Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices.

Nadine Lavan[1], Sophie K Scott[2], Carolyn McGettigan[1, 2].

[1]Department of Psychology, Royal Holloway, University of London, Egham, UK

[2]Institute of Cognitive Neuroscience, University College London, London, UK

Word count abstract: 206/250

Total word count: 7423

Correspondence should be addressed to: Nadine Lavan, Department of Psychology, Royal Holloway, University of London, Egham Hill, Egham TW20 0EX, UK. E-mail: Nadine.Lavan.2013@rhul.ac.uk

Acknowledgements: The auditory stimulus preparation for Experiment 1 was funded by a Wellcome Trust Senior Research Fellowship [grant number WT090961MA] awarded to Sophie K. Scott.

Abstract

In two behavioural experiments, we explored how the extraction of identity-related information from familiar and unfamiliar voices is affected by naturally occurring vocal flexibility and variability, introduced by different types of vocalizations and levels of volitional control during production. In a first experiment, participants performed a speaker discrimination task on vowels, volitional (acted) laughter, and spontaneous (authentic) laughter from 5 unfamiliar speakers. We found that performance was significantly impaired for spontaneous laughter, a vocalization produced under reduced volitional control. We additionally found that the detection of identity-related information fails to generalize across different types of nonverbal vocalizations (e.g. laughter versus vowels) and across mismatches in volitional control within vocalization pairs (e.g. volitional laughter versus spontaneous laughter), with performance levels indicating an inability to discriminate between speakers. In a second experiment, we explored whether personal familiarity with the speakers would afford greater accuracy and better generalization of identity perception. Using new stimuli, we largely replicated our previous findings: while familiarity afforded a consistent performance advantage for speaker discriminations, the experimental manipulations impaired performance to similar extents for familiar and unfamiliar listener groups. We discuss our findings with reference to prototype-based models of voice processing and suggest potential underlying mechanisms and representations of familiar and unfamiliar voice perception.

Keywords: voice processing; vocal flexibility; generalization; representations; familiarity

Introduction

Voices carry a wealth of information about a speaker: a person's age, sex, emotional state, state of health and identity are all encoded in a voice and can be extracted by listeners with some accuracy (Belin, Fecteau & Bédard, 2004; Lass, Hughes, Bowyer, Waters & Bourne, 1976; Linville, 1996; see also Mathias & von Kriegstein, 2014 for a recent review). Much of what we know about the extraction of identity-related information from voices, be that for explicit identification, recognition, or discrimination of familiar or unfamiliar voices, has been based on speech signals, produced under full volitional control and in a neutral voice – that is, a voice produced with minimal vocal effort, in modal register (e.g. Winters, Levi & Pisoni, 2008 [words]; Schweinberger, Herholz & Sommer, 1997, Kreiman & Papcun, 1991 [extracts from discourse]; Van Lancker & Kreiman, 1987, Perrachione, Del Tufo & Gabriele, 2011 [sentences]). This may have resulted in a somewhat skewed account of the processing of identity-related information, for two reasons.

First, speech is only one of the types of vocal signal used in human communication: non-verbal vocalizations, such as laughter, sighs, and filler sounds (e.g. "erm, uhm") permeate everyday interactions and serve many social and communicative functions. The perceptual properties of such vocalizations have, however, not been widely explored in the literature to date. It should also be noted that speech, when produced in a language familiar to the listener (see Goggin, Thompson, Strube & Simental, 1991; Winters, Levi, & Pisoni, 2008), is uniquely rich in cues to speaker characteristics, including regional accent, lexical content and individual differences in pronunciation. Such speech-specific cues have been shown to be crucial for the extraction of speaker characteristics and identity (e.g. Remez,

Fellowes & Rubin 1997) but are largely absent in non-verbal vocalizations. Restricting previous investigations to speech signals may thus have provided relatively favorable conditions for the extraction of speaker characteristics.

Second, vocal signals are not exclusively produced in a neutral voice. On the one hand, humans can readily change their voices volitionally, for example to convey particular social traits (Cartei, Cowles, & Reby, 2012; Hughes, Mogilski, & Harrison, 2014) and in audience-dependent ways (e.g. the exaggerated pitch contours of infant-directed speech; Shute & Wheldall, 1989). This pronounced flexibility in voice use is illustrated in its extreme by impressionists and voice artists, who can radically change their voices to sound convincingly like a different person – a skill which has no equivalent in, for example, the visual modality (Scott, 2008). On the other hand, transient changes in the voice introduced by involuntary and spontaneous changes in a speaker's state have been shown to drastically affect the vocal output. Authentic emotional experiences are often accompanied by spontaneous vocalizations whose production mechanisms differ dramatically from those employed to produce neutral speech (e.g. Ruch & Ekman, 2001). Due to physiological changes accompanying authentic emotional experiences, spontaneous-vocal signals are affected at both the *source* (sound production by vibration of the vocal folds) and the *filter* (shaping of the source sound by the articulators, including the lips, tongue, jaw, soft palate). Increased subglottal pressure, introduced by the contracting thoracic muscles, is known to modulate the source signal. This can, for example, increase the average fundamental frequency, modulate the spectral features, and introduce non-linearities such as glottal whistles into the vocal signal (Titze, 1988; for spontaneous laughter, see Bryant & Aktipis, 2014; Lavan, Scott, & McGettigan, 2015), while facial

expressions associated with spontaneous vocalizations are also known to modulate the filter characteristics of the vocal tract (e.g. Aubergé & Cathiard, 2003 for smiles). This is in stark contrast to the spoken stimuli of the kind typically used in previous studies of voice perception, which are produced under full volitional control; for example, in neutral speech the spoken source signal remains largely unaffected and stable due to the constant control over the subglottal pressure, and the filter characteristics of the vocal tract are modulated by fine articulation (Draper, Ladefoged, & Whitterridge, 1969). While neutral speech constitutes a significant part of everyday interactions, the apparent flexibility and sources of variability in vocal signals (i.e. different types of vocalizations, effects of ill health or affective state on voice quality, and other changes introduced by variation in control over voice production) have not been accounted for in previous voice perception research. Furthermore, listener performance in previous studies may have benefitted from cues to identity that are specific to speech, such as accent. Thus, experimental approaches to date have only explored a relatively restricted part of vocal identity perception processes and additional research into the effects of vocal variability on voice processing is required.

From a perceptual point of view, accurately extracting constant information such as person identity from a dynamic system such as the voice requires listeners to be able to generalize across highly variable vocal signals. From the speech perception literature, we know that listeners are able to generalize across highly variable realizations of linguistically meaningful information, allowing us to understand speech produced with different accents, dialects or idiosyncratic patterns, with relative ease and high accuracy (speaker normalization, e.g. Johnson, 1987). It could

thus be expected that we are also able to generalize identity-related information across other types of vocal signals, such as emotional vocalizations. There are, however, situations in which the extraction of identity-related characteristics from divergent vocal signals is less robust, for example when attempting to generalize a percept of identity for a multilingual talker producing speech in different languages (unfamiliar voices: Goggin et al., 1991; newly-learned voices: Winters et al., 2008). Earwitness studies have also shown that recognition accuracy decreases when (acted) emotional content is present in a recording at study, but emotionally neutral recordings are presented at test (Saslove & Yarmey, 1980; Read & Craik, 1995). This evidence therefore suggests that in contrast to the robustness of speech content processing (i.e. the extraction of linguistic information), accurate generalizations of identity-related information may be unreliable in the presence of variability in vocal signals.

One proposed mechanism for the extraction of speaker information from vocal signals is that voices are processed in relation to prototypical representations (Kreiman & Sidtis, 2011; Latinus, McAleer, Bestelmeyer & Belin, 2013). According to this model, unfamiliar voices are processed based on their acoustic features in a stimulus-driven way, and compared to prototypical templates based on population averages. In contrast to this, familiar voices are thought to be matched to representations of the specific speaker's vocal inventory stored in long-term memory, possibly in addition to a generic template. When determining speaker identity in perception, the prototypical templates used during unfamiliar voice processing may thus be underspecified, limiting the ability to form generalized percepts in the presence of dramatically different vocalizations and production states. When

processing familiar voices, the detailed and person-specific representations can, however, provide a better fit and would facilitate accurate identity perception despite variability in vocal signals to allow better generalisation. This prediction has been to some extent demonstrated in the familiar talker advantage for speech comprehension, where listeners performed more accurately with speech produced by familiar talkers (e.g. Nygaard & Pisoni, 1998). Whether such familiarity advantages extend from speech comprehension to the processing of identity-related information has, however, not been tested.

In two experiments, we investigated how the natural flexibility of vocal signals (introduced here by manipulating the presence or absence of authentic emotional states) affects the perception of person identity from unfamiliar (Experiments 1 and 2) and familiar voices (Experiment 2). Participants performed a speaker discrimination task judging within- and across-vocalization pairs of vowels, spontaneous laughter ($Laughter_S$) and volitional laughter ($Laughter_V$). We thus address a gap in the literature of voice processing: by using a diverse set of nonverbal vocal signals that more comprehensively represent a speaker's vocal inventory, we were able to gain novel insights into person perception from familiar and unfamiliar voices and extend the evidence from speech-based studies.

Experiments 1: Speaker discrimination in unfamiliar listeners

In a first experiment, we compared identity perception for spontaneous laughter ($Laughter_S$; produced during authentic amusement, under reduced volitional control), volitional laughter ($Laughter_V$; produced on demand, under full volitional control) and vowels (Vowels; produced volitionally as for $Laughter_V$). Participants performed

a same-different speaker discrimination task on pairs of vocalizations, including 4 within-vocalization conditions (Vowels-Vowels, $Laughter_V$-$Laughter_V$, $Laughter_S$-$Laughter_S$, $Laughter_V$-$Laughter_S$) and 2 across-vocalization conditions ($Laughter_V$-Vowels, $Laughter_S$-Vowels). Previous research has shown that laughter is highly variable in its acoustic properties, including voiced, unvoiced and snort-like variations that have been shown to shape the perception of laugher attributes (Bachorowski & Owren, 2001; Bachorowski, Smoski & Owren, 2001). Differences in volitional control over the production of laughter have furthermore been shown to affect the acoustic features of vocalizations and are perceptually salient – listeners can accurately distinguish between volitional and spontaneous vocal signals (Bryant & Aktipis, 2014; Lavan & McGettigan, in revision; Lavan, Scott & McGettigan, 2016). Based on the hypothesis that variability in vocal signals should harm our ability to accurately discriminate between speakers, we predicted that a) speaker discrimination performance would be better for within-vocalization trials compared to across-vocalization trials, b) performance would be impaired for vocalizations produced under reduced volitional control and c) this impairment would be more marked for within-pair mismatches in volitional control (i.e. one vocalization is produced under full control while the other is not). Specific predictions per condition, based on stepwise decreases in accuracy in the presence of these three factors, are illustrated in Figure 1.

--- INSERT FIGURE 1 ABOUT HERE ---

Participants

43 participants (39 female; M_{Age}: 19.2 years; *SD*: 1.1 years; range 19-21 years) were recruited at Royal Holloway, University of London and received course credit for their participation. All participants reported normal or corrected-to-normal vision, and did not report any hearing difficulties. Ethical approval was obtained from the Departmental Ethics Committee at the Department of Psychology, Royal Holloway, University of London. None of the participants were familiar with the speakers used.

Materials

Spontaneous (authentic) laughter ($Laughter_S$), volitional laughter ($Laughter_V$), and series of brief vowel sounds were recorded from 5 talkers (3 male, 2 female, age range 23 – 46 years) in a soundproof, anechoic chamber at University College London. Recordings were obtained using a Bruel and Kjaer 2231 Sound Level Meter fitted with a 4165 cartridge, recorded onto a digital audio tape recorder (Sony 60ES; Sony UK Limited, Weybridge, UK) and fed to the S/PDIF digital input of a PC sound card (M-Audio Delta 66; M-Audio, Iver Heath, UK) with a sampling rate of 22050 Hz. The speakers were seated at a distance of 30 cm at an angle of 15 degrees to the microphone. $Laughter_S$ was elicited from speakers who watched or listened to amusing sound or video clips. Details for this procedure are described by McGettigan et al. (2015). Crucially, speakers reported genuine feelings of amusement during and after the recording of $Laughter_S$. $Laughter_V$ was recorded in the same session as $Laughter_S$, with $Laughter_V$ always being recorded first to avoid carry-over effects. The speakers were instructed to produce natural sounding laughter, without inducing a specific emotional state (see McGettigan et al., 2015). Therefore, $Laughter_V$ was

produced under full volitional control over the voice (and in the absence of amusement), while $Laughter_S$ was produced under reduced volitional control, in response to viewing and listening to amusing stimuli. We were thus able to contrast the perception of the same vocalization (laughter) in these different emotional contexts. The speakers also produced series of short vowels (/a/, /i/, /e/, /u/, /o/, average vowel duration within a stimulus = .35secs) with a relatively stable pitch (*Fo Mean*: 206.4 Hz, *SD*: 78.3 Hz) to preserve a percept of neutral emotional valence. This type of non-emotional stimulus was chosen as its acoustic structure resembles laughter and crying, given all three vocalizations are based on series of vocalic bursts (visit http://www.carolynmcgettigan.com/#!stimuli/c7zu for examples of all vocalizations). Individual vocalization exemplars were extracted from the recordings, normalized for RMS amplitude using PRAAT (www.praat.org/) and saved as uncompressed WAVE files.

In a pilot study, a group of independent listeners (*N* = 13) provided ratings of arousal ("*How aroused is the person producing the vocalization?*", with 1 denoting "*the person is feeling very sleepy and drowsy*" and 7 denoting "*the person is feeling very alert and energetic*"), valence ("*How positive or negative is the person producing this vocalization feeling?*", with 1 denoting "*very negative*" and 7 denoting "*very positive*"), control over the vocalizations ("*How much control did the person have over the production of the vocalization?*", with 1 denoting "*none at all*" and 7 denoting "*full control*") and authenticity ("*How authentic is the vocalization?*", with 1 denoting "*not authentic at all*" and 7 denoting "*very authentic*"). These ratings established that participants reliably rate spontaneous laughter as higher in arousal and authenticity, lower in control over the production of the vocalization, and more positive than their

volitional laughter (Lima et al., in preparation).

Based on the ratings from the pilot study, we selected 25 stimuli per vocalization (5 per speaker). There were marked differences in perceived authenticity between $Laughter_V$ and $Laughter_S$ ($Laughter_V$ *M*: 3.60, CI[3.41, 3.79]; $Laughter_S$ *M*: 4.79, CI[4.42, 5.16]; *t*[24]= 5.829, p < .001). $Laughter_S$ and $Laughter_V$ were significantly higher in arousal than Vowels ($Laughter_V$: *t*[24] = 12.954, p < .001; $Laughter_S$: *t*[24] = 11.181, p < .001), but only marginally different from each other ($Laughter_V$ *M*: 4.39, CI[4.16, 4.62]; $Laughter_S$ *M*: 4.78, CI[4.46, 5.10]; *t*[24]= 1.944, p = .064), There was no perceived difference in valence between the laughter types ($Laughter_V$ *M*: 5.28, CI[4.93, 5.43] $Laughter_S$ *M*: 5.23, CI[4.79, 5.67]; t[24]= -.20, p = .846). Although we attempted to match the average duration of the different vocalizations, there was a marginally significant difference in duration (Vowels *M*: 2.55 secs, CI[2.43, 2.66]; $Laughter_V$ *M*: 2.32 secs, CI[2.17, 2.47]; $Laughter_S$ *M*: 2.41 secs, CI[2.32, 2.61]; one-way repeated measures ANOVA: *F*[2,48]=3.13, p = .053). An overview of the acoustic properties of the stimuli, a breakdown of perceptual measures per speaker, and example stimuli can be found in the supplementary materials.

Design and Procedure

After hearing all stimuli once in a brief task (judging speaker sex) to familiarize participants with the nature of the stimuli, participants performed a speaker discrimination task. Participants heard permutations of pairs of $Laughter_V$, $Laughter_S$, and Vowels, the two sounds being presented sequentially with a pause of 0.7 seconds between them. This yielded 6 conditions: 4 within-vocalization conditions (Vowels-

Vowels, Laughter$_V$-Laughter$_V$, Laughter$_S$-Laughter$_S$, Laughter$_V$-Laughter$_S$), and 2 across-vocalization conditions (Laughter$_V$-Vowels, Laughter$_S$-Vowels). Participants were not pre-informed about the inclusion of spontaneous and volitional laughter in the tasks. There were 50 trials, with 25 trials including the same speaker and 25 trials presenting two sounds from different speakers – this yielded 300 trials in total. The inclusion of across-vocalization conditions allowed us to explore our hypotheses regarding listener's ability to generalize identity information, while within-vocalization conditions allowed us to probe for effects of vocalization type and volitional control over production. No stimuli were repeated during the task, and none of the speakers was known to participants prior to the experiment. The order of presentation for the two sounds within a trial was counterbalanced – for instance, for Vowels-Laughter$_V$ trials, half began with a vowels stimulus and half began with Laughter$_V$. Speaker pairings were fixed across participants. After the presentation of the sounds, participants were asked to indicate via a button press on a keyboard whether they thought the two sounds were produced by the same speaker or by two different speakers.

Results

D' scores were calculated from the raw responses and entered into a one-way repeated measures ANOVA, with 6 levels for Condition. Hit and False Alarm rates of 1 and 0 were adjusted using the formula ((*n* - *0.5*) ÷ *n*) (*n* = number of trials per condition; see Stanislaw & Todorov, 1999) for all analyses. After this adjustment, d' scores can range from 0 to 4.11, with a *d'* score of zero indicating that listeners were not able to discriminate between speakers while gradually higher scores indicate a

greater ability to discriminate between speakers (Stanislaw & Todorov, 1999). There was a significant effect of condition on the *d'* scores ($F[5,220] = 61.12$, $p < .001$, $\eta_p^2 = .59$). To further explore the effects of authentic emotional content as well as the impact of within- and across-vocalization judgements between conditions, we conducted pairwise post-hoc t-tests to assess our predictions of a stepwise decrease in performance introduced by 1) across-vocalization judgements, 2) the presence of vocalizations produced under reduced volitional control and 3) a mismatch in level of volitional control within a pair (i.e. one vocalization produced under full volitional control while the other was not), see Figure 1.

Post-hoc t-tests (8 comparisons, corrected alpha = .006) tested for the predicted pattern illustrated in Figure 1. Predictions were confirmed for all within-vocalization judgements, i.e. Vowels-Vowels, $Laughter_V$-$Laughter_V$ and $Laughter_S$-$Laughter_S$ (*ps* < .001). Following our predictions, performance for $Laughter_V$-$Laughter_S$ was also significantly lower compared to $Laughter_S$-$Laughter_S$ (*p*s ≤ .001). As expected, performance levels for $Laughter_V$-$Laughter_S$ and $Laughter_V$-Vowels were similar ($p = .535$). There was, however, only a marginally significant difference between $Laughter_V$-Vowels and $Laughter_S$-Vowels ($p = .073$; see Figure 2). There was a steep decline in performance across the conditions, being not significantly different from zero for $Laughter_S$-Vowels (one-sample t-test, against 0: $p = .011$, Bonferroni-corrected alpha = .008; all other *p*s < .004), indicating an inability to discriminate signal from noise, see Figure 2.

To directly assess whether speaker discrimination was more accurate for within-vocalization trials compared to across-vocalization trials, we averaged the scores for the four within-vocalization conditions and compared them to the

averaged scores for the two across-vocalization conditions. Participants performed better at discriminating speakers for within-vocalization trials compared to across-vocalization trials ($t[43]= 12.83$, $p < .001$).

--- INSERT FIGURE 2 ABOUT HERE ---

We furthermore ran a response bias analysis to further explore the underlying processes for different trial types by adding up hit rates and false alarm rates in relation to "same" responses across conditions. A score of 1 would indicate no bias, while scores between 0 and 1 indicate a bias towards "different" responses and scores between 1 and 2 indicate a bias towards "same" responses. We entered the bias measure into a one sample t-test (testing against 1), to determine whether any biases observed were significant. This showed that for all within-vocalization conditions, with the exception of $Laughter_V$-$Laughter_S$, there was a significant bias towards responding "same" (all *p*s < .001). For the across-vocalization trials, and $Laughter_V$-$Laughter_S$, there was, however, a significant bias towards responding "different" (all *p*s < .001). This suggests that greater within-pair similarity in vocalization type affected how responses were chosen for judgements of speaker identity (for similar effects of linguistic similarity on response bias, see Narayan, Mak & Bialystock, 2016).

Discussion

For speaker discrimination, previous research using only speech vocalizations reported high probabilities of correct responses for an unfamiliar speaker

discrimination task (> 90% for healthy young adults; Van Lancker & Kreiman, 1987). Our results, however, indicate that vocal signals produced under reduced volitional control, within-pair mismatches in volitional control and the requirement to perform across-category judgements impaired participants' ability to discriminate between speakers. Performance was highest for $Laughter_V$-$Laughter_V$ and Vowels-Vowels – showing that vocalization type (laughter versus vowels) *per se* does not have an impact on performance for within-vocalization trials comprising sounds produced under full volitional control. In the presence of vocalizations produced under reduced volitional control, but in the absence of an across-vocalization judgement and a mismatch in levels of volitional control, performance for $Laughter_S$-$Laughter_S$ was lower compared to Vowels-Vowels and $Laughter_V$-$Laughter_{V,}$ but higher than $Laughter_V$-$Laughter_S$ (additional presence of a mismatch in levels of volitional control) and $Laughter_V$-Vowels (additional across-vocalization judgement). Finally, performance was not significantly different from zero for $Laughter_S$-Vowels, for which all three detrimental factors (presence of vocalizations produced under reduced volitional control, across-vocalization judgement, and a mismatch in degree of volitional control) impaired performance.

This points towards listeners' limited ability to generalize the markers of identity-related information in the presence of natural and meaningful variability (introduced here by differences in volitional control over the production, and communicating emotional content) across different vocal signals from unfamiliar individuals. Studies looking at earwitness accuracy have reported similar findings regarding accurate speaker recognition from speech: when being asked to identify a voice from a line up, identification accuracy in these studies decreases if the

(volitional) emotional content signaled in the voice differs between study and test (Saslove & Yarmey, 1980; Read & Craik, 1995). There is furthermore a body of research that has shown that by manipulating specific acoustic properties of a vocal signal, the processing of identity-related information can be harmed (see Kreiman & Sidtis, 2011, Chapter 5 for a review). The aim of these previous studies was to identify sets of salient acoustic features used by listeners to make inferences about a speaker; these studies can, however, also be interpreted as showing evidence for a lack of generalization across variability in vocal signals (in those cases, introduced by acoustic manipulations). For successful generalization, the effect of manipulations on one parameter should be compensated for with little impact on performance, as listeners are known to rely on a number of potentially speaker-specific acoustic cues when extracting identity-related information (Lavner et al., 2000; Sell, Suied, Elhilali, & Shamma). While studies comparing speaker recognition across different languages have shown decreases in recognition accuracy when speech is presented in a language unfamiliar to the listener (see Introduction), there is nonetheless some retention of ability in the absence of intelligibility (i.e. listeners do not perform at floor).

The specific impairment for the processing of $Laughter_S$, a vocalization produced under reduced volitional control, could be explained in two ways, which are not mutually exclusive. First, from a voice production point of view, our results suggest that in situations involving competing communicative purposes (in this case, encoding indexical properties versus authentic emotional state), the more immediate and salient communicative purpose may be preferentially encoded in the vocal output through changes in the acoustic properties of the sound. These salient

acoustic markers are potentially conveyed at the cost of other otherwise informative cues in the voice. Authentic emotional content may thus be encoded in preference to reliable cues to speaker identity, impacting on our ability to detect and extract such cues. Second, from a perception point of view, authentic emotional content is a highly salient signal that automatically captures attention (Öhmann, Flykt & Esteves, 2001): in the presence of such authentic emotional content in a vocal signal, the processing of highly salient emotional content may be adaptively and automatically prioritized over the extraction of (in this context) minimally salient identity information (Goggin, Thompson, Strube & Simental, 1991; see Stevenage & Neil, 2014 for a review), hence impairing performance on our task. Alternatively, listeners may simply be less frequently exposed to such spontaneous vocalizations in general, resulting in less expertise in processing these vocal signals in fine-grained ways.

Belin and colleagues (2004, see also Belin, Bestelmeyer, Latinus & Watson, 2011) have proposed a model of voice perception based on Bruce and Young's model of face perception (1986). The model is hierarchical in nature: following low-level auditory analyses, affective, speech (linguistic content) and identity-related information are processed in partially dissociable but interacting pathways. With regard to this model, our data provide empirical evidence for interactions between affect and identity processing pathways – crucially, with specifically *authentic* emotional information impairing identity processing. The underlying mechanism driving this interaction may, according to our results, lie with a failure to generalize identity information accurately across variable vocal signals.

Experiments 2: Speaker discrimination in personally familiar and unfamiliar listeners

In the previous experiment, we show that unfamiliar listeners' ability to extract indexical speaker characteristics from a range of vocalizations is strongly affected by the variability in vocal signals introduced by reduced volitional control during production, and further by the demand to perform across-vocalization generalizations in the presence of such variability. However, it is possible that these costs to performance would be reduced for listeners already familiar with the speakers being heard. The face perception literature shows familiarity advantages for identity processing: studies suggest that assessments of identity information from photographs are more accurate for familiar than unfamiliar viewers (Bruce, Henderson, Newman & Burton, 2001; Jenkins, White, van Montfort & Burton, 2011; Ramon & Van Belle, 2016). For speech, it has been shown that speech comprehension is more accurate when listeners are familiar with the talking voice than when they are unfamiliar (Nygaard, 2005) – whether, and how, such a familiarity advantage might extend to the extraction of identity-related information remains to be established.

Given our finding that unfamiliar listeners fail to successfully generalize identity-related information across variable non-verbal vocalizations, we explored whether listeners familiar with the voices would be similarly affected by vocal variability. We recorded a new set of vowels, volitional laughter and spontaneous laughter from 5 lecturers working in the Psychology department at Royal Holloway. Given the presence of male and female speakers in Experiment 1, performance may have been inflated as male and female voices can be easily distinguished from

another (in neutral speech – e.g. Owren, Berkowitz & Bachorowski, 2007). We therefore only included female speakers in this set of stimuli to address this issue. In Experiment 1, the two types of laughter furthermore differed in arousal. In the new set of stimuli, this possible confound was addressed by matching the laughs for arousal. In a replication of Experiment 1, we tested a group of listeners familiar with these speakers (students and other members of the Psychology department) as well as an unfamiliar listener group. Based on the previous research showing familiarity advantages across visual and auditory signals, we predicted overall better performance on speaker discrimination for familiar listeners. Based on the hypothesis that familiar listeners should have a well-formed mental representation of the voices (Kreiman & Sidtis, 2011), we further predicted that familiar listeners should demonstrate a greater ability to generalize across vocalizations.

Participants

46 participants were recruited at Royal Holloway, University of London and received course credit for their participation or were paid at a rate of £7.50 per hour. 23 (16 female; M_{Age}: 31.7 years; *S*D: 10.1 years; range 19-65) of the participants were familiar with the voices of the speakers represented in the stimuli set by virtue of having been lectured by these individuals for between 12 and 28 hours in the past 2-3 terms (dependent on the timing of the testing session) as part of their degree course or having worked in the department for more than 2 two years. 23 unfamiliar participants (17 female; M_{Age}: 20.2 years; *S*D: 1.9 years; range 19-27) were recruited from other departments around campus and had had no exposure to the voices used in the study. All participants reported normal or corrected-to-normal vision, and did

not report any hearing difficulties. Ethical approval was obtained from the Departmental Ethics Committee at the Department of Psychology, Royal Holloway, University of London. One participant from the familiar group was excluded as they reported having general difficulties with recognizing individuals from their faces and voices. One participant from the unfamiliar group was excluded as their average performance across all conditions was at zero, indicating random responses.

Materials

New stimuli were recorded for this experiment. The vocalization types included were identical to the ones used in Experiment 1: $Laughter_S$, $Laughter_V$, and Vowels. The sounds were recorded using the same elicitation procedure described above. 5 talkers (all female, ages range from 29 – 42 years), all lecturers, selected based on their exposure to a subgroup of undergraduate degree students at the department, were recorded in a sound-treated recording booth at Royal Holloway, University of London. Recordings were obtained using a Røde condenser microphone (NT-A) with a sampling rate of 44100 Hz. The output of the microphone was fed into a PreSonus Audiobox which was connected to the USB port of the recording computer. Participants were asked to remain as still as possible during the recordings, but were seated at a distance of about 50cm from the microphone to avoid that any movement associated with intense laughter would interfere with the recordings or move the microphone. All laughs and vowels were extracted from the raw recordings and saved as uncompressed WAVE files. All stimuli of a duration between 1.2 and 3.3 seconds were taken forward into a pilot study to measure the perceptual properties of the stimuli: in a design identical to the one reported for the pilot study and

stimulus selection for Experiments 1, 12 participants rated the perceived arousal of 104 spontaneous laughs, 92 volitional laughs and 105 series of vowels on a 7-point Likert scale. They additionally rated the perceived authenticity of laughter on a 7-point Likert scale.

Based on the ratings from this pilot study (see Experiment 1), we selected 30 stimuli (6 per speaker) per vocalization. There were marked differences in perceived authenticity between $Laughter_V$ and $Laughter_S$ ($Laughter_V$ *M*: 3.17, CI[2.94, 3.41]; $Laughter_S$ *M*: 4.98, CI[4.79, 5.18]; *t*[29]= 12.922, p < .001). $Laughter_S$ and $Laughter_V$ were significantly higher in arousal than Vowels ($Laughter_V$: *t*[29] = 28.590, p < .001; $Laughter_S$: *t*[29] = 35.451, p < .001), but were matched for arousal with each other ($Laughter_V$ *M*: 4.67, CI[4.53, 4.81]; $Laughter_S$ *M*: 4.77, CI[4.63, 4.91]; *t*[1,24]= .929, p = .360). We furthermore matched all vocalizations for overall duration (Vowels *M*: 2.55 secs, CI[2.43, 2.66]; $Laughter_V$ *M*: 1.90 secs, CI[1.71, 2.09]; $Laughter_S$ *M*: 1.92 secs, CI[1.70, 2.14]; one-way repeated measures ANOVA: *F*[2,48]=.501, p = .604). An overview of the acoustic properties, a breakdown of perceptual measures per speaker can be found in the supplementary materials.

Design and Procedure

Participants were tested in individual sessions lasting around one hour. Participants were seated in front of a computer screen, with stimuli being presented at a comfortable volume via headphones (Sennheisser HD 201), using MATLAB (Mathworks, Inc., Natick, MA) with the Psychophysics Toolbox extension (http://psychtoolbox.org/). The testing session comprised three tasks:

Task 1: Perceived number of speakers

This task was designed to introduce listeners to the stimuli used in the main task (speaker discrimination) and thus results are not reported here. Participants were initially presented with all stimuli in randomized order and were asked to listen to all sounds attentively. After the presentation of the sounds, participants were prompted to estimate the number of different speakers they had heard. They were then presented with the stimuli blocked by vocalization (vowels, spontaneous laughter and volitional laughter) and prompted to provide the same judgements. The order of these three blocks was randomized.

Task 2: Speaker recognition from speech

This task was included to assess the familiarity of participants with the speakers. After the completion of Task 1, participants were informed that they had heard 5 different speakers and were asked if they were familiar (yes/no answer) with these speakers based on pictures and the names of the individuals. All familiar participants reported to be familiar with each of the speakers, while none of the unfamiliar listeners reported familiarity. Following this, participants underwent a brief voice (re)familiarization task: they were presented with a brief speech sample of each of the five speakers (a brief excerpt from the rainbow passage [mean duration: 6.6 secs, SD = .49 secs]) while the speaker's name and picture were presented on the screen. After this, participants were presented with 6 sentences (from the BKB corpus; Bench, 2006) from each speaker, as well as their time-reversed versions (i.e. 30 sentences of forward speech and 30 sentences in reversed speech; 60 trials in total, presented in a random order). Reversed versions were included to reduce interference from

speaker-specific accents. Following this, participants were asked to identify the speakers from the speech samples in a 5-way forced choice paradigm, via a prompt on the screen. Trials were timed, giving participants 6 seconds to make a response to each sample.

Task 3: Speaker discrimination from non-verbal vocalizations

The design and procedure of this task were both as used in Experiment 1. Following these tasks, familiar participants were asked to report how familiar they thought they were with each lecturer's speaking voice and laughter, on a scale from 1 (not familiar at all) to 7 (very familiar). These data confirm that familiar listeners indeed perceived themselves to be familiar with the speaking voices ($M_{all\ speakers}$ = 5.04; $SD_{all\ speakers}$ = 1.73; means for individual speakers ranging from 5.91 to 4.54) and their laughter ($M_{all\ speakers}$ = 4.28; $SD_{all\ speakers}$ = 2.03; means for individual speakers ranging from 5.71 to 3.54). Overall, listeners thought they were more familiar with the speaker's speaking voices than their laughter (t[21]= 4.203, p < .001).

Results

Speaker recognition from speech

Raw accuracy responses in percent were analyzed in a 2 (familiar listeners, unfamiliar listeners) x 2 (backward speech, forward speech) repeated measures ANOVA. There were significant main effects of listener group (F[1,42] =61.641, p < .001, η_p^2 = .595) as well as of condition (F[1,42] = 183.959, p < .001, η_p^2 = .814) but no interaction (F[1,42] = .007, p = .935, η_p^2 < .001). Familiar listeners were significantly better at identifying speakers from both backward and forward speech than unfamiliar listeners. For

forward speech, the familiar listeners' performance was close to ceiling (*M* = 89.9% *SD* = 11.5%), again confirming a high familiarity with the speech of the individuals recorded for this stimulus set. Clear above-chance performance (i.e. >20% correct) for unfamiliar listeners (*M* = 59% *SD* = 18.3%) can be explained by the brief familiarization phase that preceded this task. For backward speech, the performance of unfamiliar listeners was close to chance level (*M* = 26% *SD* = 11.2%), while familiar listeners' performance was much higher (*M* = 56%; *SD* = 18.4%), indicating that familiarity with the voice of the speaker for the familiar group goes beyond identification based on idiosyncratic linguistic cues, such as regional accents.

Speaker discrimination from non-verbal vocalizations

D' scores were computed and entered into a 2 (group) x 6 (condition) repeated measures ANOVA. There were significant main effects of listener group ($F[1,42] = 371.399$, $p < .001$, $\eta_p^2 = .898$) as well as of condition ($F[5,210] = 65.004$, $p < .001$, $\eta_p^2 = .607$) but no interaction ($F[5,210] = .263$, $p = .933$, $\eta_p^2 = .006$).

Post-hoc t-tests further explored the effects of condition and listener group. We expected significant advantages for familiar listener across all conditions. Our predictions for condition effects were identical to those for Experiment 1 (see Figure 1). The post-hoc paired t-tests (8 comparisons, corrected alpha = .006) largely replicated the pattern of results per condition shown in Experiment 1 (see Figure 2). In contrast to the findings of Experiment 1, performance for Vowels-Vowels was low and significantly worse compared to $Laughter_V$-$Laughter_V$ ($p < .001$) but consequently was similar to $Laughter_S$-$Laughter_S$ ($p = .774$). Furthermore, performance for $Laughter_V$-$Laughter_S$ and $Laughter_V$-Vowels was different ($p < .001$)

and there was no difference between Laughter$_V$-Vowels and Laughter$_S$-Vowels (p = .573; see Figure 2). In line with our previous findings, participants performed better at discriminating speakers for within-vocalization trials compared to across-vocalization trials (t[44]= 13.23, p < .001), with performance dropping to 0 for unfamiliar listeners in the two across-vocalization conditions (one-sample t-tests, both ps > .116).

Post-hoc independent-samples t-tests (6 comparisons, corrected alpha = .008) were run to explore the effect of group for each condition. This showed, as predicted, a significant advantage for familiar listeners over unfamiliar listeners for all conditions (all $p \leq .004$), with the exception of marginally significant advantages for Vowels-Vowels (p = .011) and Laughter$_V$-Laughter$_V$ (p = .015).

--- INSERT FIGURE 3 HERE ---

In parallel to Experiment 1, we ran a response bias analysis. Collapsing across listener group, one-sample t-tests showed that for all within-vocalization conditions, with the exception of Laughter$_V$-Laughter$_S$, there was a significant bias towards responding "same" (all ps < .001). For the across-vocalization trials there was, however, a significant bias towards responding "different" (all ps < .001). No bias was found Laughter$_V$-Laughter$_S$ (t[43]= .482, p = .632). This replicates the findings of the previous experiment.

Discussion

Experiment 2 formed a replication of Experiment 1 that additionally explored the effect of familiarity on listeners' abilities to generalize identity-related information across diverse vocalizations. Our results were very similar to those in Experiment 1, in terms of overall levels of performance as well as a stepwise decline across conditions. This replication suggests that despite the relatively small number of speakers used, stimulus set effects and influences of speaker idiosyncrasies are limited in our study. No formal analysis of perceptual distinctiveness of the voices was performed for the current set of experiments, since individual participants in the current study were only presented with a subset of all possible speaker and stimulus pairings - in order to adequately assess the distinctiveness of each speaker/stimulus within the context of a pair, we would require data from all participants on all possible pairings (see e.g. Baumann & Belin, 2008). Future studies should, however, explicitly explore how perceptual distinctiveness of different voices (and different vocalizations) interacts with vocal variability. Performance for Vowels-Vowels was noticeably lower in Experiment 2, which could be attributed to the vowel tokens being relatively similar across the Experiment 2 speakers (who were all female and with relatively low average fundamental frequency) in contrast to the differences between male and female speakers in Experiment 1 (e.g. in Fo). Otherwise, no striking differences were found between Experiment 1 and 2, indicating a limited effect of speaker sex in our study.

In line with findings from face and speech perception (Bruce, Henderson, Newman & Burton, 2001; Jenkins, White, van Montfort & Burton, 2011; Nygaard, 2005; Ramon & Van Belle, 2016), we found a consistent advantage for familiar

listeners over unfamiliar listeners, where the former have a greater ability to generalize identity-related information across a range of spontaneous and volitional non-verbal vocalizations in a speaker discrimination task. Our data can be interpreted in line with Kreiman and Sidtis' (2011) proposal regarding the differential processing of familiar and unfamiliar voices: given prior experience with the heard voices, familiar listeners can additionally compare the pairs of vocalizations to speaker-specific templates that entail idiosyncrasies and are based on a range of vocal outputs. Unfamiliar listeners have to rely on averaged, prototypical voice templates only, which may serve well as a heuristic but are underspecified compared to a familiar listener's speaker-specific representations. The increased specificity of representations for familiar voices allows for a more precise fit between the incoming vocal signal and the perceptual template for a speaker, thus listeners can assess identity-related information more accurately, compared to unfamiliar listeners.

No interaction between groups was found, which suggests that despite the general advantage for familiar listeners, the factors implicated in impairing performance in the previous experiment (across-vocalization judgements, the presence of vocalizations produced under reduced volitional control and mismatches in volitional control within a pair) have a similar effect on familiar and unfamiliar listeners. It should, however, be noted that unfamiliar listeners were not able to discriminate between speakers for the across-vocalization conditions ($Laughter_V$-Vowels and $Laughter_S$-Vowels), which could potentially mask interactions. Another consideration here could be the nature of the familiarity of our listeners. We show that familiar listeners were able to recognize the five speakers with very high accuracy based on their speech, which serves as an objective measure of familiarity,

and we also show perceived familiarity with the speaker's voices from self-report: however, the familiar listeners in this study had engaged with these speakers in specific contexts (lectures, professional settings), which may have resulted in a familiarity with the voices that is skewed towards certain kinds of vocal signals (e.g. speech and other volitional vocalizations, with high-intensity spontaneous laughter being rare). This possibility is reflected in subjective familiarity ratings, where familiarity with the speaking voice of each lecturer was rated higher than familiarity with that person's laughter. Arguably, listeners presented with vocal signals from speakers they know in a wider range of contexts (i.e. close friends, partners) may be able to more easily generalize across vocalizations, based on having experienced the speakers' full vocal inventory in a way that is more representative of having learned voice identity through social interaction.

Future studies should attempt to create groups of speakers with different profiles of familiarity and personal relationships to listeners, e.g. partners, friends, acquaintances, celebrities, and strangers. Differential profiles per group could be expected. Familiarity is furthermore a broad concept, encompassing many subcomponents. Additional measures that may tap into these subcomponents, for example perceived distinctiveness of a voice, likability of a voice or speaker, level of personal engagement with the speakers, or frequency of exposure to different vocalizations, could yield further insights into the nature of listener familiarity and point towards some of the underlying factors driving the familiarity advantage.

General Discussion

The human voice is a rich and uniquely variable communicative signal. Its potential for flexibility has been largely neglected in studies of voice perception to date, as these have almost exclusively used speech stimuli produced in a volitional, highly controlled manner. The current study addressed this gap in the literature, by examining voice perception across nonverbal vocalizations that are representative of the flexibility and variability in vocal signals (exemplified here by the degree of volitional control over their production).

We found that the presence of vocalizations produced under reduced volitional control, mismatches in volitional control within a pair, and across-vocalization comparisons decreased performance for speaker discrimination, at times indicating that listeners were not able to discriminate between voices at all. While listeners can display relatively high accuracy in extracting speaker characteristics from a single type of vocalization, our findings strikingly illustrate that they have a rather more limited ability to generalize speaker identity across different kinds of vocal signals. For familiar voices, an overall advantage in the processing of indexical speaker properties can be observed, although performance was affected in similar ways by the condition manipulations. Our findings thus put into perspective our ability to extract identity information from a speaking voice: while speech can encode a wealth of cues to identity, our vocal repertoire is highly variable. Accurately attributing these divergent vocal signals to a single individual becomes challenging without prior familiarity with the person's full vocal inventory.

This familiarity advantage observed in our task may be based on the retrieval and matching of the incoming vocal signal to underlying representations

(prototypical representations for unfamiliar listeners vs. speaker-specific representations of voices for familiar listeners, cf. Kreiman & Sidtis, 2011). It is to date unclear what the nature and degree of abstraction of these prototypical and speaker-specific representations of voices might be. We suggest that listeners encode voices based on abstract representations of the vocal tract and its source and filter properties. With increasing exposure to a voice and its full repertoire, knowledge of speaker-specific vocal tract morphology, and of variation in how the articulators shape vocal outputs under varying levels of volitional control (e.g. speaking different languages versus producing sounds in extreme emotional states or in ill health) are integrated into this percept, allowing listeners to gradually build more robust estimates of the dynamics of the vocal system of that speaker. Representations of voices, be they for familiar individuals or generic prototypes, are furthermore likely to be formed and shaped based on long-term exposure to vocal outputs. It is unclear if representations of familiar voices are qualitatively different from the generic prototypes associated with unfamiliar voice processing. Over time and exposure, the initial perceptual assessment of an unfamiliar voice may evolve to be underpinned by a new speaker-specific representation, while the original generic prototype to which this voice may been compared could remain largely unaffected.

It should be noted that our study explored *familiar voice discrimination*. Theoretical and empirical investigations have traditionally considered voice identity perception in familiar voice recognition and unfamiliar voice discrimination tasks (see Kreiman & Sidtis, 2011; Mathias & Von Kriegstein, 2013 for recent reviews). Thus familiarity as a factor in voice perception has been strongly associated with task type. There is, however, evidence that familiarity with a voice can be perceived in the

absence of recognition (see Hanley, Smith & Hadfield, 1998), and our data suggest that familiarity with voices can affect performance on a speaker discrimination task. In the context of our study, we propose that familiar listeners' prior exposure to the voices has led to the development of speaker-specific expertise, which may interact with different aspects of voice processing, and across a range of tasks.

Evidence from the current study seems to suggest an advantage in identity processing for vocalizations produced under full volitional control for unfamiliar and (moderately) familiar listeners. Volitional vocalizations form the vast majority of human communication, leading to greater exposure and expertise, while vocalizations produced under reduced volitional control are not only comparatively rare but also diverge in terms of production mechanisms from volitional vocalizations. The representations used during unfamiliar voice processing are thought to be averaged voice templates (Kreiman & Sidtis, 2011; Latinus, McAleer, Bestelmeyer & Belin, 2013). It is thus not surprising that the performance for spontaneous laughter, a vocalization diverging from prototypical vocal signals, is impaired. Only intimate familiarity with a speaker's vocal inventory may enable listeners to form sufficiently detailed and reliable representations of a voice, including representations of non-prototypical vocal signals. Without reliable representations of such non-prototypical signals, generalization of identity-related information across a range of vocal signals only seems to be possible to a limited extent.

References

Aubergé, V., & Cathiard, M. (2003). Can we hear the prosody of smile?. *Speech Communication*, *40*(1), 87-97.

Bachorowski, J. A., & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *The Journal of the Acoustical Society of America*, *106*(2), 1054-1063.

Bachorowski, J. A., Smoski, M. J., & Owren, M. J. (2001). The acoustic features of human laughter. *The Journal of the Acoustical Society of America*, *110*(3), 1581-1597.

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

Belin, P., Bestelmeyer, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology*, *102*(4), 711-725.

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences*, *8*(3), 129-135.

Bench, J. (2006). The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *Foundations of Pediatric Audiology*, *13*, 145.

Bruce, V., & Young, A. (1986). Understanding face recognition. *British journal of psychology*, *77*(3), 305-327.

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

Bryant, G. A., & Aktipis, C. A. (2014). The animal nature of spontaneous human laughter. *Evolution and Human Behavior*, *35*(4), 327-335.

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

Draper, M. H., Ladefoged, P., & Whitteridge, D. (1959). Respiratory muscles in speech. *Journal of Speech, Language, and Hearing Research*, *2*(1), 16-27.

Hanley, J. R., Smith, S. T., & Hadfield, J. (1998). I recognise you but I can't place you: An investigation of familiar-only experiences during tests of voice and face recognition. The Quarterly Journal of Experimental Psychology: Section A, 51(1), 179-195.

Herald, S. B., Xu, X., Biederman, I., Amir, O., & Shilowich, B. E. (2014). Phonagnosia: A voice homologue to prosopagnosia. *Visual Cognition*, *22*(8), 1031-1033.

Hughes, S. M., Mogilski, J. K., & Harrison, M. A. (2014). The Perception and Parameters of Intentional Voice Manipulation. *Journal of Nonverbal Behavior*, *38*(1), 107-127.

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

Johnson, K. (2008). Speaker Normalization in Speech Perception. *The handbook of speech perception*, 363.

Kreiman J. & Sidtis D. (2011). *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. Hoboken, NJ: Wiley-Blackwell.

Kreiman, J., & Papcun, G. (1991). Comparing discrimination and recognition of unfamiliar voices. *Speech Communication*, *10*(3), 265-275.

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *The Journal of the Acoustical Society of America*, *59*(3), 675-678.

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

Lavan, N. & McGettigan, C. (under review). *Increased discriminability of authenticity from multimodal laughter is driven by auditory information*. Quarterly Journal of Experimental Psychology.

Lavan, N. Scott, SK. & McGettigan, C. (2016). Laugh like you mean it: Authentic emotional experience modulates acoustic, physiological and perceptual properties of laughter. *Journal of Nonverbal Behaviour*.

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

Lima, C. F., Lavan, N., Evans, S., Chen, S., Boebinger, D., & Scott, S.K. (in preparation). Acoustic correlates of authenticity in laughter and weeping.

Mathias, S. R., & von Kriegstein, K. (2014). How do we recognise who is speaking?. *Front Biosci (Scholar edition)*, *6*, 92.

McAllister, H. A., Dale, R.H.I., & Keay, C. E. (1993). Effects of lineup modality on witness credibility. *Journal of Social Psychology*, 133(3), 365-376.

McGettigan, C., Walsh, E., Jessop, R., Agnew, Z. K., Sauter, D. A., Warren, J. E., & Scott, S. K. (2015). Individual differences in laughter perception reveal roles for mentalizing and sensorimotor systems in the evaluation of emotional authenticity. *Cerebral Cortex*, *25(1),* 246-257.

Narayan, C. R., Mak, L., & Bialystock, E. (2016) Words Get in the Way: Linguistic Effects on Talker Discrimination. *Cognitive Science* doi: 10.1111/cogs.12396.

Nygaard, L. C. (2005). Linguistic and paralinguistic factors in speech perception. *Handbook of speech perception. Oxford: Blackwell Publishers*.

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

Öhman, A., Flykt, A., & Esteves, F. (2001). Emotion drives attention: detecting the snake in the grass. *Journal of experimental psychology: general*, 130(3), 466.

Owren, M. J., Berkowitz, M., & Bachorowski, J. A. (2007). Listeners judge talker sex more efficiently from male than from female vowels. *Perception & psychophysics*, 69(6), 930-941.

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

Ramon, M., & Van Belle, G. (2016). Real-life experience with personally familiar faces enhances discrimination based on global information. *PeerJ*, *4*, e1465.

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

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception and Performance*, *23*(3), 651.

Ruch, W., & Ekman, P. (2001). The expressive pattern of laughter. *Emotion, qualia, and consciousness*, 426-443.

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

Scherer, K. R. (1989). Vocal correlates of emotional arousal and affective disturbance. *Handbook of social psychophysiology*, 165-197.

Schweinberger, S. R., Herholz, A., & Sommer, W. (1997). Recognizing Famous Voices: Influence of Stimulus Duration and Different Types of Retrieval Cues. *Journal of Speech, Language, and Hearing Research*, *40*(2), 453-463.

Scott, S. K. (2008). Voice processing in monkey and human brains. *Trends in cognitive sciences*, *12*(9), 323-325.

Sell, G., Suied, C., Elhilali, M., & Shamma, S. (2015). Perceptual susceptibility to acoustic manipulations in speaker discrimination. *The Journal of the Acoustical Society of America*, *137*(2), 911-922.

Shute, B., & Wheldall, K. (1989). Pitch alterations in British motherese: some preliminary acoustic data. *Journal of Child Language*, *16*(03), 503-512.

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

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, *31*(1), 137-149.

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.

Titze, I. R. (1989). On the relation between subglottal pressure and fundamental frequency in phonation. *The Journal of the Acoustical Society of America*, *85*(2), 901-906.

Van Lancker, D. R., Cummings, J. L., Kreiman, J., & Dobkin, B. H. (1988). Phonagnosia: a dissociation between familiar and unfamiliar voices. *Cortex*, *24*(2), 195-209.

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

Vettin, J., & Todt, D. (2004). Laughter in conversation: Features of occurrence and acoustic structure. *Journal of Nonverbal Behavior*, *28*(2), 93-115.

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, M. J., & Yarmey, A. L. (1996). Accuracy of eyewitness identifications in showups and lineups. *Law and Human Behavior*, *20*(4), 459.

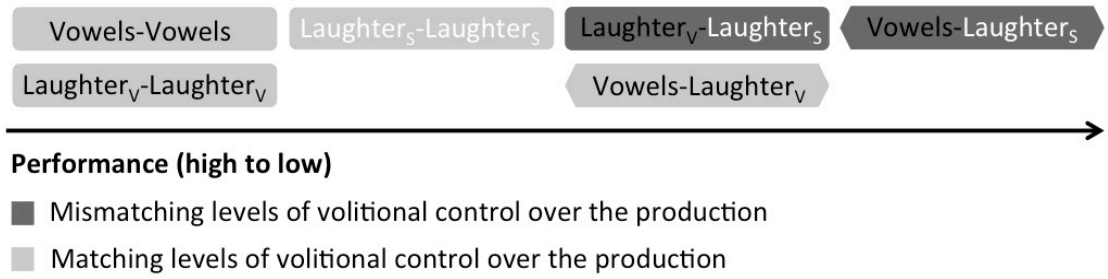

Figure 1 Predicted pattern for performance on the speaker discrimination task (from high performance to low performance). Boxes with rounded edges represent within-vocalization pairs, hexagons represent across-vocalization pairs. Black text: vocalizations produced under full volitional control; white text: vocalizations produced under reduced volitional control. Specific predictions follow the pattern Vowels-Vowels (full volitional control, within-vocalization, matching levels of volitional control) = Laughter$_V$-Laughter$_V$ (full volitional control, within-vocalization, matching levels of volitional control) > Laughter$_S$-Laughter$_S$ (reduced volitional control, within-vocalization, matching levels of volitional control) > Laughter$_V$-Laughter$_S$ (reduced volitional control, within-vocalization, mismatching levels of volitional control) = Laughter$_V$-Vowels (full volitional control, across-vocalization, mismatching emotional content) > Laughter$_S$-Vowels (reduced volitional control, across-vocalization, mismatching levels of volitional control).

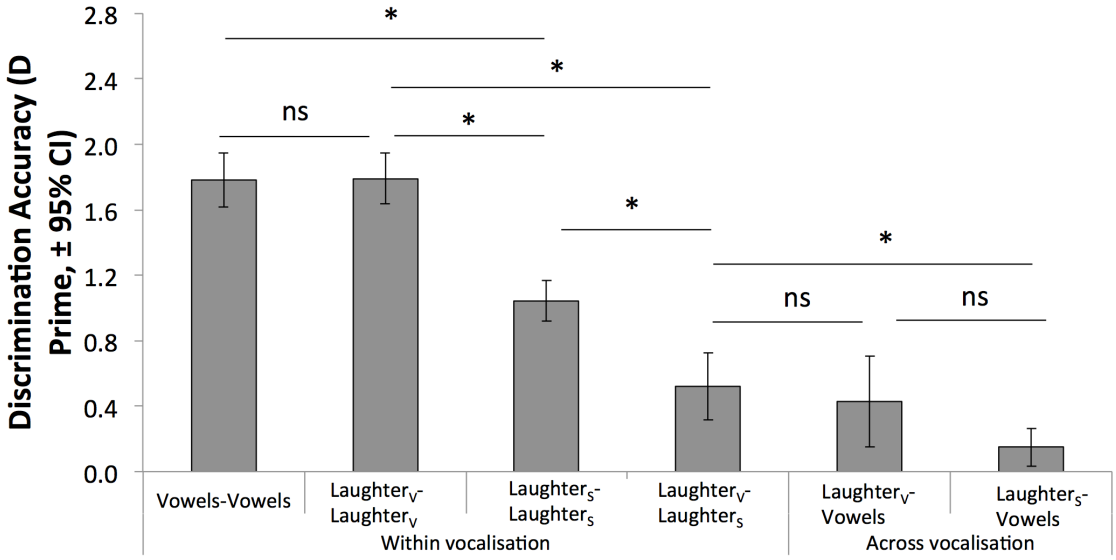

Figure 2 Results for Experiments 1 a) average d' scores per vocalization for the sex identification task, b) average d' scores per condition for the speaker discrimination task. Significant comparisons (Bonferroni-corrected, see Results for alpha levels) are highlighted with an asterisk; marginally significant results are highlighted with an asterisk in brackets.

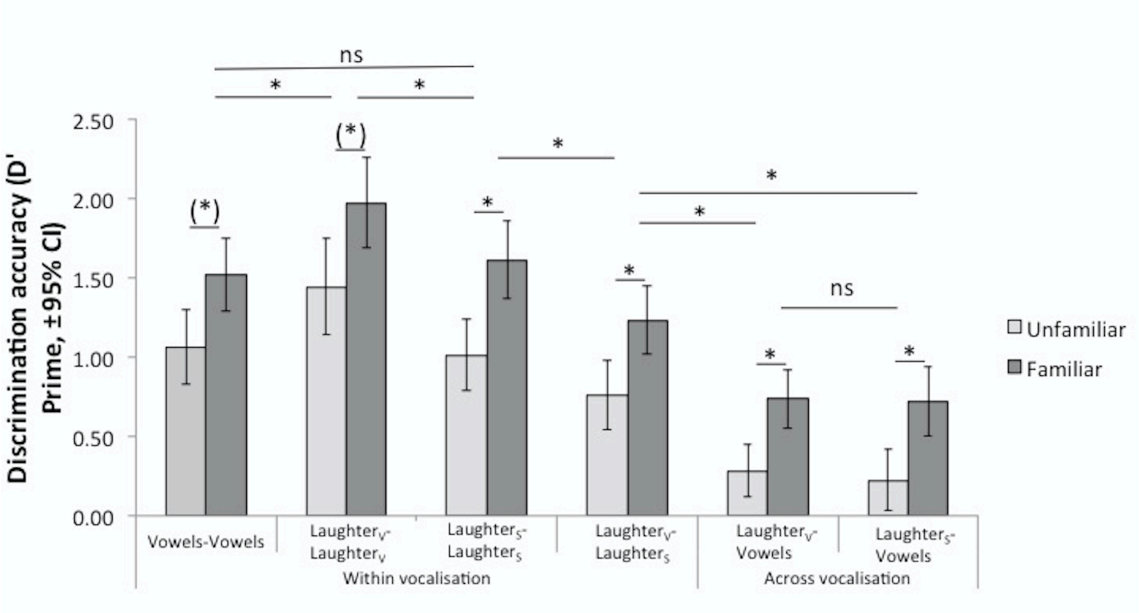

Figure 3 Results for Experiment 2. Average d' scores per condition for the speaker discrimination task. Significant comparisons (Bonferroni-corrected, see Results for alpha levels) are highlighted with an asterisk; marginally significant results are highlighted with an asterisk in brackets.