, there are nonetheless striking examples of when familiar voice processing
11
12
13 fails: familiar individuals are not well recognised when speaking in a falsetto voice
14
15
16
17 (Wagner & Köster, 1999) or when listeners make speaker discrimination judgements
18
19
20
21 across different types of vocalisation (Lavan et al., 2016).
22
23
24
25
26
27

28 Recently, a voice sorting task has reported striking differences between familiar and
29
30
31 unfamiliar voice identity processing in the context of natural-within person variability
32
33
34 within the same task (Lavan, Burston & Garrido, 2018c). Listeners were asked to
35
36
37
38 sort 30 exemplars of two voices into clusters according to perceived identity. These
39
40
41
42 stimuli crucially varied naturally, that is, stimuli were sampled from scenes in a
43
44
45 popular TV show and thus featured different speaking styles and environments (see
46
47
48 ‘ambient images’ for faces, e.g. Jenkins, White, Van Montfort & Burton, 2011).
49
50
51
52 Listeners who were unfamiliar with the TV show formed significantly more clusters
53
54
55 than familiar listeners (i.e. they perceived more identities). While familiar and
56
57
58
59 unfamiliar listeners’ performance for “telling people apart” was comparable, the
60

1
2
3 differences in the number of clusters formed could be linked to a selective failure in
4
5
6
7 “telling people together” for unfamiliar listeners (i.e. failing to perceive different
8
9
10 exemplars of the same voice as belonging to the same identity). This study thus
11
12
13 replicated previous findings from face sorting tasks (Jenkins et al., 2011, Zhou &
14
15
16
17 Mondloch, 2016). Sorting tasks provide a powerful method to explore identity
18
19
20 processing for naturally-varying voices, while also allowing for comparisons of
21
22
23
24 familiar and unfamiliar participants’ behaviour within the same task. For face
25
26
27
28 perception, sorting tasks have recently been used to probe more nuanced aspects of
29
30
31
32 identity processing: Zhou and Mondloch (2016) report an other-race effect in a face
33
34
35 sorting task for unfamiliar but not familiar participants. Redfern and Benton (2017)
36
37
38 used a sorting task to investigate the role of facial expressiveness on identity
39
40
41
42 perception using naturally-varying pictures of individuals unknown to the participants:
43
44
45
46 when contrasting high-expressiveness with low-expressiveness faces in two sorting
47
48
49 tasks, viewers made significantly more errors for “telling people apart” when sorting
50
51
52 highly expressive faces, by mixing pictures of different people into a single perceived
53
54
55
56 identity. There was no effect on the overall number of clusters made.
57
58
59
60

1
2
3
4 The current study is a novel exploration of the role of expressiveness in voice identity
5
6
7 perception, building on and extending Redfern and Benton's (2017) face sorting
8
9
10 study. We contrasted speech that was either low-expressiveness
11
12
13 (neutral/conversational speech) or high-expressiveness (speech that deviates from
14
15
16 neutral/conversational speech) using voice sorting tasks. When voices become
17
18
19 expressive, their acoustic and perceptual properties change dramatically compared
20
21
22 to neutral, conversational speech (e.g. Juslin & Laukka, 2003; Banse & Scherer,
23
24
25 1996 for emotional speech). For example, angry shouting may raise the average
26
27
28 pitch of speech, increase loudness and introduce 'roughness' (Arnal, Flinker,
29
30
31 Kleinschmidt, Giraud & Poeppel, 2015). For a fearful whisper, on the other hand, no
32
33
34 (or few) canonically voiced speech segments are present, but the speech rate may
35
36
37 increase compared to neutral speech (Ito, Takeda & Itakura, 2005). Aside from such
38
39
40 acoustic and perceptual differences, low-expressiveness and high-expressiveness
41
42
43 speech differ in their prevalence in everyday life: highly expressive speech is likely to
44
45
46 occur less frequently than low-expressiveness speech, possibly leading to
47
48
49 impoverished representations of this type of speech (e.g. Lavan et al. 2016, Lavan et
50
51
52 al., 2018a for discussions).
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7 In the current study, familiar and unfamiliar listeners completed two voice sorting
8
9
10 tasks: in each, we asked listeners to sort 30 exemplars of either high-expressiveness
11
12
13 or low-expressiveness speech from two voices (15 exemplars per voice) into
14
15
16 clusters, according to perceived identity. We predicted that familiar listeners would
17
18
19
20 form fewer clusters than unfamiliar listeners, and that unfamiliar listeners would
21
22
23 selectively fail to accurately “tell people together” (Lavan et al., 2018c; Jenkins et al.,
24
25
26
27 2011). We furthermore predicted that while expressiveness would not affect the total
28
29
30 number of clusters formed, unfamiliar listeners in particular would make more errors
31
32
33 in “telling people apart”, by mixing identities within clusters (see Redfern & Benton,
34
35
36
37 2017). Making judgements across different types of vocalisation has been shown to
38
39
40 affect familiar and unfamiliar listeners alike (Lavan et al., 2016): we therefore finally
41
42
43 predicted that familiar listeners would also be affected by high expressiveness, a
44
45
46 relatively less frequent type of speech.
47
48
49
50
51
52
53
54
55

56 **Methods**

57 58 59 *Participants* 60

1
2
3
4 68 participants completed the study. Sample size was determined to match Lavan et
5
6
7 al. (2018). Participants were recruited via social media and the participant pool of the
8
9
10 Department of Psychology at Royal Holloway, University of London. Participants
11
12
13 were either entered into a prize draw, received course credit or were paid £5 for their
14
15
16 participation. The study was approved by the local ethics committee. We recruited
17
18
19 familiar and unfamiliar listeners: if participants reported to have watched at least one
20
21
22 season of *Breaking Bad*, they were assigned to the familiar group: these participants
23
24
25 had watched 4.6 seasons on average, with last viewing times ranging from a recently
26
27
28 as the day of testing to around 5 years ago. Participants who reported to have not
29
30
31 seen any episodes of the TV show were assigned to the unfamiliar group. A number
32
33
34 of participants were excluded based on the following criteria: familiar participants
35
36
37 were excluded if they reported that they had recognised or remembered more than 3
38
39
40 of the specific exemplars included in the sorting tasks ($N = 3$). The average number
41
42
43 of exemplars remembered after exclusions was matched across sorting tasks, with
44
45
46 listeners remembering on average 0.69 exemplars for the low expressiveness task
47
48
49 and 0.62 exemplars for the high expressiveness task, and is thus unlikely to bias our
50
51
52 data with regard to the main contrast of high vs low expressiveness. Unfamiliar
53
54
55
56
57
58
59
60

1
2
3 participants were excluded if they recognised the voice of one of the actors from
4
5
6
7 another TV show/movie ($N = 5$ – all of them recognised the actor Bryan Cranston’s
8
9
10 voice). Additionally, we excluded participants whose performance (indexed by
11
12
13 number of perceived identities; see below) differed by more than 3 standard
14
15
16 deviations from the mean of their listener group ($N = 2$) and/or who failed the
17
18
19 attention checks (see below) in either task ($N = 2$). This resulted in a final data set of
20
21
22
23
24 29 familiar (21 female, mean age: 22.52 years, SD: 6.64 years) and 27 unfamiliar
25
26
27 participants (15 female, mean age: 20.4 years, SD: 2.26 years).
28
29
30
31
32
33
34

35 *Materials*

36
37

38 Short audio clips, containing speech of low and high expressiveness from two of the
39
40
41 prominent characters of TV show *Breaking Bad* (Hank Schrader and Walter White),
42
43
44 were used in this experiment. To create an initial set of stimuli with high-
45
46
47 expressiveness versus low-expressiveness speech, we extracted sound clips that
48
49
50 ranged between 1.2 and 4 seconds in duration and contained meaningful utterances,
51
52
53
54 with only minimal background noise and no interference from other voices.
55
56
57
58
59 Exemplars did not include iconic catchphrases or otherwise diagnostic linguistic
60

1
2
3 information (e.g. referring to a character's job, etc). Exemplars were normalized for
4
5
6
7 peak amplitude (to 0.400 Pa), and low-pass filtered at 10kHz (using a Hann pass-
8
9
10 band filter with upper and lower edges 0Hz and 10000Hz, smoothing 20Hz) using
11
12
13 Praat (Boersma & Weenink, 2018). Long silences were cut.
14
15
16
17
18
19
20

21 *Pilot ratings: High versus low expressiveness*

22
23
24 40 low-expressive speech exemplars (19 exemplars of Hank Schrader, 21
25
26
27 exemplars of Walter White) and 50 high-expressiveness speech exemplars (30
28
29
30 exemplars of Hank Schrader, 20 exemplars of Walter White) were included in a
31
32
33 ratings experiment. 21 listeners (2 male; mean age: 22.2 years) rated all exemplars
34
35
36 for their perceived arousal (*'How emotionally aroused was the speaker?': 1 = very*
37
38
39 *drowsy and sleepy to 7 = very alert and energetic*), valence (*'How positive was the*
40
41
42 *clip you heard?': 1 = very negative to 7 = very positive*) and expressiveness (*'How*
43
44
45 *much did the voice sound different from normal speech? For example: shouting,*
46
47
48 *laughing and whispering would be more expressive than conversational speech.'*: 1
49
50
51
52 = normal to 7 = very expressive) using the online platform Qualtrics. Ten "catch"
53
54
55
56 trials were also included: a set of additional sound clips was generated using the
57
58
59
60

1
2
3
4 online text-to-speech app (<https://www.naturalreaders.com/online/>), where a
5
6
7 synthetic voice asked listeners to give a specific rating for the current trial. One
8
9
10 participant was excluded from further analyses as they did not follow the spoken
11
12
13
14 instructions on any of the 10 “catch” trials.
15

16
17
18
19
20
21 Based on these ratings, a final stimulus set was selected with 15 high-
22
23
24 expressiveness and 15 low-expressiveness exemplars per identity. Independent
25
26
27 samples t- tests confirmed that the two identities were matched for arousal,
28
29
30 expressiveness, and valence within the low-expressiveness and high-
31
32
33 expressiveness stimulus sets (all p s > .311). We furthermore ensured that high-
34
35
36 expressiveness and low-expressiveness stimulus sets are maximally different from
37
38
39 each other in perceived expressiveness and arousal (both p s < .001). Total duration
40
41
42 was additionally matched across high-expressiveness and low-expressiveness
43
44
45 stimulus sets (p s > .211). To minimise systematic differences in the overall variability
46
47
48 between low-expressiveness and high expressiveness sets, we took care to match
49
50
51 standard deviations across conditions. We furthermore primarily chose negative- to
52
53
54 neutral-valence items for the high variability condition (all rated between 1 and 4, one
55
56
57
58
59
60

1
2
3 item: 4.9) to broadly match the range of ratings to the low expressiveness condition.
4
5
6

7 Overall, the high expressiveness exemplars thus mainly consisted of shouting or
8
9
10 strained speech. All exemplars had significant voiced portions with the exception of
11
12
13 one fully whispered exemplar¹. The properties of these exemplars are reported in
14
15
16

17 Table 1.
18
19

20
21 --- Insert Table 1 about here ---
22
23
24
25
26
27

28 *Procedure*

29
30

31 The selected exemplars (2 identities [Hank, Walter] x 2 expressiveness [high, low] x
32
33 15 exemplars) were then embedded into two Microsoft Powerpoint slides, one
34
35 including the 30 high-expressiveness exemplars, the other including the 30 low-
36
37
38 expressiveness exemplars. Additionally, each slide included 2 identical exemplars
39
40
41
42
43
44
45 spoken by a synthetic female voice (created via the natural reader text-to-speech
46
47
48
49

50
51 ¹ This exemplar did not stand out as being particularly difficult to process for familiar or unfamiliar
52 listeners: for unfamiliar listeners, the item's "telling apart" probability (see methods) was .14 (grand
53 average = .13, SD = .05, range = .04 - .25) and its "telling together" probability was .16 (grand average
54 = .19, SD = .05, range = .10 - .27). For familiar listeners, the whispered item's "telling apart" probability
55 was .13 (grand average = .15, SD = .11, range = .04 - .60) and its "telling together" probability was .80
56 (grand average = .61, SD = .18, range = .17 - .81).
57
58
59
60

1
2
3 synthesis app, see above), saying either “Hello. My name is Laura” or “Hello. My
4
5
6
7 name is Sarah”. These items were included as attention checks to verify that
8
9
10 participants were completing the task correctly (i.e. by forming a single identity
11
12
13 cluster for the 2 female voice exemplars on each slide; see exclusion criteria). On
14
15
16 the two slides, each embedded sound was represented by a number on the screen.
17
18
19
20
21 These numbers were evenly distributed across the slide, with no clusters being
22
23
24 obvious from the outset (see also Lavan et al., 2018c).
25
26
27
28
29
30

31
32 Participants completed this task online via Qualtrics, where they were asked to
33
34
35 download the Powerpoint slides described above. Participants were then asked to
36
37
38 sort the exemplars into clusters, so that each cluster included the exemplars
39
40
41 produced by a single speaker, thus representing a perceived speaker identity.
42
43
44

45
46 Clusters were formed by dragging and dropping exemplars on the slide. There was
47
48
49 no limit on how many times participants could play the sounds, nor was there a time
50
51
52 limit on completing the task. The ordering of the tasks was counterbalanced across
53
54
55 participants. Please see the supplementary materials for plots showing that there we
56
57
58
59 no meaningful order or learning effects. After completing each ask, listeners then re-
60

1
2
3
4 uploaded the now sorted Powerpoint slides onto Qualtrics and completed a number
5
6
7 of debrief questions (see exclusion criteria).
8
9

10 11 12 13 14 **Results**

15 16 17 *Number of perceived identities*

18
19
20
21
22
23
24 --- *Insert Figure 1 about here* ---
25
26
27
28
29

30
31 The number of clusters formed by each participant on each of the two sorting tasks
32
33 was analysed (after removing the “catch” items). Shapiro-Wilk tests indicated that
34
35 data were not normally distributed in most cases. We therefore used non-parametric
36
37 tests for the following analyses in the R environment using the *coin* package.
38
39
40
41
42
43
44

45 Familiar listeners perceived significantly fewer identities than unfamiliar listeners for
46
47 both sorting tasks (High expressiveness. Familiar: Mode = 2, Median = 3, Range =
48
49 2-9; Unfamiliar: Mode = 9, Median = 8, Range = 4-15. Low expressiveness. Familiar:
50
51 Mode = 3, Median = 3, Range = 2-9; Unfamiliar: Mode = 6, Median = 9, Range = 3-
52
53
54
55
56
57
58
59 16). Mann-Whitney U tests confirmed that these differences were significant (High
60

1
2
3
4 expressiveness: $Z = 5.27$, $p < .001$; low expressiveness: $Z = 5.27$, $p < .001$).

5
6
7 However, there was no difference between the number of clusters formed for high
8
9
10 versus low expressiveness, in either familiar or unfamiliar listeners (Familiar: $Z = .73$,
11
12
13
14 $p = .768$; Unfamiliar: $Z = -.06$, $p = .476$, see Figure 1).
15
16
17
18
19
20

21 *“Telling people apart” versus “telling people together”*
22

23
24 To further investigate *how* listeners formed clusters, we created 30x30 item-wise
25
26
27 response matrices for each participant, sorted by identity (catch items were
28
29
30
31 excluded). In these participant-wise response matrices, each cell codes for whether
32
33
34 the relevant pair of exemplars was placed within the same cluster (coded as 1) or
35
36
37 placed in two separate clusters (coded as 0). These matrices are symmetrical across
38
39
40
41 the diagonal and can be conceptually divided into within-person submatrices
42
43
44 indexing listeners’ performance for “telling people together” and across-person
45
46
47 submatrices, indexing listeners’ performance for “telling people apart” (see Figure
48
49
50
51
52 2b).
53
54
55
56
57
58
59
60

1
2
3
4 The group-averaged response matrices are shown in Figure 2a. To explore the
5
6
7 effects of familiarity and expressiveness on listeners' performance for "telling people
8
9
10 together" and "telling people apart", we computed the participant-wise averages of
11
12
13 the within-person and across-person submatrices respectively (see Figure 2b).
14
15
16 Perfect performance (i.e. forming two clusters of 15 exemplars, with correct
17
18
19 assignment of all exemplars to their corresponding identity) would thus result in an
20
21
22 average of 1 for the within-person submatrices and an average of 0 for across-
23
24
25 person submatrix (for a detailed description of the analyses, see Lavan et al.,
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
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,

1
2
3 Familiar: Median = .75; Unfamiliar: Median = .18; High expressiveness: Familiar:
4
5
6 Median = .64; Unfamiliar: Median = .19). These differences were significant as
7
8 confirmed by Mann-Whitney U tests (Low expressiveness: $Z = 5.74$, $p < .001$; High
9
10
11 expressiveness: $Z = 5.32$, $p < .001$). No obvious differences were apparent for
12
13
14 “telling exemplars apart” (Low expressiveness: Familiar = .06; Unfamiliar = .08; High
15
16
17 expressiveness: Familiar = .13; Unfamiliar = .11): Mann-Whitney U tests confirmed
18
19
20 that these differences were not significant (Low expressiveness: $Z = .98$, $p = .163$;
21
22
23 High expressiveness: $Z = .14$, $p = .445$). Familiar listeners are thus better at “telling
24
25
26 people together” than unfamiliar listeners, for both high- and low-expressiveness
27
28
29 exemplars, while performance was comparable between groups for “telling people
30
31
32 apart”.

33
34
35 We then investigated the effect of expressiveness on task performance within each
36
37
38 listener group, by comparing differences in the within- and across-person matrices
39
40
41
42 between the two tasks. For “telling exemplars together”, familiar listeners were
43
44
45 significantly worse for high-expressiveness speech (Wilcoxon’s signed rank: $Z =$
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60 3.09, $p = .001$), while performance for unfamiliar listeners was comparable across

1
2
3 the two tasks (Wilcoxon's signed rank: $Z = .913$, $p = .819$). For "telling exemplars
4
5
6 apart", performance was significantly worse for high-expressiveness speech, for both
7
8
9 the familiar and unfamiliar listener groups (Wilcoxon's signed rank, Familiar: $Z =$
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

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 ---

1
2
3
4
5 Participants appeared to systematically misperceive a number of exemplars as the
6
7
8 second voice identity (e.g. Hank perceived as Walter; see Figure 2a for lines that are
9
10
11 darker in the within-person submatrices and lighter in the across-person
12
13
14 submatrices). The most striking examples of this can be found for familiar listeners in
15
16
17 the high-expressiveness task. Here, familiar listeners clustered these exemplars
18
19
20 more frequently with the exemplars of the other identity than with the exemplars of
21
22
23 the correct identity. To quantify this observation, a misperception index was
24
25
26 computed for an exploratory analysis: for each participant, we computed an item-
27
28
29 wise average for the across-person cells (see Figure 2b, dark grey areas) and
30
31
32 subtracted these from the corresponding item-wise average from the within-person
33
34
35 cells (see Figure 2b, light grey areas). This resulted in an index ranging between
36
37
38 possible endpoints of -1 to 1 per item, per participant: 1 indicates that a particular
39
40
41 item was consistently grouped with all the items from the same voice. In contrast, -1
42
43
44 indicates that a particular item was consistently grouped with all items of the other
45
46
47 voice (i.e it was consistently misperceived as the other voice; see Figure 3). A
48
49
50 comparison of mean misperception scores per participant for high-expressiveness
51
52
53
54
55
56
57
58
59
60

1
2
3 and low-expressiveness conditions confirmed that identities were overall more
4
5
6
7 confusable for highly expressive speech, in both familiar and unfamiliar listeners
8
9
10
11 (Wilcoxon's signed rank test; familiar listeners: $Z = 6.52$, $p < .001$, unfamiliar
12
13
14 listeners: $Z = 6.29$, $p < .001$).

21 Discussion

22
23
24 Using naturally varying clips from a popular TV show, we investigated how voice
25
26
27 identity perception is affected by expressiveness in speech. For both high- and low-
28
29
30 expressiveness speech, familiar listeners perceived fewer identities in a voice sorting
31
32
33 task than unfamiliar listeners: familiar listeners most frequently perceived the
34
35
36 veridical number of two identities for highly expressive speech, and three identities
37
38
39 for low-expressiveness speech, while unfamiliar listeners most frequently perceived
40
41
42 9 identities in the highly expressive speech compared to 6 identities for less
43
44
45 expressive speech. This study replicates previous findings highlighting that
46
47
48 unfamiliar identity perception is highly susceptible to the effects of within-person
49
50
51 variability, while familiar voice/face processing remains relatively unaffected (Lavan
52
53
54 et al., 2018c; Jenkins et al., 2011). This advantage for familiar listeners was linked to
55
56
57
58
59
60

1
2
3 being better able to “tell together” different exemplars from the same voice.
4
5

6
7 Unfamiliar listeners, on the other hand, frequently split the exemplars of a single
8
9
10 voice identity into different clusters. In contrast, both listener groups performed
11
12
13 equally well for “telling people apart”. This pattern of results may indicate a bias in
14
15
16 unfamiliar listeners, who in the absence of a person-specific representation of a
17
18
19 voice are likely to assess any acoustic differences between exemplars as cues to
20
21
22 dealing with separate identities, thus frequently perceiving within-person variability
23
24
25 as between-person variability. In contrast, familiar listeners have access to a person-
26
27
28 specific representation that is likely to include information on how this voice varies
29
30
31 (see Burton, Kramer, Ritchie & Jenkins, 2014 for faces). By accessing a person-
32
33
34 specific representation, familiar listeners are thus able to largely overcome this
35
36
37 perceptual bias, allowing them to “tell people together”, leading to more accurate
38
39
40 perception of identity.
41
42
43
44
45
46
47
48
49
50
51

52 While expressiveness did not have an effect on the total number of identities
53
54
55 perceived, more detailed analyses of listeners’ responses revealed that for highly
56
57
58 expressive speech both familiar and unfamiliar listeners more frequently failed to “tell
59
60

1
2
3
4 people apart” – that is, listeners more frequently mixed exemplars from the two
5
6
7 different voices within a cluster. Familiar listeners’ performance for “telling people
8
9
10 together” furthermore decreased for highly expressive speech, while there was no
11
12
13 significant change in unfamiliar listeners’ performance for “telling together”. These
14
15
16 results align well with previous findings in the face perception literature (Redfern &
17
18
19 Benton, 2017), where unfamiliar participants more frequently mixed clusters for high-
20
21
22 expressive faces, thus making more errors in “telling people apart”. Here, we extend
23
24
25 these findings to the auditory modality, and to familiar listeners, which Redfern and
26
27
28 Benton (2017) did not include in their study. We therefore show that highly
29
30
31 expressive speech also detrimentally affects performance even for listeners who are
32
33
34 familiar with the voices.
35
36
37
38
39
40
41
42
43
44

45 Which properties of the high-expressiveness clips affected listeners’ behaviour? It
46
47
48 could be argued that high-expressive and low-expressiveness sets may differ in
49
50
51 exemplar variability: despite attempting to match standard deviations and ranges of
52
53
54 perceptual properties (see methods section), low-expressiveness exemplars in the
55
56
57 current study broadly contained only one speaking style (neutral/conversational
58
59
60

1
2
3
4 speech) with subtler differences in tone (e.g. dismissive, patronising) while the high-
5
6
7 expressiveness exemplars included a number of broad styles (e.g. shouting,
8
9
10 growling or strained voice). If multiple speaking styles are present, it could be
11
12
13
14 predicted that within-talker variability should be higher and it should thus be harder
15
16
17 for listeners to generalise identity information across such variable exemplars (see
18
19
20 Lavan et al., 2016). Along this line of reasoning, “telling people together” should
21
22
23
24 therefore be *harder* for high-expressiveness speech due to the increased within-
25
26
27 person variability between exemplars: listeners should perceive this within-person
28
29
30 variability as between-person variability. Crucially, however, this would also predict
31
32
33
34 that “telling people apart” should be *easier*, as individual exemplars are likely to be
35
36
37 more acoustically distinct from one another. However, this explanation was not fully
38
39
40 supported by our data: familiar listeners’ performance for “telling people together”
41
42
43
44 was indeed negatively affected, but so was “telling people apart”. For unfamiliar
45
46
47 listeners, performance did not change for “telling people together”, while more errors
48
49
50 occurred for “telling people apart”. An alternative explanation for the current patterns
51
52
53
54 of results is that listeners have greater exposure in everyday life (and also within
55
56
57 *Breaking Bad*) to low-expressiveness speech, while high-expressiveness speech
58
59
60

1
2
3 (particularly the negatively-valenced speech used here) is relatively less frequent.
4
5

6
7 This interpretation makes predictions in line with the current findings: identity
8
9
10 representations are less well-formed for highly expressive speech, resulting in worse
11
12
13 performance for all aspects of identity processing (here: “telling people apart” and
14
15
16 “telling people together”) compared to what can be achieved from the relatively more
17
18
19 exposed conversational, neutral speech. Variability and exposure are, however,
20
21
22 notoriously difficult to adequately describe and quantify in naturally-varying stimuli.
23
24
25
26 Furthermore, both are complex concepts in themselves: there are many different
27
28
29 types of variability and exposure, some potentially more informative and helpful
30
31
32 during identity learning and perception than others. Not much is known to date about
33
34
35 these aspects of identity processing and more work is needed to be able to better
36
37
38 explain the mechanisms behind effects as the ones reported here.
39
40
41
42
43
44
45
46
47
48
49

50 Intriguingly, we observed exemplars for which familiar listeners consistently
51
52
53 misidentified Hank as Walter (and vice versa). None of the exemplars was
54
55
56 deliberately selected to mislead in this way, nor are the actors likely to have intended
57
58
59
60

1
2
3 to sound like one another. Yet these were striking examples of systematic failure in
4
5
6
7 familiar voice perception. Human voices can be extremely variable, such that two
8
9
10 exemplars of the same voice can differ dramatically from each other. Conversely this
11
12
13 flexibility also means that within-person voice spaces are extensive, and may
14
15
16 partially overlap across different voice identities (Lavan et al., 2018a). Thus, a given
17
18
19 vocal signal produced by one person may both match a listener's mental
20
21
22 representation of the corresponding voice identity, but might also be a sufficiently
23
24
25 good fit for another person's voice space (for a mechanistic account of voice identity
26
27
28 processing, see Maguinness, Roswandowitz & Von Kriegstein, 2018). Whether this
29
30
31 effect is driven by the properties of the stimuli or the listener (or both) remains
32
33
34 unclear. It is possible that voices are acoustically more similar to one another when
35
36
37 highly expressive: Expressiveness may erase or change idiosyncratic properties of
38
39
40 voices at the production stage. Similarly, highly expressive voices are less frequently
41
42
43 encountered in everyday life or, if encountered, not primarily processed with regard
44
45
46 to the identity (Goggin, Thompson, Strube & Simental, 1991; see Stevenage & Neil,
47
48
49 2014 for a review). Listeners may therefore be less expert at decoding identity from
50
51
52 such signals, being less able to perceive the diagnostic differences in the acoustic
53
54
55
56
57
58
59
60

1
2
3 properties of highly expressive voices compared to the over-exposed less expressive
4
5
6
7 voices. Another factor that may increase errors across all aspects of the task is a
8
9
10 lack of informative contextual cues that could help to disambiguate identity in more
11
12
13 naturalistic settings (e.g. visual identity cues, preceding speech context). There may
14
15
16
17 also be exemplar- or voice pair-specific effects at play: the voice identities used in
18
19
20 the study were relatively similar (middle-aged males, with similar accent) and thus
21
22
23
24 more vulnerable to confusion.
25
26
27
28
29
30
31

32 The current study demonstrates the power of natural variation in the voice to
33
34
35 significantly disrupt identity perception, even in listeners experienced with the
36
37
38 dramatic variations of TV characters' speech. Much is still to be learned about how
39
40
41 the physiology and acoustics of the voice are shaped by communicative contexts,
42
43
44 and the limits of our capacity to generalise across these. Future studies should
45
46
47 further explore how familiarity with a voice or certain vocal signals interacts with
48
49
50 stimuli properties (e.g. variability or frequency of occurrence): can a listener reach a
51
52
53
54 level of familiarity with a voice that would lead to perfect performance, no matter
55
56
57
58
59
60

1
2
3 what the stimulus? Similarly, it is unclear what drives the sizeable individual
4
5
6
7 differences in performance observed for familiar (and unfamiliar) listeners: any
8
9
10 number of factors could be at play, ranging from the recency of exposure, type of
11
12
13 exposure (binge-watching vs watching the show over a number of months or years),
14
15
16 overall engagement with the TV show or simply individual differences voice identity
17
18
19 processing (Aglieri, Watson, Pernet, Latinus, Garrido & Belin, 2017). Conversely, we
20
21
22 need to better establish how much (or little) variability unfamiliar listeners can cope
23
24
25
26
27
28 with before making the substantial “telling people together” errors observed here and
29
30
31 in previous studies. Overall, our findings again highlight the pressing need for within-
32
33
34 person variability to be incorporated in theoretical accounts of how voice identities
35
36
37
38 are represented in the human brain (Lavan et al., 2018a).
39
40

41 **Supplementary Material**

42
43
44
45 The Supplementary Material is available at: qjep.sagepub.com
46
47
48
49
50
51

52 **References**

53
54
55
56 Aglieri, V., Watson, R., Pernet, C., Latinus, M., Garrido, L., & Belin, P. (2017). The
57
58 Glasgow Voice Memory Test: Assessing the ability to memorize and recognize
59
60 unfamiliar voices. *Behavior research methods*, 49(1), 97-110.

- 1
2
3
4 Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A. L., & Poeppel, D. (2015).
5 Human screams occupy a privileged niche in the communication
6 soundscape. *Current Biology*, *25*(15), 2051-2056.
7
8
9
10 Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion
11 expression. *Journal of personality and social psychology*, *70*(3), 614-636.
12
13
14 Bartle, A., & Dellwo, V. (2015). Auditory speaker discrimination by forensic
15 phoneticians and naive listeners in voiced and whispered speech. *International*
16 *Journal of Speech, Language & the Law*, *22*(2), 229-248.
17
18
19
20 Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: common
21 dimensions for different vowels and speakers. *Psychological Research*, *74*(1), 110-
22 120.
23
24
25
26
27 Burton, A. M. (2013). Why has research in face recognition progressed so slowly?
28 The importance of variability. *Quarterly Journal of Experimental Psychology*, *66*(8),
29 1467-1485.
30
31
32
33 Burton, A. M., Kramer, R. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from
34 variation: Representations of faces derived from multiple instances. *Cognitive*
35 *Science*, *40*(1), 202-223.
36
37
38
39 Goggin, J. P., Thompson, C. P., Strube, G., & Simental, L. R. (1991). The role of
40 language familiarity in voice identification. *Memory & cognition*, *19*(5), 448-458.
41
42
43
44 Ito, T., Takeda, K., & Itakura, F. (2005). Analysis and recognition of whispered
45 speech. *Speech Communication*, *45*(2), 139-152. Boersma, Paul & Weenink, David
46 (2018). Praat: doing phonetics by computer [Computer program].
47
48
49
50 Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos
51 of the same face. *Cognition*, *121*(3), 313-323.
52
53
54
55 Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expression
56 and music performance: Different channels, same code?. *Psychological*
57 *bulletin*, *129*(5), 770-814.
58
59
60

- 1
2
3 Lavan, N., Burston, L., & Garrido, L. (2018c). How many voices did you hear?
4 Natural variability disrupts identity perception in unfamiliar listeners. *British Journal of*
5 *Psychology*.
6
7
8
9
10 Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2018a). Flexible voices:
11 identity perception from variable vocal signals. *Psychonomic Bulletin and Review*.
12
13
14 Lavan, N., Scott, S. K., & McGettigan, C. (2016). Impaired generalization of speaker
15 identity in the perception of familiar and unfamiliar voices. *Journal of Experimental*
16 *Psychology: General*, *145*(12), 1604-1614.
17
18
19
20 Lavan, N., Short, B., Wilding, A., & McGettigan, C. (2018b). Impoverished encoding
21 of speaker identity in spontaneous laughter. *Evolution and Human Behavior*, *39*(1),
22 139-145.
23
24
25
26 Maguinness, C., Roswadowitz, C., & Von Kriegstein, K. (2018). Understanding the
27 mechanisms of familiar voice-identity recognition in the human
28 brain. *Neuropsychologia*. [e-pub ahead of print].
29
30
31
32
33 Narayan, C. R., Mak, L., & Bialystok, E. (2017). Words get in the way: Linguistic
34 effects on talker discrimination. *Cognitive science*, *41*(5), 1361-1376.
35
36
37
38 Peynircioğlu, Z. F., Rabinovitz, B. E., & Repice, J. (2017). Matching Speaking to
39 Singing Voices and the Influence of Content. *Journal of Voice*, *31*(2), 256-e13.
40
41
42 Redfern, A. S., & Benton, C. P. (2017). Expressive Faces Confuse Identity. *i-*
43 *Perception*, *8*(5), 2041669517731115.
44
45
46 Reich, A. R., & Duke, J. E. (1979). Effects of selected vocal disguises upon speaker
47 identification by listening. *The Journal of the Acoustical Society of America*, *66*(4),
48 1023-1028.
49
50
51
52 Stevenage, S. V., & Neil, G. J. (2014). Hearing faces and seeing voices: The
53 integration and interaction of face and voice processing. *Psychologica*
54 *Belgica*, *54*(3), 266-281.
55
56
57
58
59
60

1
2
3
4 Wagner, I., & Köster, O. (1999). Perceptual recognition of familiar voices using
5 falsetto as a type of voice disguise. *Proceedings of the XIVth International Congress*
6 *of Phonetic Sciences, San Francisco*, 1381-1385.

7
8
9
10 Wester, M. (2012). Talker discrimination across languages. *Speech*
11 *Communication*, 54(6), 781-790.

12
13
14 Winters, S. J., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of
15 bilingual talkers across languages. *The Journal of the Acoustical Society of*
16 *America*, 123(6), 4524-4538.

17
18
19
20 Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., & Parliament, L. (2001).
21 Commonsense beliefs and the identification of familiar voices. *Applied Cognitive*
22 *Psychology: The Official Journal of the Society for Applied Research in Memory and*
23 *Cognition*, 15(3), 283-299.

24
25
26
27
28 Zarate, J. M., Tian, X., Woods, K. J., & Poeppel, D. (2015). Multiple levels of
29 linguistic and paralinguistic features contribute to voice recognition. *Scientific*
30 *reports*, 5, 11475.

31
32
33
34 Zhou, X., & Mondloch, C. J. (2016). Recognizing “Bella Swan” and “Hermione
35 Granger”: No own-race advantage in recognizing photos of famous
36 faces. *Perception*, 45(12), 1426-1429.

Figure Captions

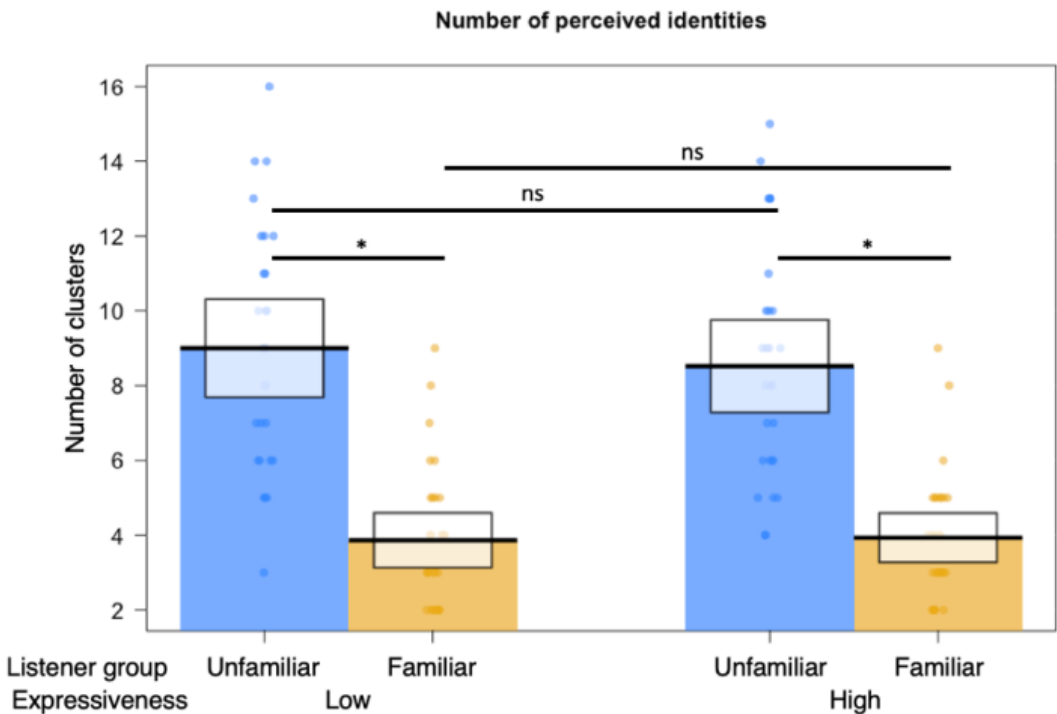
Figure 1. Number of perceived identities per task for familiar and unfamiliar listeners.

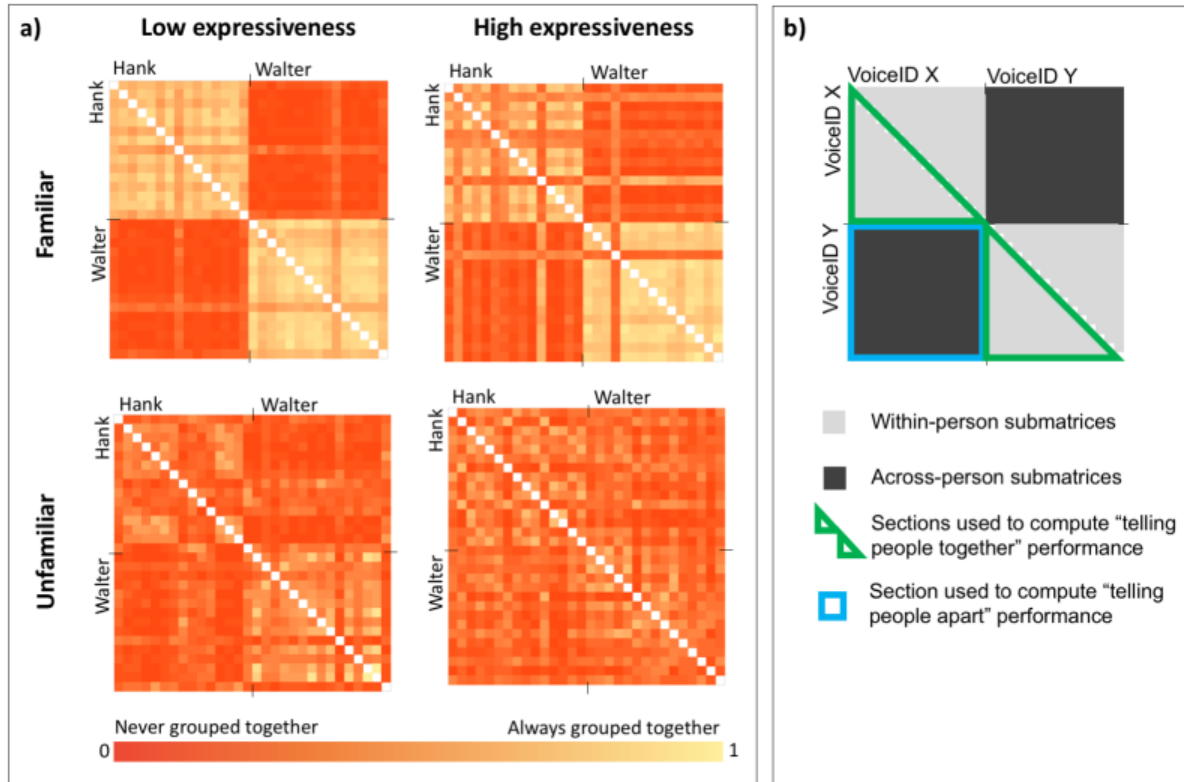
Bars show the means across participants, and each dot shows the data for one participant. Boxes show the 95% confidence intervals for the means. Stars show significant differences between familiar and unfamiliar listeners ($\alpha = 0.0125$ after correcting for multiple comparisons).

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.

1
2
3
4
5
6
7
8 Figure 3. Illustration of the misperception index averaged across participants for
9
10
11 each of the 30 exemplars for familiar and unfamiliar listeners.
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41





1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41

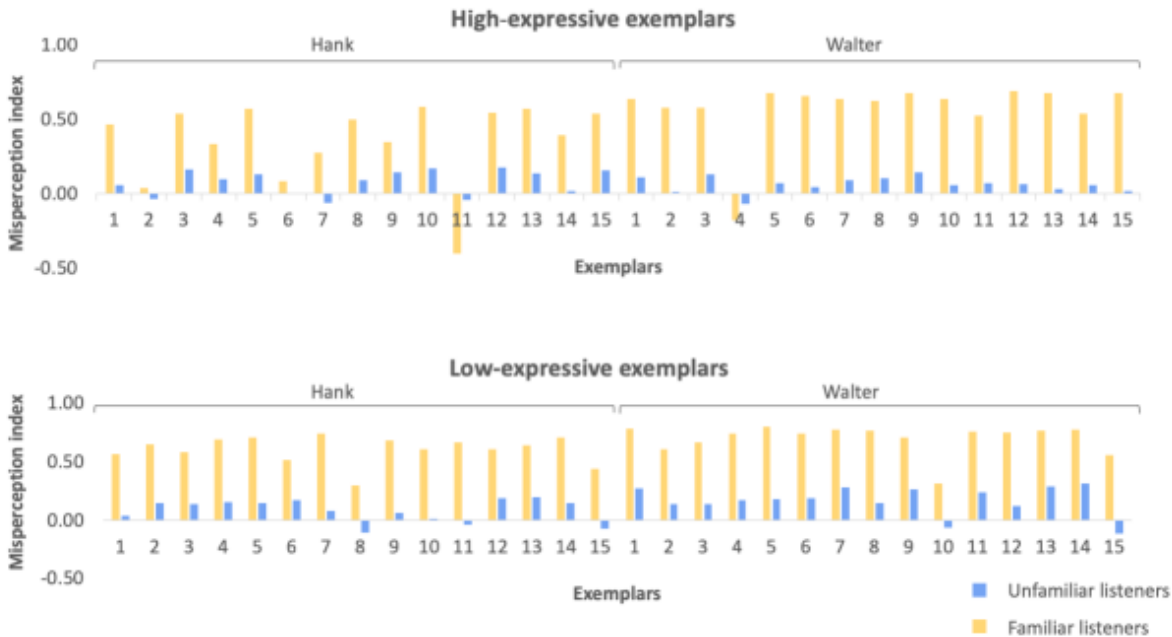


Table 1 Breakdown of the means and standard deviations of arousal, expressiveness and valence ratings as well a mean total duration per speaker and per conditions for the final set of exemplars.

	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