Flexible voices: Implications of variability in vocal signals for the perception of speaker characteristics from familiar and unfamiliar voices

Nadine Lavan

Thesis submitted for the degree of Doctor of Philosophy
Royal Holloway, University of London

Declaration of Authorship

I, Nadine Lavan, hereby declare that this thesis and the work presented in it is entirely my own. Where I have consulted the work of others, this is always clearly stated.

Signed:

Date:

Abstract

Listeners are able to extract a wealth of information from voices: linguistic content, psychological states and speaker characteristics, such as age and sex, can all be decoded from vocal signals. While human vocal communication is uniquely flexible and variable, studies looking at the extraction of speaker characteristics to date have, however, mainly used neutral speech samples. This thesis explores the perception of speaker characteristics from variable vocal signals outside of neutral speech: Non-verbal vocalisations, produced under different levels of volitional control (vowels, volitional laughter, spontaneous laughter and crying) and whispered speech were used to investigate familiar and unfamiliar listeners' abilities to extract and generalise information about speaker characteristics from such variable vocal signals.

Experiments 1-2 show that speaker sex perception is impaired for spontaneous vocal signals compared to volitional signals. Experiments 3-4 reveal that speaker identity discrimination is impaired for pairs of spontaneous vocalisations (spontaneous laughter and crying) compared to volitional vocalisations (volitional laughter and vowels), and performance decreases dramatically for pairs requiring generalisation across variable social signals (e.g. vowels versus spontaneous laughter). Experiment 5 shows that while familiarity with a voice can to some extent offset these effects, generalisation is still drastically impaired. Experiment 6 further suggests that familiar listeners are afforded a greater advantage over unfamiliar listeners when extracting identity-related information from voiced vocals signals, compared to whispered signals. Experiments 5-6 thus suggest that a familiarity advantage only generalises to a certain extent for relatively unfamiliar vocal signals (spontaneous laughter, whispered speech). Finally, an fMRI study (Experiment 7) explored the neural

underpinnings of the effects described above. This thesis thus shows that 1) the perception of speaker characteristics is affected in a differential manner for different vocalisations and that 2) generalisations of identity-related information across variable vocal signals is only possible to a limited extent – even in familiar listeners.

Content

Figures

Tables

Acknowledgements

The last three years that I spent doing this PhD were an incredible time and I am indebted to all of the people who helped me along the way in many many ways.

A huge thanks goes to Carolyn for being extremely patient, knowledgeable, generally amazing – thanks for putting up with me. More thanks go to Sophie Scott for starting me off in science a few years ago with the opportunity work in her lab and to Manos Tsakiris for being a great advisor.

None of the studies would not have been possible without the help of many a volunteer. Therefore thanks go to the laughers, speakers and whisperers, Alana, Becca, Carolyn, Dawn, Laura and Polly, who very kindly came up to my recording booth more than once and, despite the strange requests, produced wonderful vocalisations that resulted in some really interesting findings (and amused my participants). Thanks also to the undergraduate students who helped me collect data for some of the projects in this thesis.

I also have to thank all of my colleagues and friends, especially Cesar, Kyle, Saloni, Sinead and Sam, who were always ready to have a drink, a rant, a dance and/or some glorious renditions of ABBA or Adele songs to mark successes, failures and everything falling between these two. Many thanks also go to Dan, Lou, Mirjam and my housemates, past and present, who did an excellent job at being voices of reason whenever things felt like they were getting a little bit out of hand: You were right, it all worked out ok in the end!

1 Introduction

Human voices are uniquely variable and flexible: Aside from (conversational) speech, which is one of the most prominent and frequently produced human vocal signals, human vocal communication includes many other vocalisations, such as laughter, sighs, and filler sounds (e.g. "erm, uhm") that permeate everyday interactions and serve diverse social and communicative functions. Thus, humans routinely produce a wide range of vocalisations that differ vastly from each other in how they are produced, their acoustic properties, perceptual qualities and meaning.

While evidence for vocal flexibility has been found in some animals (Pisanksi, Cartei, McGettigan, Raine & Reby, 2016), humans are exceptional in their ability to change their voices volitionally, for example to convey particular social traits (such as masculinity/femininity or confidence; 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 the volitional use of the voice 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). Further, transient changes in the voice introduced by involuntary or spontaneous changes in a speaker's state have also been shown to drastically affect the vocal output. Authentic emotional experiences are often accompanied by emotional intonation patterns in speech or spontaneous vocalisations, whose production mechanisms differ dramatically from those employed to produce neutral speech (e.g. Ruch & Ekman, 2001, see Section 1.2). Due to physiological changes apparent in spontaneous vocalisations produced during authentic emotional

experiences, the production of vocal signals is affected at both the *source* (sound production by vibration of the vocal folds in the larynx) and the *filter* (shaping of the source sound by the articulators, including the lips, tongue, jaw, soft palate). Thus, humans produce highly variable and flexible vocal signals – with both volitional and spontaneous processes modulating the features of the vocal output.

A large body of work has shown that a wealth of information about a speaker, such as a person's age, sex, emotional state, state of health and identity are all encoded in vocal signals and can be extracted by listeners with some accuracy (Belin, Fecteau & Bédard, 2004; Kreiman & Sidtis, 2011; Lass, Hughes, Bowyer, Waters & Bourne, 1976; Linville, 1996; Mathias & von Kriegstein). Much of what we know about the extraction of speaker characteristics and identity-related information from voices, be that for explicit identification, recognition or discrimination of familiar and unfamiliar persons, has, however, been based on speech signals, produced under full volitional control and in a neutral voice (e.g. Winters, Levi & Pisoni, 2008 [words]; Schweinberger, Herholz & Sommer, 1997, Kreiman & Papcun, 1991 [extracts from discourse]; Van Lancker & Kreiman, 1987, Perrachione, Del Tufo & Gabrieli, 2011 [sentences]). However, such speech vocalisations, produced in a neutral voice, are only a limited subset of the vocal signals, that humans regularly produce in everyday settings and do not reflect the variability [1] and flexibility of human vocal communication: the extraction of speaker characteristics in the context of vocal flexibility, that is speaker perception based on a range of diverse vocal signals, has only received limited attention in the literature to date. The current thesis will

[1] Variability as a concept is here used to describe between vocalization type variability, that is differences between, for example, laughter and vowels and not within vocalization type variability.

therefore attempt to address these aspects of voice perception through a series of behavioural and neuroimaging experiments.

The following section will provide an overview of how differences in voice production affect vocal signals and encode information. Spontaneous and volitional vocal production will be contrasted in term of the neural and physiological underpinnings as well as their acoustic consequences. This will be followed by a detailed review of voice processing based on Belin et al.'s (2011) model of voice perception. The introduction will thus synthesise the literature on voice perception and production that is relevant to further investigate perception of speaker characteristics outside of neutral-speech signals.

1.1 Voice production: a dual pathway model

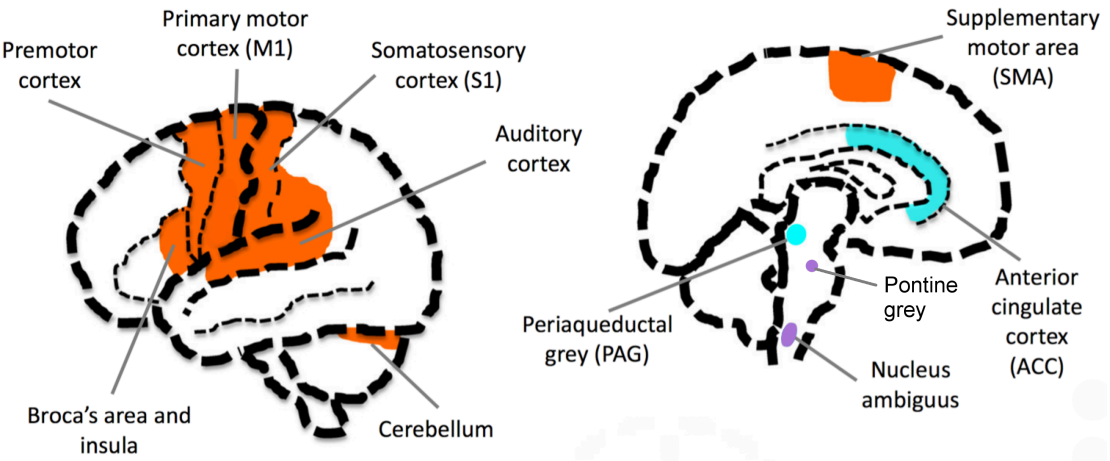

Figure 1 Illustration of the regions implicated in the production of volitional vocalisations (orange) and spontaneous vocalisations (turquoise). Purple indicates that the structure is thought to play a role in both pathways. On the left, the lateral surface of the brain is illustrated, on the right a midline sagittal slice is shown. Adapted from Pisanksi et al. (2016).

Variability in vocal signals can be introduced through volitional as well as spontaneous changes in voice production. There is evidence that the production of spontaneous and volitional vocalisations relies on at least partly distinct neural networks: For

example, some aphasic patients with disrupted volitional speech production are still able to produce spontaneous vocalisations such as laughter or emotionally-charged speech through swearing (e.g. Rohrer, Warren & Rossor, 2009; Cappa, Guidotti, Papagno & Vignolo, 1987; Van Lancker & Cummings 1999). A model of vocal production proposes two neural pathways that underlie the production of innate and learned vocalisations, respectively (see **Figure 1**; e.g. Ackermann, Hage & Ziegler, 2014; Pisanski et al. 2016; Owren, Rendall & Amoss, 2011): A pathway including primary motor cortex, ventrolateral and insular parts of the frontal lobes, connected to subcortical structures, such as the reticular formation, pontine grey and the phonatory motor neurons (located in, for example, the nucleus ambiguous) is involved in the production of learnt vocal behaviours, such as speech. A second pathway runs from the anterior cingulate cortex (aCC) via the periaqueductal grey (PAG) and adjacent ventral tegmentum to the reticular formation and pontine grey, finally to the phonatory motor neurons in the nucleus ambiguus. This pathway is thought to be involved in the production of innate and automatic behaviours, such as non-verbal emotional vocalisations (Jürgens, 2009; see Ackermann et al., 2014 for a review). The contrast between innate and learned vocal behaviours is closely linked to the notion of spontaneous and volitional vocalisations used in this thesis. The dual pathway model thus provides valuable insights how these two types of vocal behaviours may arise and differ from each other.

A clear separation between pathways?

The evidence from animal and human studies thus suggests a phylogenetically continuous neural pathway for spontaneous vocalisations produced under reduced

volitional control, with a second neural pathway in humans supporting the production of volitional and learned vocalisations, such as speech. Some research has recently challenged the independence of these two pathways, noting that they may interact under certain circumstances, such as during the production of emotional speech: while 'spontaneous' emotional prosody is present, speech is also 'volitionally' produced at the same time (Ackermann et al., 2014; McGettigan & Scott, 2014). This thus also implies that differences in volitional control are likely to reflect a continuum instead of a clearly defined binary (see McKeown et al., 2015 for a discussion). Nonetheless, given the presence of these basic differences in the neural underpinnings of volitional and spontaneous vocalisation production, one can expect differences in the physiological processes underpinning volitional and spontaneous voice production, the vocal output and as a consequence the perception of such variable signals. These differences will be discussed in the following sections.

1.2 How are vocal signals produced?

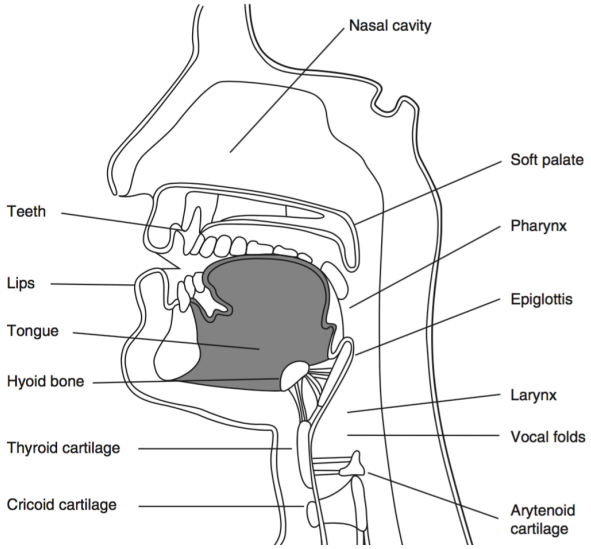

Figure 2 Anatomy of the vocal apparatus, adapted from Kreiman and Sidtis, 2011

In order to understand how information can be extracted from vocal signals, it is crucial to be aware of the physiological mechanisms underlying voice production that lead to the encoding of information in a vocal signal. **Figure 2** shows the basic anatomical layout of the human vocal tract. Voice production involves intricate interactions between breathing patterns, control of voicing at the vocal folds (source) and further modulations of the source signal through the shape of the vocal tract and movements of the articulators (filter; Fant, 1960). In order to phonate (i.e. produce voiced vocal signals) the air that is expelled from the lungs passes between the vocal folds, causing them to oscillate (see the Bernoulli effect or myoelastic-aerodynamic theory of speech; Titze, 1994). This results in a quasi-periodic buzzing sound, the source signal. Depending on the configuration of the vocal folds and the pressure with which air is passing through them, humans can regulate the fundamental frequency (the lowest frequency of a periodic signals, perceived as pitch, see **Figure 3**), intensity (perceived as loudness) and quality of their voice (e.g. falsetto, breathy, creaky or whispered). This source signal that is produced in the larynx is then further modulated by the shape of the supralaryngeal vocal tract, in the oral and nasal cavity: Depending on the type of vocal signal produced, the articulators assume different configurations and gestures during voice production.

Speech

The production of connected speech requires a stable subglottal pressure, which is exerted on the lungs to maintain a slow release of air from the lungs (Draper, Ladefoged & Whitteridge, 1959). This steady pulmonary airflow is essential for controlled and prolonged phonation as is best illustrated during attempts to speak

after intense exercise when breathing heavily: In this context, air is rapidly forced out of the lungs to supply the body with fresh oxygen. As a result of this type of heavy breathing, speaking is difficult, allowing the speaker to only produce only relatively brief vocal signals that are characterised by intakes of breath mid-utterance rather than the complex vocal signals lasting several seconds that are characteristic of human speech.

In order to create voiced and unvoiced realisations of speech sounds (e.g. creating phonemic contrasts between /p/ and /b/ to create, for example, the minimal pair 'pet' and 'bet'), vocal folds rapidly alternate between oscillating and letting air pass through them without oscillation. Articulator movements of, for example, the tongue, jaw and lips are precise and quick, shaping the airflow via articulatory gestures into a range of different speech sounds, such as vowels, fricatives, plosives and nasals (Scott, Sauter & McGettigan, 2010). These articulator movements or articulatory gestures, partially or fully obstruct the airflow leaving the lungs. For example, forcing the air through a narrow channel between tongue and another articulator, such as the hard palate, introduces turbulent airflow. This affects the spectral properties of the vocal output, introducing high-frequency aperiodic properties in the vocal signal to form consonants such as /s/ and /z/. Other articulator movements during, for example, vowel production, do not fully obstruct the air flow but change the shape of the supralaryngeal vocal tract which then introduces formants (i.e. bands of high spectral energy) into the vocal signal – these formants are crucial for making the different vowels in speech (Ladefoged & Disner, 2011; see **Figure 3**).

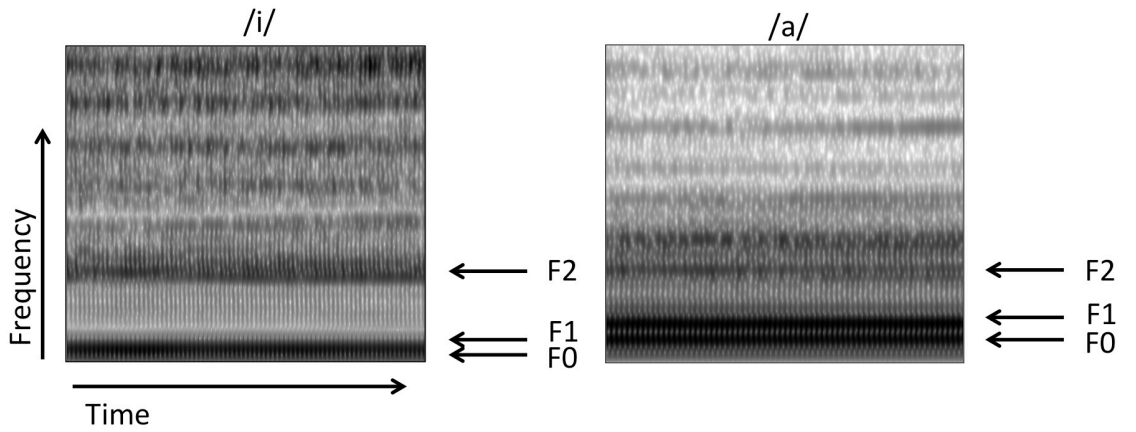

Figure 3 Spectrogram of the vowel /i/, showing the fundamental frequency (Fo) and first formants (F1 and F2) volitionally of each other. Darker shading on the spectrogram represents higher intensity.

This intricate control over respiration and the articulators required to produce speech is unique to humans. The groups of abdominal and thoracic muscles (e.g. intercostal muscles located between the ribs) that are used to exert respiratory control are used in quadrupedal primates during locomotion, breathing is coupled with movement and thus limits respiratory control. Bipedalism in humans has, however, freed up these groups of muscles to a large extent, increasing breath control and thus enabling the production of long speech utterances with one intake of breath (MacLarnon & Hewitt, 1999, see also Provine, 2016). Further, compared to other primates, humans have a descended larynx, a descended tongue root and more domed palate. These anatomical differences allow for a more diverse inventory of articulatory gestures, which subserve speech sounds (e.g. Fitch, 2010; see also the discussion of differences between humans and non-human primates' LMC in Section 1.1). Finally, (adult) humans are also able to control source and filter volitionally and independently of each other and vocalise predominantly without any associated affective or need state being present (Pisanski et al., 2016). Volitional voice and especially speech production is thus a uniquely human skill arising from evolutionary

changes in anatomy and vocal control that requires intricate interactions between respiratory, laryngeal and orofacial muscles.

Non-verbal emotional vocalisations

Non-verbal emotional vocalisations are thought to pre-date speech and are considered to be the most phylogenetically continuous means of expressing emotion in the voice (e.g. Davila-Ross, Owren & Zimmermann, 2010 for laughter). They have been considered to be more similar to animal calls than to speech due to their origins that seem to predate speech (Bryant & Aktipis, 2014; Scott, et al., 2010; MacLarnon & Hewitt, 1999). In contrast to speech, non-verbal emotional vocalisations emerge earlier in development (Scheiner, Hammerschmidt, Jürgens & Zwirner, 2006) and their production does not require any auditory experience (Eibl-Eibesfeldt, 1972; Scheiner et al., 2006). They arguably also constitute clearer examples of emotional expression in the voice than speech-based emotional vocal signals, indicated by higher emotion category recognition rates (e.g. ~50% for 14 emotions encoded in speech prosody [Banse & Scherer, 1996] and ~80% for 10 non-verbal vocalisations [Sauter, Eisner, Calder & Scott, 2010a]). This may be in part because they are not constrained by, for example, the production of speech sounds or other linguistic cues that are encoded in intonation contours (e.g. rising pitch at the end of an utterance to signal a question; Scott et al., 2010). Emotional displays affect the whole body as evolutionary accounts of the production of emotional displays see them as vestiges of formerly adaptive behaviours, such as initiating flight responses in the presence of a threatening stimulus, thus enhancing an individuals' chances of survival (Darwin, 1872; Izard, 1992; Ekman, 1992). Therefore, facial expressions frequently occur

simultaneously with the production of spontaneous vocalisations directly affecting the filter characteristics of the vocal output. For example, smiles modulate the acoustic characteristics of vocalisations, which can then be perceived by listeners (Aubergé & Cathiard, 2003).

Non-verbal emotional vocalisations tend to be characterised by few uniform segments (e.g. a scream or gasp, but also see laughter and high arousal crying that are characterised by multiple onsets, Lloyd, 1938; Sauter, Eisner, Calder & Scott, 2010b). The relatively uniform nature of segments within these vocalisations is indicative of the absence of the fast and precise movements of the articulators that are typical for speech. The relatively short duration and small number of individual segments in non-verbal emotional vocalisation further suggest a lack of control over respiration as emotional states modulate breathing through the automatic contraction of abdominal and thoracic muscles (e.g. Heim, Knapp, Vachon, Globus & Nemetz, 1968).

Spontaneous and volitional vocalisations: the case of laughter production

In humans, vocalisations that occur in the context of a genuine emotional experience may be produced under reduced volitional control. For example, the production mechanisms underlying spontaneous laughter – which differ drastically from the mechanisms underlying volitional speech production: Intense laughter produced involuntarily in response to an intense underlying emotional state is characterised by an initial forced exhalation (Ruch & Ekman, 2001), expelling most of the air in the lungs due to spasming of the diaphragm and the intercostal muscles that continues for the duration of the laugh. Breathing during intense laughter is characterised by the sharp inspirations between bouts of laughter when the "inspiratory muscles overcome

the expiratory muscles" (Lloyd, 1938, p. 188). While the abdominal and thoracic muscles are engaged during laughter, very few supralaryngeal modulations occur during non-verbal emotional vocalisations. Ruch and Ekman (2001) suggest that laughter produced in the presence of an underlying emotional experience is an inarticulate vocalisation, with articulators being mostly in their resting positions (note however, that, for example the lips tend to be spread due to smiling, and/or the jaw may be open). The more intense a laugh, the less control the laugher has over these physiological changes that affect phonation as outlined above. In humans, these production mechanisms have been proposed to introduce 'hard-to-fake' acoustic features to spontaneous vocalisations, marking them as reliable, authentic signals for receivers, which contrasts with volitional laughter being an unreliable signal, potentially evolved to deceive receivers (Bryant & Aktipis 2014; McKeown, Sneddon & Curran, 2015). Bryant and Aktipis (2014) have consequently proposed that due to an evolutionary arms race, humans have become experts both in producing maximally authentic-sounding volitional laughter by approximating the production mechanisms of spontaneous laughter but have also developed a fine-tuned perceptual system to discriminate between potentially well-matched volitional and spontaneous laughter – both beneficial skills for the individual.

1.3 Acoustic descriptions of volitional and spontaneous changes in vocal signals

The field of phonetics is dedicated to describing the acoustics of speech sounds and intonation contours to convey linguistic information (e.g. differences between /s/ and /z/ or rising intonation at the end of an utterance for a question). Another large body

of research has focussed on how non-verbal information, such as affective content, is encoded in the voice. The following section will focus on this literature to illustrate the drastic acoustic changes present in vocal signals outside of neutral voice production. Due to ethical issues during stimulus recordings (e.g. inducing negative states in listeners that would result in spontaneous vocalisations of for example, fear or anger) or quality of spontaneous recordings (e.g. background noise or verbal content within recordings made in naturalistic settings) only few studies to date have provided acoustic descriptions of spontaneous vocal signals. There is, however, a large literature discussing acoustic properties of emotional speech and non-verbal emotional vocalisations produced under volitional control. Since these vocalisations are (potentially stereotyped) close approximations of spontaneous vocalisations (see Bryant & Aktipis, 2014; McKeown et al. 2015), these descriptions may nonetheless serve as a useful heuristic to describe the acoustic changes in spontaneous emotional vocalisations.

Studies looking to obtain acoustic descriptions of vocal signals, usually use acoustic parameters, such as descriptions of fundamental frequency (e.g. Fo mean, Fo variation and Fo range), spectral measures (e.g. spectral centre of gravity, proportion of high frequency energy above a certain threshold), measures of periodicity (e.g. harmonics-to-noise-ratio, shimmer and jitter) and descriptors of the amplitude envelope (e.g. total duration, rate of articulations and intensity). It should be noted that one challenge to date has been to compare results across different studies: studies use distinct sets of acoustic parameters, extracted in study-specific ways. To make studies using acoustic parameters more comparable to each other, Eyben et al.

(2016) have proposed a minimalistic set of acoustic parameters meaningful for emotional voice analyses based on computational modelling.

Finding emotion-specific acoustic markers

Studies of emotional displays in vocal signals have attempted to define minimal sets or a small number of emotion-specific acoustic cues for emotional prosody, that is emotionally-inflected speech, and non-verbal emotional vocalisations. Depending on the specific study, a range of emotion categories have been explored, such as fear, anger, sadness and happiness, as well as occasionally different levels of emotional intensity signified by category labels such as 'hot anger' versus 'cold anger' and 'fear' versus "panic" (see Juslin & Laukka, 2003 overview of the emotions used). Materials in studies looking at emotional prosody range from vowels to pseudo-words and brief sentences of neutral linguistic content. Acoustic parameters are then extracted from vocal signals and used to describe the acoustic profile vocal signals in relation to each other While individual studies report distinct profiles for different emotions based on such acoustic analyses (e.g. Banse & Scherer, 1996; Sauter et al., 2010b), an exhaustive meta-analysis of studies of acoustic features of vocal emotions by Juslin and Laukka (2003) highlights that such claims may be problematic: Acoustic markers for distinct emotion categories largely overlap and can differ vastly for low compared with high intensity displays from the same emotional category, thus within-vocalisation differences may be as large a across vocalisation differences in cues. Juslin and Laukka (2003) argue these heterogeneous findings may partly be due to the lack of control and matching in the materials, for example, for arousal and other affective features.

Acoustic markers of continuous affective properties

Some studies have looked, more generally, for acoustic markers of affective features, based on Russell's (1980) circumplex model of emotion: This model assumes that all emotions can be mapped onto a multidimensional space, defined by a small number of orthogonal axes. Studies of affective features across emotion categories have thus investigated acoustic correlates of for example arousal and valence, which have been proposed as axes in Russell's (1980) original account. In this context, arousal describes a continuum from a person being very drowsy and sleepy to someone feeling highly alert, while valence describes a continuum between very pleasant and very unpleasant experiences. Such studies provide clear evidence for arousal-specific cues encoded in vocal signals: higher fundamental frequency (F0), higher intensity (in dB), a faster speech rate and an increase in harmonic energy have all been associated with high arousal emotions compared to low arousal and neutral vocalisations (Scherer, Johnstone & Klasmeyer, 2003; Sauter et al., 2010b). Intriguingly, acoustic analyses have been unable to distinguish between emotional displays of different valence (Scherer, 2003; Juslin & Laukka, 2003): Very different emotions, such as happiness and anger, show the same profile of acoustic features compared to neutral speech, i.e. higher F0, faster speech rate, higher intensity and more energy in the high frequencies among other factors. Studies using regression analyses to predict perceptual ratings of arousal and valence based on acoustic measures furthermore report that relatively little variance in valence ratings is accounted for by acoustic measures that are reliable predictors of arousal ratings: Laukka, Juslin & Breslin (2005) found that only 25% of the variance in valence ratings for emotional prosody could be explained by acoustic parameters in contrast to 74% of the variance of the arousal ratings (for similar

findings, see Bachorowski, 1999 and Bänziger & Scherer, 2005). Similarly, in Sauter et al.'s study (2010b) looking at the recognition of non-verbal emotional vocalisations, acoustic predictors accounted for only 17% of the variance in valence ratings as opposed to nearly 60% of the variance in the arousal ratings. Therefore, while listeners are reliably able to judge the valence of vocal signals (see Section 1.5.2), analyses nonetheless fail to describe acoustic cues underlying these judgements. This may indicate that the traditional acoustic measures typically applied to the analysis of neutral speech samples may not be sufficient for the investigation of emotional information encoded in the voice.

Alternative descriptions of the signals

Some attempts were made to find alternative descriptions of the acoustics of emotional speech and non-verbal vocalisations: Bänzinger and Scherer (2005) hypothesised that the intonation profiles of emotional prosody (as opposed to global, averaged Fo measures) may encode information about the valence of the emotion conveyed in an utterance. The authors analysed the intonation contour based on tones, measured steepness of slopes in the Fo contour, general Fo declination over the course of the utterance and counted the number of falls and rises during various segments of the utterance. They did not, however, find reliable evidence for emotion or valence specific features within these measures. Other studies have attempted to find alternative frameworks to describe the acoustic signal. For example, Gobl and Nì Chasaide (2003) created synthetic stimuli of varying types of voice quality and asked participants to judge the emotional content of the stimuli. The authors conclude that certain voice qualities seem to be associated with affective qualities, but similar to

traditional acoustic analyses, no distinct profiles of voice quality were found for individual emotion categories. In another small-scale study, we investigated whether ratings of phonatory and articulatory features could form a framework for better describing perceptual qualities in laughter (Lavan, Scott & McGettigan, 2016). Using trained listeners (phoneticians as well as speech and language therapists), we collected ratings of nasality, breathiness and mouth opening for a set of volitional and spontaneous laughs. Regression analyses within and across these two types of laughter then explored the relationship between such ratings and perceptual qualities, such as authenticity and arousal. The ratings accounted for a significant amount of variance for most regression models of arousal, valence and authenticity ratings – even though traditional acoustic measures accounted for a larger amount of variance for most regression models.

In sum, different approaches and measures of investigating the acoustic cues to different emotions in the voice have, at times, have shown that emotional content reliably affects vocal signals. While there is some evidence for emotion-specific acoustic profiles in the voice as assessed by traditional acoustic measures, other studies suggest that differences in continuous properties (such as arousal and valence) may explain these results, within and across different emotion categories. Even though attempts to describe emotion-specific cues in the voice can be to some extent problematic, acoustic analyses, voice quality judgements of phonatory and articulatory features of vocal signals confirm that emotional content – especially when modulating levels of arousal – has a significant impact on the acoustic features of vocal signals compared to neutral vocalisations.

1.4 Voice perception: A multi-step model

Listeners are able to extract a wealth of information from vocal signals. A model of voice processing, based on Bruce & Young's model of face processing (1986), has been proposed as a heuristic, offering a framework for understanding the different aspects that may underlie voice processing (Belin, Fecteau & Bédard, 2004; Belin, Bestelmeyer, Latinus & Watson, 2011; see Figure 4). The model proposes that voices are processed in a hierarchical manner, starting with the basic acoustic analysis of incoming sounds in subcortical nuclei and primary auditory cortex. This is followed by voice-signal specific analyses in the temporal lobes, in the so-called Temporal Voice Areas (TVAs; Pernet et al., 2015). Evidence from neuroimaging studies supports the notion of voice-selective areas in bilateral mid and anterior superior temporal gyri (STG) and superior temporal sulci (STS). These areas have been shown to respond more strongly to vocal signals compared to other non-vocal signals, such as environmental sounds (Pernet et al., 2015; see also Belin et al., 2011). It should be noted that a large part of this selectivity for voices can, however, be explained by the physical properties of the signals, as activations in TVAs largely disappear when the acoustic properties of the sounds have been accounted for. This suggests that within the TVAs, voice processing is still closely linked to specific acoustic features and activation of these regions may thus at least partially reflect the processing of complex or highly salient sounds, instead of (or in addition to) abstracted voice-specific processing (Leaver & Rauschecker, 2010).

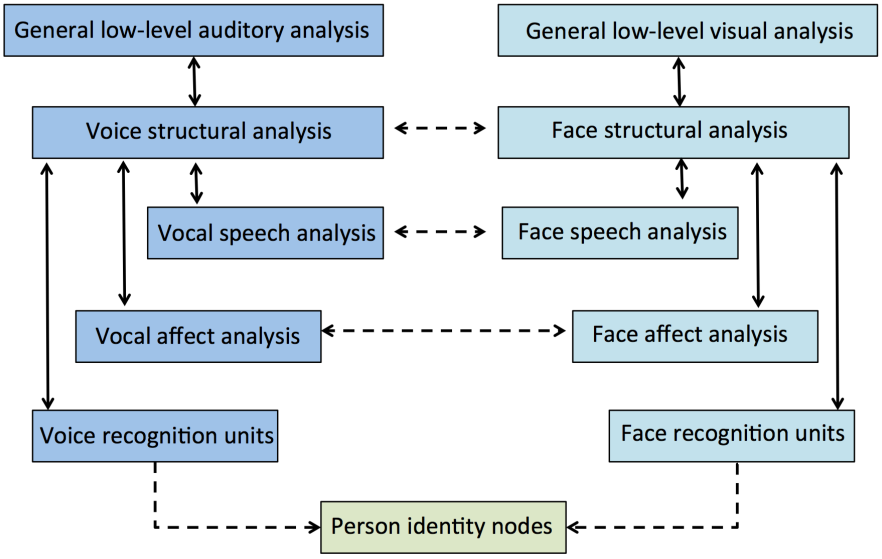

Figure 4 Hierarchical model of voice perception, adapted from Belin et al. (2011). Light blue boxes contain information about auditory processing, turquoise boxes refer to visual processes, light green boxes refer to amodal processing steps. Arrows denote interactions within (solid) and across (dotted) modalities.

Independence of voice processing pathways

The model outlined in Figure 4 proposes that within the TVAs, speech-, affect- and identity-related information is extracted from the voice and processed along three partially independent pathways. Neuropsychological evidence confirms this partial dissociation: Studies of individuals with phonagnosia, that is individuals who are unable to recognise others from their voices, show a selective disruption of the processing of identity-related information (e.g. recognising a speaker or discriminating voices from each other) while emotion and speech processing from verbal and non-verbal emotional vocal signals remain intact (Garrido et al., 2009; Hailstone, Crutch, Vestergaard, Patterson & Warren, 2010). Van Lancker and Canter (1982) further report that 20 out of 21 left hemisphere stroke patients, all of them aphasic (according to the authors: "seven Broca's aphasics, three mixed anteriors, two Wernicke's, one amnesic, and five global aphasics. Three patients had mixed

symptomatologies and were not clinically classifiable" [Van Lancker & Canter, 1982, p. 189]), could still process identity information encoded in voices. This study thus provides evidence for a dissociation between speech (linguistic content) and identity processing. With regard to the independent processing of emotion and speech from the voice, Barrett, Crucian, Raymer and Heilman (1999) report the case of a patient with global aphasia after a left hemisphere stroke. While the patient's speech comprehension was disrupted, her ability to match emotional prosody to facial expressions (and vice versa) in a nonverbal task remained intact. Heilman, Scholes and Watson (1975) further report the results of a study of six patients showing normal sentence comprehension but impaired on emotion recognition from emotionally inflected sentences after right hemisphere strokes. By finding evidence for selective impairments of one pathway versus the other, these patient studies thus demonstrate dissociations for all three proposed processing pathways, providing evidence for at least partial independence.

Interactions between voice (and face) processing pathways

There is also some evidence that the three pathways for emotion-, speech- and identity-related information interact with each other in healthy listeners: Studies of speech intelligibility, for example, show that listeners are better at understanding speech from familiar talkers compared to unfamiliar talkers (Goggin, Thompson, Strube, Simental, 1991; Nygaard & Pisoni, 1998; Pisoni, 1993). Speaker recognition has also been shown to be language dependent: Listeners are better at recognising and learning vocal identities when exposed to speech samples produced in their native language (e.g. Perrachione, Pierrehumbert & Wong, 2009; Perrachione et al.,

2011) – this is the case even when the listeners have only been passively exposed to the language without understanding it (Orena, Theodore & Polka, 2015).

Belin et al.'s (2004, 2011) model furthermore proposes that voice and face processing pathways interact at each of the different stages during the processing of multimodal information. There is a wealth of evidence showing that signals from faces interact with information encoded in a voice. The McGurk effect (McGurk & MacDonald, 1976), where incongruent visual and auditory information (e.g. (seeing a person saying /ga/ while simultaneously hearing them saying /ba/) for a speech sound interact resulting in an auditory illusion (perceiving the audiovisual stimulus as /da/), shows such interactions for speech comprehension. Emotional information in one modality has been shown to influence the perception of emotion in other modalities: During the presentation of incongruent audiovisual pairings of emotional displays, the perception of the emotional content of one modality has been shown to be shifted towards the emotional content of the other modality (Barrett & Kensinger, 2010; Collignon et al., 2008; De Gelder & Vroomen, 2000; Lavan, Lima, Harvey & McGettigan, 2014). The posterior STS has been suggested as a hub for the integration of interactions between vocal and visual information (see Brück, Kreifelts & Wildgruber, 2011 for a review of emotional displays; MacSweeney et al., 2001 for speech reading). As a final stage of the model, information from the identity pathway in both faces and voices (plus any other identity related information available) finally culminates in an amodal person identification node (PIN).

1.5 Voice perception: what's in a voice?

According to Belin et al.'s (2004, 2011) model, listeners can extract information about speech, emotional states and person-related characteristics from (speech-based) vocal signals. The following sections will review the evidence supporting these claims.

1.5.1 Speech

A large body of work has shown that despite considerable variability within speech signals, listeners are experts at extracting linguistic information from vocal signals – even in challenging listening situations. Theoretical models have described the neural underpinnings of speech perception (e.g. Hickok & Poeppel, 2000; Scott & Johnsrude, 2003) and have proposed models for the underlying cognitive mechanisms (e.g. McClelland & Elman, 1986). Since most of the current thesis is not concerned with the processing of linguistic information in speech signals, this literature will not be reviewed in detail here.

1.5.2 Emotion

Much of the research on emotion perception from voices has focussed on whether participants can accurately categorise emotions from vocal displays. Similar to emotional expressions from other modalities, such as facial expressions (e.g. Ekman & Friesen, 1971), there is conclusive evidence that emotional prosody and non-verbal emotional vocalisations can be recognised and reliably categorised both across and within cultures (Bryant & Barrett, 2008; Scherer, Banse & Wallbott, 2001; Pell, Monetta, Paulmann & Kotz, 2009; Paulmann & Uskul, 2014; Sauter & Scott, 2007; Sauter et al., 2010a). All studies report above-average categorisation accuracy for

anger, fear, sadness, disgust and amusement. It should be noted that in-group advantages are regularly found in cross-cultural studies, where recognition rates are generally higher in cases where sounds are presented to listeners from the same culture as the speakers – indicating some cultural specificity in emotion expressions (Paulmann & Uskul, 2014; Sauter et al., 2010a). Aside from the basic emotion categories (anger, fear, sadness, surprise, disgust and joy) initially proposed by Ekman (1992), it has also been shown that other emotions can as well be reliably recognised by listeners within-culture: participants were for example shown to be able to distinguish between vocalisations denoting different positive emotional states such as relief, achievement and pleasure (Sauter & Scott, 2007).

Beyond accurate categorisation of vocal emotions, studies have also found that participants are able to reliably judge other affective qualities within and across different vocal emotions. Based on Russell's circumplex model of emotion (1980, see Section 1.3 for a description), it has been shown that participants can give ratings of arousal and valence among other dimensional affective judgements for vocal emotions with high interrater reliability (Juslin & Laukka, 2003, Laukka, Juslin & Bresin, 2005). Crucially, listeners can furthermore accurately assess other nuanced aspects of affective information in the voice, for example, they can detect whether vocalisations were produced volitionally or spontaneously (Bryant & Aktipis, 2014; Lavan & McGettigan, 2016).

Laughter: The perception of meaningful distinctions within a vocalisation category

Emotional vocalisations are not unitary in their meaning: one type of vocalisation, such as laughter, can signal a range of meanings, depending on the context in which it

is produced (or perceived). Recent research on laughter has shown that participants are able to make accurate within-vocalisation judgements of affective features of certain vocal signals: Participants are able to reliably judge the authenticity of a laugh, i.e. whether the laugh was produced in response to genuine amusement or whether it was produced without an underlying emotional state (Bryant & Aktipis, 2014; McGettigan, Walsh, Jessop, Agnew, Sauter, Warren & Scott, 2015; Lavan et al., 2016). Szameitat and colleagues (2009a, 2009b) presented participants with recordings of ticklish laughter, taunting laughter, joyful laughter as well as laughter denoting *schadenfreude* (a laugh signalling both joy and taunting). The authors found that participants could not only distinguish between the different laughs and categorise them accurately but that they also perceived these laughs to be significantly different in valence, arousal and dominance; the affective ratings could in turn be predicted by acoustic measures (Szameitat et al., 2009b). The laughter categories used in Szameitat et al.'s (2009a, 2009b) studies are, however, to some extent problematic: the authors used professional actors to produce the stimuli, asking them to portray these four specific types of laughter and assuming ecological validity of all four types – the actors thus created maximally discriminable, potentially stereotypical laughter sounds that listeners were then asked to classify, which may have inflated recognition performance. In another study of laughter, Bachorowski and Owren (2001) showed that authentic voiced laughter was rated as more positive, friendlier and more attractive compared to authentic unvoiced laughter (i.e. grunt-like or snort-like laughter). Further, Bryant et al. (2016) report the results of a large-scale cross-cultural study, which indicated that listeners from diverse cultures can reliably judge whether brief samples of laughter were produced by two friends or by two strangers. These

studies thus show that a wealth of nuanced affective features are encoded within a single vocalisation type (here: laughter) and can be reliably decoded by listeners.

1.5.3 Identity (and other speaker characteristics)

Listeners are able to extract a wealth of information about a speaker from vocal signals: information about physical characteristics of a speaker (age, height, weight, state of health, sex, arousal, etc.), psychological characteristics (arousal, emotional state, stress, etc.) and social characteristics (education, regional origin, social status, sexual orientation, occupation, etc.; see Kreiman & Sidtis, 2011 for an overview) can all be readily perceived by listeners. While some of these speaker characteristics can be established from voices with relatively high accuracy (e.g. speaker sex, estimates of speaker age and emotional state), other characteristics are less well circumscribed and mainly socially but not biologically marked, making listener judgements less reliable (e.g. occupation, regional origin and sexual orientation). In addition to these relatively objective speaker characteristics, many judgements pertaining to a speaker's personality as well as subjective assessments of specific qualities of a speaker's voice can be extracted from vocal signals: Studies have shown that listeners are able to make judgements about personality features, such as attractiveness, trustworthiness, dominance, aggressiveness and likeability of a speaker among others with high interrater reliability (Cronbach's alpha > .88; McAleer, Todorov & Belin, 2014). While listeners seem to agree on judgements such as these, it still remains to be established whether these judgements accurately reflect a speaker's personality traits or whether such judgements are associated with certain acoustic features that are, however, not necessarily linked to the speaker's personality.

Aside from judgements about specific speaker characteristics, such as age and sex, listeners are also able to extract holistic information about a person's identity from voices only: listeners can recognise and identify a familiar person from their voice only and they can discriminate between different (unfamiliar) speakers. Speaker identification (most frequently) requires listeners to explicitly name the speaker after being presented with the relevant vocal signals. Speaker recognition is tested by using forced choice paradigms including a range of speaker identities whereas speaker discrimination tasks are based on same-different judgements of pairs of vocal signals. Reliable speaker identification and recognition require prolonged prior exposure to a speaker's voice – with duration and type of exposure (incidental versus explicit) affecting performance: Earwitness studies, for example, report chance performance for speaker recognition after a brief incidental 15 second exposure to a voice, i.e. listeners talked to a speaker prior to the experiment without knowing that they will be asked about the speaker's voice later (Yarmey, Yarmey & Yarmey, 1994). However, after listening to a speaker for ~1:30 minutes with the explicit instruction to remember the voices, Papcun, Kreiman and Davies (1989) report clear above chance performance for speaker recognition. Abberton and Fourcin (1978) further report very high performance of speaker identification for voices of classmates that had known each other for 5 months (98%), while other studies show evidence for good recognition and identification accuracy of celebrities and other famous voices (e.g. Hanley, Smith & Hadfield, 1998; Schweinberger, Herzholz & Sommer, 1997).

Successful speaker recognition – especially for speakers that are not highly familiar to listeners – has furthermore been shown to depend on the duration of the test stimuli, the information encoded in the stimuli as well as the retention interval

between exposure and test time: Papcun, Kreiman and Davis (1989) report a significant drop off in speaker recognition rates after 4 weeks. In two well-controlled studies, Bricker and Pruzansky (1966) and Schweinberger et al. (1997) show that the longer the test stimuli and also the more (linguistic) information is encoded in a stimulus, the higher the recognition or identification rates. The authors link these findings to listeners being able to more thoroughly sample a speaker's vocal and phonemic inventory. Yarmey and Matthys (1992) and Kerstholt, Jansen, Van Amelsvoort and Broeders (2004) similarly find complex interactions between listener performance, stimulus duration and retention intervals.

Other studies have investigated speaker discrimination abilities of (unfamiliar) listeners. Van Lancker and Kreiman (1987) note that speaker recognition and discrimination are separate abilities with discrimination not necessarily preceding recognition. The authors support their argument with patient data showing a double dissociation between speaker discrimination and recognition. Studies of speaker discrimination in general show that, for healthy listeners in good listening conditions, accuracy of (unfamiliar) speaker discrimination is very high (> 90% for healthy young adults; Reich & Duke, 1979; Van Lancker & Kreiman, 1987; Wester, 2012).

It has thus been shown that listeners can extract a wide range of speaker characteristics, ranging from specific features, such as speaker sex to holistic judgements about speaker identity as well as (subjective) impressions of a speaker's personality. Performance for the tasks varies (see high performance for speaker sex compared to at times low performance for speaker recognition) as a function of task, listener, stimulus and speaker characteristics.

1.5.3.1 Mechanisms underlying successful identity perception in the voice

In the literature on the perception of speaker characteristics from vocal signals, some authors have proposed candidate processes or mechanisms to help listeners to extract information from vocal signals in a reliable way. Prototype-based processing has been suggested by some authors to underlie voice processing. In addition, a framework based on auditory expertise and auditory perceptual learning aligns with the effects reported in the literature to date. These processes will be discussed below.

Prototype-based processing

Kreiman and Sidtis (2011, 2012) propose a model of voice identity processing that is centred on the differential processing of familiar and unfamiliar voices. The authors offer a basic framework with regard to the underlying mechanisms involved in the higher order aspects of voice processing. They propose that familiar and unfamiliar voices are processed in relation to voice prototypes. The notion of prototype-based processing of signals is a mechanism commonly proposed by models in psychology and linguistics to explain the categorisation and recognition of abstract information (e.g. Homa, Sterling & Trepel, 1981; Osherson & Smith, 1981; Posner & Keele, 1968). While many versions of prototype theories exist, they all revolve around the notion that prototypical concepts of objects are stored in long-term memory. When encountering an object, imagined or real, this object is compared to the most likely or suitable prototypes available in long-term memory. Objects within a certain category can be more or less representative of these prototypical concepts. For example, while a sparrow might be highly representative of the category 'bird' as it has wings, feathers, a beak, lays eggs and can fly, a penguin is less representative of this category

as it cannot fly. Prototypes are thought to be shaped by an individual's experience – what is highly familiar and encountered frequently will be highly representative of an existing and potentially very specific prototype.

For voices, Kreiman and Sidtis (2011) propose that unfamiliar voices are processed based on their acoustic features in a stimulus-driven way and then compared to prototypical templates based on population averages (see also Kreiman & Papcun, 1991 and Papcun, Kreiman & Davis, 1989). In contrast to this, familiar voices are thought to be processed in a more holistic way, based on salient idiosyncratic features in a vocal signal. After detecting the idiosyncratic features of a familiar voice, vocal signals produced by familiar speakers are thought to be matched to representations of the specific speaker's vocal inventory that are stored in long-term memory.

A small number of studies report evidence supporting this prototype-based model of voice processing: Papcun, Kreiman and Davis (1989, for a similar study, see also Kreiman & Papcun, 1991) explored the long-term memory for three unfamiliar voices that were rated (by independent listeners) as being easy, medium or hard to remember – the interpretation here was that easy-to-remember voices are very distinct and dissimilar to average voices, while hard-to-remember voices are very similar to the average. In a between-subjects design, three groups of listeners were each initially exposed to one of these three target voices and asked to remember it. In a follow-up testing session (after 1, 2 or 4 weeks), listeners were then asked to identify their target voice from within a set of 10 voices (containing their target voice, plus 9 probes including the 2 other targets and 6 additional distractors). The authors report that correct identification rates for the target voice were similar for listeners in the

hard-, medium- and easy-to-remember groups. Listeners did, however, make significantly more false identifications (i.e. false alarms) in the easy-to-remember group than the hard-to-remember group. A further analysis looked at false identifications on the same 3 voices when listeners encountered them as *probes* (i.e. for listener who did not hear the voices as targets): this analysis showed that fewer false identification were made for the easy-to-remember probe while more false identifications were made for the hard-to-remember probe. Papcun et al. (1989) thus note on the one hand that the easy-to-remember target voice diverges from population average-voices and therefore has a relatively unstable representation in long-term memory. Over time, the memory of such distinct or atypical voices will degrade, thus resulting in frequent false identifications of other voices as the target voice. In contrast to this, when an easy-to-remember voice (i.e. atypical) voice is encountered as a probe and the listener has to compare it to their remembered target voice, all acoustic features of this distinctive probe are immediately available to the listener. This allows listeners to then dismiss this probe voice as a potential target, thus making fewer false identifications. On the other hand, the authors argue that hard-to-remember voices have relatively stable mental representations, because they are more similar to the already established population-average prototype and thus easier to more robustly encode on first presentation. The representation in memory of such relatively average voices does therefore not degrade as rapidly over time. This then results in fewer false identifications when hard-to-remember voices are targets (and the listener has to recall the features), but will also be associated with a higher number of false identifications when these voices are used as probes as their physical

features are not as distinct from population-average prototypes compared to easy to remember voices.

In another study looking at unfamiliar voice processing, Latinus and Belin (2011) report the results of an adaptation paradigm exploring perceptual after-effects from exposure to 'anti-voices'. This work was modelled on studies exploring person recognition from faces, which showed that adaptation to an "anti-face" results in facilitation of identification of the original (e.g. Leopold, O'Toole, Vetter & Blanz, 2001). Anti-faces and anti-voices are described as a caricature of the average identity, created by morphing an individuals' face/voice toward, and then beyond, the average stimulus. Intriguingly, anti-faces (and anti-voices) are perceived as categorically distinct from the original person identity. Latinus and Belin (2011) found perceptual after-effects on the identification of voice morphs (between an original speaker and the population average) following adaptation with the matching anti-voice. That is, after several seconds of exposure to Anti-voice C, listeners were relatively more accurate (in a 3-way forced choice task) at identifying a 30% morph between Speaker C and the prototype as Speaker C (and not A and B). However, this effect did not hold for adaptation with non-matching anti-voices (i.e. exposure to Anti-voice A followed by test on the Speaker C morph. Similar to Leopold et al. (2001), Latinus and Belin (2011) propose that the exposure to voices on the same 'identity trajectory' - that is, passing through a prototypical, average voice on the perceptual 'voice space' - can cause aftereffects by adapting the 'anti-features' of the talker and thus making the original features more prominent in perception. The authors take this as evidence showing that prototypical averaged voices play a central role in voice processing.

In an fMRI study, Latinus, McAleer, Bestelmeyer and Belin (2013) further report that activation in TVAs is modulated as a function of acoustic distance of a voice to an average prototypical voice, that is the less prototypical a voice, the stronger the activation in the TVAs. Conceptually, this is similar to the easy- and hard-to-remember voices found in Papcun et al.'s (1989) study of voice prototypicality, albeit that Latinus and colleagues defined this using acoustic properties as opposed to listeners ratings. Based on their fMRI results, Latinus et al. (2013) argue that voice processing occurs in bilateral TVAs with reference to prototypical (or averaged) voice representations. While these compelling findings could be regarded as evidence for prototype-based processing of voices, this could also reflect an expertise effect – TVAs may respond more strongly to non-prototypical or unusual voices that occur less frequently in perceptual experience, which by definition will lie further away from population averages in their acoustic properties – this calls into question whether it is the acoustic properties of average voices that play a prominent role during voice perception, or merely the listener's experience with those properties. Further, it is also unclear how increasing BOLD (blood oxygen level dependent) responses in a cortical area should be interpreted in this case – do these reflect greater novelty, or the presence of more elaborate neural computations? Nonetheless, these empirical studies do provide some evidence for a role of prototypical or averaged voices during unfamiliar speaker identity processing, further supported by the alignment with findings from studies of face perception (Leopold et al., 2001; Leopold, Rhodes, Müller & Jeffery, 2005; Rhodes & Jeffery, 2006). While there is some evidence from studies regarding the processing of learned voices (Papcun et al., 1989), prototype models in the context of familiar voices have not been explored in detail.

Auditory expertise, familiarity and perceptual learning

It has been proposed that humans are experts at face and voice processing based on the vast amount of information that can be extracted from these signals (Belin et al., 2004; 2011) – in their discussion of familiar voice processing, Sidtis and Kreiman (2012) write that human voice perception is a "prodigious cognitive ability" (p.147), and that the capacity to recognize multiple vocal identities has no known limit. Through repeated exposure and engagement, humans can become experts in processing certain stimuli. In the auditory domain in general, most studies of auditory expertise have focused on musicians: through prolonged exposure and engagement with sounds, many studies have shown advantages for musicians in auditory tasks, such as pitch discrimination, when directly compared to non-musicians (Spiegel & Watson, 1981; Kishon-Rabin, Amir, Vexler & Zaltz, 2001). Further evidence for perceptual learning comes, for example, from studies showing that listeners can improve their ability to understand drastically degraded noise-vocoded speech (with only minimal spectral information in the auditory signal; Shannon, Zeng, Kamath, Wygonski & Ekelid, 1995) after some training. Other studies have shown that listeners are able to accurately remember the pitch of a television series theme tune they have been watching, while listeners who are not familiar with the theme tune fail to identify the original pitch (Schellenberg & Trehub, 2003). With regard to auditory expertise in identity perception, only very few studies are available. One study has shown that musicians are better than non-musicians at identifying individuals from their musical performances (e.g. Koren & Gingras, 2014 for harpsichord performances; but see Gingras, Lagrandeur-Ponce, Giordano & McAdams, 2011 who do not report a listener effect). Chartrand, Peretz and Belin (2008) note that, in the visual domain, studies

have shown that dog, bird and car experts and chick sexers can discriminate between, recognise and identify individual exemplars within their expertise while lay people struggle to do so. Similar evidence exists for people being able to identify specific types of cars or train by the sound of their engines, while pet owners can identify their pet by its name – although all of this evidence appears to be purely anecdotal rather than empirical.

These few examples thus provide evidence for auditory perceptual learning and expertise, which has an impact on how stimuli are perceived. In the framework of a prototype-based account of voice processing (Kreiman & Sidtis, 2011), this thus suggests that listeners have formed robust and highly specific prototypes through perceptual learning of the sounds that fall within their area of expertise. This will then allow them to make more fine-grained distinctions between these sounds. Listeners who have no particular expertise with these sounds, lack well-defined representations of the category of stimulus in question, resulting in less fine-grained and thus less accurate judgements for relatively similar and non-distinctive sounds.

1.6 The current thesis

The previous sections of this chapter have outlined the unique flexibility and variability of human vocal signals, described how these properties emerge due to volitional and spontaneous changes in voice control and production, and considered how listeners extract linguistic, emotional and speaker-related information from the voices they hear. The flexibility in vocal signals has to date been largely neglected by research investigating how speaker characteristics are perceived from voices, since

these studies have largely investigated voice perception in the context of speech sounds produced in a neutral tone of voice. This thesis will thus explore voice perception outside of neutral speech signals in familiar and unfamiliar listeners. Natural, unmanipulated non-verbal vocalisations, produced under different level of volitional control (vowels, volitional laughter, spontaneous laughter and spontaneous crying) as well as whispered speech will be used to investigate familiar and unfamiliar listeners' ability to extract and generalise information related to speaker identity through behavioural testing, functional magnetic resonance imaging and acoustic analyses.

2 Speaker sex recognition from volitional and spontaneous non-verbal vocalisations

Experiments 1 and 2 explore how a basic judgement of speaker characteristics, speaker sex recognition, is affected by vocal flexibility, introduced by contrasting volitional and spontaneous vocalisations. In Experiment 1, participants judged speaker sex from two spontaneous vocalisations, laughter and crying, and volitionally produced vowels. Listeners' performance was significantly impaired for spontaneous vocalisations compared to volitional ones, a pattern that was also reflected in longer reaction times for spontaneous vocalisations. Within spontaneous vocalisations, performance for laughter was additionally impaired compared to crying. In Experiment 2, different stimuli were used: spontaneous laughter, volitional laughter and (volitional) vowels. Results indicate that performance was impaired for spontaneous laughter but not for volitional laughter and vowels. Experiment 2 therefore provides further evidence that differences in volitional control over production but not, for example, differences in arousal or vocalisation-specific effects drive these effects. For both experiments, acoustic analyses did not show clear relationships between stimulus properties and participant's performance. The results are discussed in the light of modulations of the salience of acoustic cues across variable vocal signals as well as potential modulations of attention during the perception of emotionally salient vocal signals.

2.1 Experiment 1

2.1.1 Introduction

Listeners are able to reliably determine speaker sex from audio-only vocal signals with high accuracy (Coleman, 1971; Lass et al., 1976). Speaker sex can be assessed rapidly, with listeners being able to identify sex from vowel segments lasting under 15ms (< 10 glottal cycles; Owren, Berkowitz & Bachorowski, 2007). Speaker sex can furthermore be successfully extracted even from degraded or manipulated vocal signals, such as sine-wave speech (which retains only the amplitude-modulated formants from the original speech signal) and noise-vocoded speech (which retains the amplitude-modulated temporal information for a limited number of frequency bands; Shannon et al., 1995) with as few as 3 channels (Gonzalez & Oliver, 2005). While speaker sex is traditionally conceptualised as a binary with two distinct categories, studies have found that when presented with stimuli morphed from male to female voices, listeners perceive speaker sex in a continuous (as opposed to a categorical) manner (Mullennix et al., 1995) – in contrast to this, phonemic contrasts show clear category-based perception (Liberman, Safford-Harris, Hoffman & Griffith, 1975). This difference between speech and speaker sex processing may thus indicate that the extraction of speaker characteristics may involve different mechanisms to speech perception.

The perceptual cues assumed to allow listeners to distinguish male from female voices are linked to sex-specific anatomical features of the vocal tract: Due to the pronounced sexual dimorphism of the human larynx and vocal folds, males tend to on average have longer and thicker vocal folds as well as longer vocal tracts than females (Titze, 1989). These two features mainly lower Fo and affect the spacing of frequencies of formants in vocal signals, thus making male and female voices

relatively distinct from each other. This sexual dimorphism is greater in humans compared to other apes and has frequently been linked to sexual selection, with low Fo predicting perception of (male) attractiveness and dominance (Puts et al., 2016).

Studies have indeed shown that acoustic cues, mainly differences in Fo and formant characteristics, are crucial for determining speaker sex from vocal signals that have been produced in a neutral voice (Bachorowski & Owren, 1999; Bachorowski, et al., 2001; Skuk & Schweinberger, 2014). The salience of these cues for speaker sex identification is furthermore highlighted in a study by Mullennix and colleagues (1995), who synthetically shifted Fo and formant frequencies in vocalisations and were thus able to successfully create continua of vocalisations that were perceived by listeners to morph from male to female. While both formant frequencies and Fo – and potentially other less explored acoustic factors – play an important role in determining speaker sex, it has been argued that Fo may be the more salient cue for speaker sex judgements: Lass et al. (1976) have shown that removing the source signal (which encodes Fo information) by using whispered speech affects participants' judgements of speaker sex more drastically than when stimuli are low-pass filtered (thus removing all filter information, which includes all formants [apart from Fo]). Honorof and Whalen (2010) report that when Fo is volitionally manipulated by a speaker within their natural Fo range while producing isolated vowels, misidentifications of speaker sex occur at the extremes of the Fo range, with high Fo vocalisations being identified as female and low Fo vocalisations as male. These studies therefore show that changes in salient acoustic cues through natural volitional voice modulations as well as synthetic manipulations of the stimuli can affect the accuracy of speaker sex judgements from voices, underlining the perceptual salience of these cues.

Given these findings, it can be expected that speaker sex identification should be impaired for other examples of natural vocal flexibility: spontaneous vocalisations differ from neutral vocal signals in their acoustic (notably in Fo which has been identified as a important cue to speaker sex judgements) and perceptual properties, and even from volitionally produced exemplars of the same vocalisation (for laughter, see Bryant & Aktipis, 2014; Lavan et al., 2016). The current study explored sex identification from variable vocal signals: Participants performed a speaker sex identification task on Spontaneous Laughter ($Laughter_S$), Spontaneous Crying ($Crying_S$) and Vowels ('staccato vowels'; see **Figure 5** for example waveforms and spectrograms). It was hypothesised that the perception of speaker sex would be impaired for spontaneous vocalisations, with listeners' performance for $Laughter_S$ and $Crying_S$ being significantly lower than for Vowels. Performance for $Laughter_S$ and $Crying_S$ should be similar as both vocalisations a similar across a range of acoustic measures (see Materials).

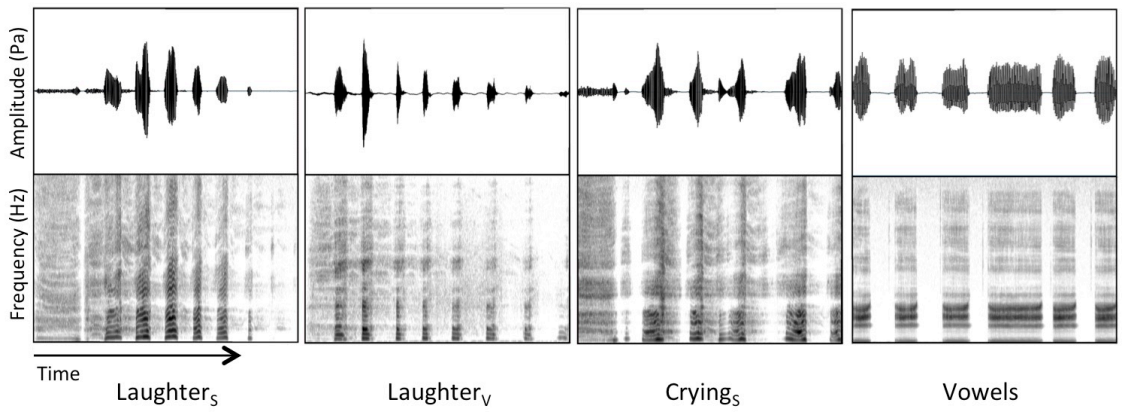

Figure 5 Waveforms (top panels) and spectrograms (bottom panels) of the vocalisation types used in Experiment 1-5 and 7: Spontaneous Laughter ($Laughter_S$), Volitional Laughter ($Laughter_V$), Spontaneous Crying ($Crying_S$) and Vowels ('staccato vowels'). Darker shading on the spectrogram represents higher intensity.

2.1.2 Participants

44 participants (24 female; M_{Age}: 20.9 years; *S*D: 1.2 years; range 19-24 years) were recruited at the Department of Psychology at Royal Holloway, University of London and received course credit for their participation. Testing for this study (and Experiment 3) was conducted by undegraduate students as part of their final year project. Each student was asked to test 15 participants, which resulted in the final sample size of 44. All participants had 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. Average performance across conditions for each participants was within 2 standard deviations from the mean and therefore all participants were included in the following analyses.

2.1.3 Materials

Laughter$_S$, Crying$_S$ and Vowels were recorded from 5 speakers (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 22.050 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 while watching or listening to amusing sound or video clips (see McGettigan et al. [2015] for a detailed description of

the recording procedure). For $Crying_S$, speakers recalled upsetting events and/or initially posed crying to encourage a transition into spontaneous crying associated with genuine felt sadness. Crucially, speakers informally reported genuine feelings of amusement and sadness during and after these recording sessions. No formal measurements of the speakers' emotional states were collected as obtaining such measures would have required to interrupt the recording session, which would have been detrimental to the elicitation of spontaneous vocalisations.

In a pilot study, a group of listeners *(N* = 13) provided ratings of arousal ("*How aroused is the person producing the vocalisation?*", 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 vocalisation feeling?*", with 1 denoting "*very negative*" and 7 denoting "*very positive*"), control over the vocalisations ("*How much control did the person have over the production of the vocalisation?*", with 1 denoting "*none at all*" and 7 denoting "*full control*") and authenticity ("*How authentic is the vocalisation?*", with 1 denoting "*not authentic at all*" and 7 denoting "*very authentic*"). Note that volitional laughter and crying were included in this pilot study as well (see Experiments 2, 4-5 and 7 for volitional laughter stimuli; volitional crying was not included in this thesis). These pilot ratings established that participants reliably rate spontaneous laughter and crying as higher in arousal and authenticity, lower in control over the production of the vocalisation, and more extreme in valence (more positive for laughter and more negative for crying, respectively) than their volitional counterparts. The speakers also produced series of short vowels ('staccato vowels'; /a/, /i/, /e/, /u/, /o/, average vowel duration within a series = .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 volitional, non-emotional stimulus was chosen as its acoustic structure resembles laughter and crying, given all three vocalisations are based on series of vocalic bursts (see **Figure 5**). Individual vocalisation exemplars were extracted from the recordings and normalised for RMS amplitude using PRAAT (Boersma & Weenink, 2010).

Perceptual features of the stimuli

Based on the ratings collected for a larger set of vocalisations in the pilot study, 25 stimuli per vocalisation (5 per speaker) were selected, choosing series of vowels that were neutral in valence ($M_{Valence}$: 3.92; CI[3.85, 3.99]) and low in arousal ($M_{Arousal}$: 2.68; CI[2.56, 2.81]) and spontaneous laughter and crying exemplars that were high in arousal ($M_{CryingS}$: 3.79, CI[3.61 3.96]; $M_{LaughterS}$: 4.78, CI[4.46, 5.10]; $t[48] = 5.691$, $p < .001$, Cohen's $d = 1.643$), and authenticity ($M_{CryingS}$: 3.58, CI[3.25, 3.91]; $M_{LaughterS}$: 4.79, CI[4.42, 5.16]; $t[48] = 5.022$, $p < .001$, Cohen's $d = 1.5$) – note that the stimulus set did not allow for a match of arousal or authenticity for $Laughter_S$ and $Crying_S$. All three vocalisation sets were matched for duration (M_{Vowels}: 2.55 secs, CI[2.43, 2.66]; $M_{CryingS}$: 2.61 secs, CI[2.50, 2.73]; $M_{LaughterS}$: 2.41 secs, CI[2.32, 2.61]; $F(2,48) = 1.31$, $p = .280$).

Acoustic features of the stimuli[2]

Table 1 shows an overview of the means for the acoustic properties of the stimuli. The following acoustic measures were used:

(1) **Duration:** The interval between the first zero-crossing of the onset to the final zero crossing after the offset of the vocalisation.

(2) **Burst duration:** The interval between the first zero-crossing of the onset to the final zero crossing of a vocalic burst.

(3) **Percentage of unvoiced segments:** Percentage of frames lacking harmonic structure.

(4) **Fo mean:** Computed using the auto-correlation method in PRAAT. Fo floor was set at 75 Hz and the Fo ceiling at 1000 Hz.

(5) **Fo standard deviation:** The standard deviation of the Fo mean

(6) **Spectral centre of gravity:** Measure for the mean height of the frequencies for each vocalisation, which captures the weighting of energy in the sound across the frequency range.

(7) **Mean harmonics-to-noise-ratio (HNR):** The mean ratio of quasi periodic to non-period signals across time segments.

(8) **Jitter:** The average absolute difference between consecutive periods, divided by the average period, i.e., micro-fluctuations in the duration of each period.

(9) **Shimmer:** The average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude.

T-tests were performed to assess acoustic differences between vocalisations (corrected for multiple comparisons, $\alpha = .006$). These tests showed that while Laughter_S and Crying_S were acoustically similar for all acoustic measures (all $ps \geq .016$) with the exception of spectral centre of gravity ($t[48] = 4.389$, $p < .001$, Cohen's d =

[2] No formant measures were extracted from the stimuli. Previous studies have extracted formant measures from non-verbal emotional vocalisations, such as laughter (e.g. Szameitat, Darwin, Szameitat, Wildgruber, Sterr, Dietrich & Alter, 2007a; Bachorowski, Smoski & Owren, 2001) but came to conflicting conclusions. For most vocalisations, especially for spontaneous ones, the authors report that it was difficult to extract reliable formant measures from a representative portion of the sounds (see Bachorowski et al., 2001, for a discussion). An analysis of such formant measures would thus have been biased and was therefore omitted from the current studies.

1.267), $Crying_S$ differed from Vowels in all acoustic measures (*p*s < .001) with the exception of total duration (*t*[48] = .805, *p* = .425, Cohen's *d* = .232), Fo variability (*p* = .029) and spectral centre of gravity (*t*{48] = .013, *p* = .994, Cohen's *d* = .001). The acoustic properties of $Laughter_S$ were significantly different from Vowels for all measures (*p* ≤ .001), except total duration (*t*{48] = .917, *p* = .364, Cohen's *d* = .131) and Fo standard deviation (*t*[48] = 2.706, *p* = .009; Cohen's *d* = .781). Despite constituting two different vocalisations, $Laughter_S$ and $Crying_S$ can be thus considered acoustically more similar to each other, while Vowels were acoustically very dissimilar to both $Laughter_S$ and $Crying_S$. For a detailed breakdown of the acoustic properties of the stimuli by speaker and by gender, see Table 1.

2.1.4 Methods

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/). All trials were timed, giving participants 2.5 seconds to make a response before moving on to the next trial. Participants were presented with 75 stimuli in total (25 per vocalisation; Vowels, $Laughter_S$ and $Crying_S$) in fully randomised order. During the presentation of the sounds, a fixation cross was presented on the screen, which was then replaced by a prompt asking participants to indicate whether the speaker was male or female (two-way forced choice) via a keyboard press. The task lasted for approximately 10 minutes.

					By Gender				By Speaker									
			All		Male		Female		CL		SE		CR		DB		SS	
Vocalisation	Acoustic Measure	Unit	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Vowels	Duration	secs	2.55	0.28	2.41	0.25	2.64	0.27	2.57	0.26	2.25	0.05	2.51	0.36	2.69	0.23	2.71	0.21
	Burst Duration (mean)	secs	0.36	0.10	0.34	0.12	0.37	0.08	0.40	0.15	0.29	0.04	0.34	0.08	0.41	0.11	0.35	0.05
	Unvoiced Segments	percent	22.24	13.77	25.48	10.89	20.08	15.37	16.73	5.37	34.24	6.81	36.74	16.27	11.74	4.68	11.77	4.43
	Fo (mean)	Hz	206.42	63.21	140.12	28.50	250.61	33.08	163.96	19.83	116.27	3.50	263.30	22.20	230.35	45.52	258.19	21.90
	Fo (SD)	Hz	78.33	52.95	54.10	44.29	94.48	53.35	85.51	40.19	22.68	18.22	84.94	31.35	77.09	79.35	121.42	35.82
	Spectral center of gravity	Hz	688.13	370.35	786.15	551.51	622.79	167.79	404.20	184.11	1168.10	534.56	618.39	143.85	677.25	233.56	572.74	128.19
	HNR	Hz	17.91	4.87	13.89	3.77	20.59	3.50	13.76	3.08	14.02	4.74	21.46	3.82	21.07	4.31	19.25	2.50
	Jitter	dB	1.28	0.54	1.57	0.48	1.08	0.50	1.91	0.48	1.24	0.11	1.10	0.67	0.77	0.36	1.38	0.25
	Shimmer	dB	0.58	0.22	0.67	0.23	0.53	0.21	0.82	0.19	0.51	0.14	0.51	0.11	0.45	0.34	0.62	0.09
$Crying_S$	Duration	secs	2.61	0.30	2.70	0.33	2.56	0.27	2.75	0.13	2.64	0.47	2.65	0.30	2.40	0.31	2.62	0.13
	Burst Duration (mean)	secs	0.19	0.13	0.15	0.05	0.21	0.16	0.16	0.01	0.14	0.07	0.17	0.02	0.16	0.07	0.30	0.27
	Unvoiced Segments	percent	53.86	16.69	52.59	15.65	54.70	17.84	43.09	10.89	62.09	14.39	65.14	9.97	60.74	6.66	38.22	21.19
	Fo (mean)	Hz	454.95	102.64	387.88	85.39	499.67	89.54	417.11	20.76	358.65	117.65	547.58	48.50	399.24	26.55	552.19	77.96
	Fo (SD)	Hz	108.21	40.08	117.22	35.90	102.21	42.77	120.43	17.15	114.02	50.80	83.81	36.78	118.89	41.37	103.92	50.62
	Spectral center of gravity	Hz	687.55	172.33	577.97	146.54	760.60	151.03	687.97	66.96	467.98	116.57	803.84	177.74	741.60	98.47	736.35	187.14
	HNR	Hz	11.16	5.05	9.81	2.73	12.06	6.07	10.83	1.80	8.79	3.30	13.04	1.86	7.65	2.35	15.49	8.92
	Jitter	dB	2.98	1.35	3.76	1.36	2.46	1.11	2.92	0.74	4.61	1.34	1.97	0.66	3.39	0.38	2.03	1.45
	Shimmer	dB	1.12	0.37	1.20	0.17	1.06	0.45	1.14	0.18	1.27	0.15	0.79	0.20	1.48	0.26	0.91	0.52
$Laughter_S$	Duration	secs	2.47	0.35	2.75	0.14	2.28	0.31	2.76	0.08	2.74	0.19	2.22	0.27	2.36	0.20	2.25	0.47
	Burst Duration (mean)	secs	0.11	0.06	0.14	0.08	0.10	0.04	0.15	0.11	0.13	0.05	0.08	0.02	0.14	0.02	0.07	0.04
	Unvoiced Segments	percent	58.82	14.75	57.27	14.32	59.85	15.45	49.99	14.97	64.56	10.22	61.15	8.19	43.37	4.41	75.04	10.89
	Fo (mean)	Hz	490.23	115.89	417.65	109.96	538.61	94.77	418.19	120.55	417.10	112.57	559.47	110.00	499.35	56.20	557.03	115.24
	Fo (SD)	Hz	113.22	36.78	115.46	34.70	111.73	39.23	136.49	30.66	94.42	25.75	105.75	14.56	115.42	32.60	114.01	63.58
	Spectral center of gravity	Hz	1047.98	372.70	838.38	260.52	1187.72	377.08	687.21	161.96	989.54	263.35	1060.85	276.67	1293.35	141.86	1208.94	605.29
	HNR	Hz	9.36	2.72	9.61	1.69	9.19	3.28	9.45	1.29	9.77	2.17	11.43	3.31	8.56	2.14	7.60	3.48
	Jitter	dB	3.05	1.11	3.13	0.64	3.00	1.36	3.31	0.80	2.95	0.44	2.21	0.97	2.62	1.08	4.16	1.30
	Shimmer	dB	1.15	0.25	1.16	0.14	1.14	0.30	1.11	0.17	1.20	0.10	0.96	0.28	1.03	0.19	1.44	0.20

Table 1 Table of means and standard deviation of acoustic descriptors of vocalisations used in Experiment 1 and 3.

2.1.5 Results

Sex recognition from volitional and spontaneous non-verbal vocalisations

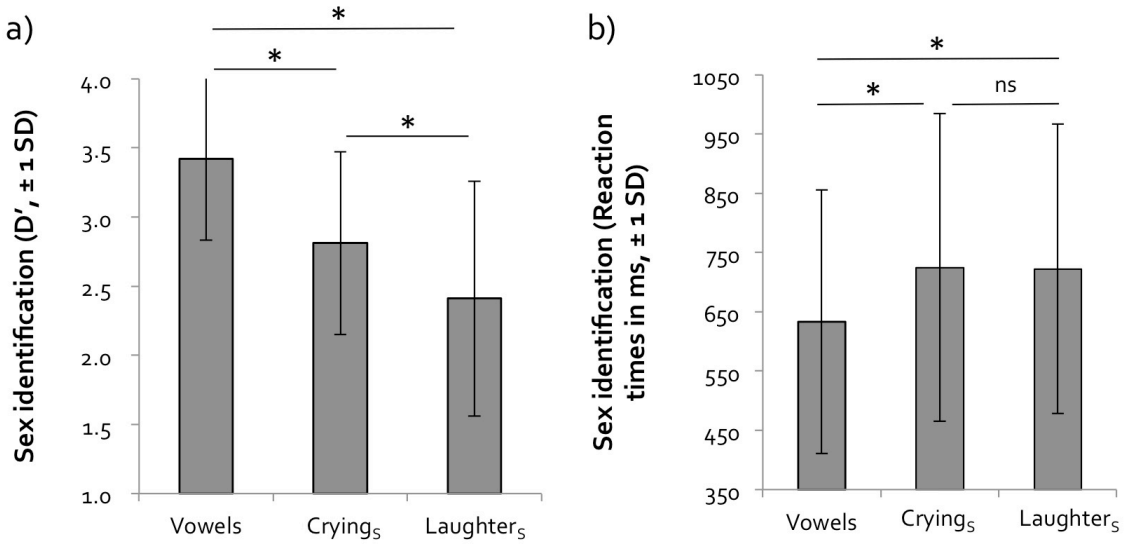

Figure 6 a) Average *d'* scores per vocalisation for the sex identification task, b) average reaction times per vocalisation for the sex identification task of Experiment 1. Significant results ($p < .017$) are highlighted with an asterisk.

For an analysis by subject, *d'* scores were calculated from the raw responses. For this study, responding "female" to a vocalisation produced by a female speaker was scored as a hit, responding "male" to a vocalisation produced by a female speaker was scored as a miss. Responding "male" to a vocalisation produced by a male speaker was scores as a correct rejection and responding "female" to a vocalisation produced by a male was scored as a false alarm. 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 could range from zero to 4.11, with a *d'* score of zero indicating that listeners were not able to discriminate between speaker sex while gradually higher scores indicate a greater ability to discriminate between speaker sex (Stanislaw & Todorov, 1999).

D' scores were entered into a one-way repeated measures ANOVA. There was a main effect of vocalisation on performance of sex identification (F[2,86] = 25.47, p < .001, η_p^2 = .37). Three post-hoc paired t-tests (alpha = .017, Bonferroni corrected) showed that performance was lower for identifying speaker sex from Laughter$_S$ compared to Vowels (t[44]= 6.22, p < .001, Cohen's d = 1.875) and as well as Crying$_S$ (t[44]= 2.72, p = .009, Cohen's d = .82). Furthermore, performance was also significantly lower for Crying$_S$ compared to Vowels (t[44]= 5.26, p < .001, Cohen's d = 1.586) (**Figure 6**a). A one-way repeated measures ANOVA on reaction times confirmed a main effect of vocalisation (F[2,86] = 16.35, p < .001, η_p^2 = .28). Post-hoc t-tests showed that reaction times were significantly faster for Vowels compared to Laughter$_S$ (t[44]= -4.60, p < .001, Cohen's d = 1.39) and Crying$_S$ (t[44]= -4.89, p < .001, Cohen's d = 1.474), while reaction times for Crying$_S$ and Laughter$_S$ were similar (t[44]= .11, p = .913, Cohen's d = .03) (**Figure 6**b). *D'* values are high for all vocalisations and correspond to average accuracy scores of 92.65% for Vowels, 90.21% for Crying$_S$ and 80.36% for Laughter$_S$.

A response bias analysis was conducted: C was calculated by averaging the z scores of the hit and false alarm rate and multiplying it by minus (see Stanislaw & Todorov, 1999). Zero indicated no bias, negative values indicate a bias towards responding 'female', positive values indicate a bias towards respond 'male'. For Vowels, a significant bias towards responding 'male' was found, whereas for Crying$_S$ and Laughter$_S$ significant biases towards responding 'female' were found (all ps ≤ .003).

Differences in perception of speaker sex for male and female vocalisations

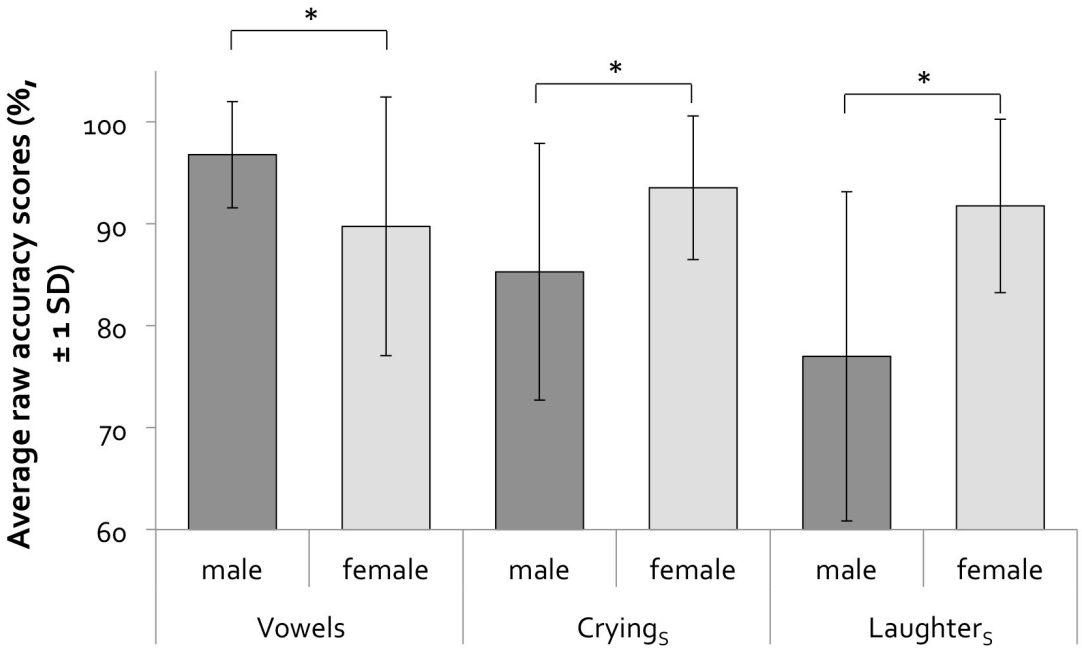

Figure 7 Raw accuracy scores per item split for male and female vocalisations for the speaker sex identification task

As the speaker set contained male and female voices, further analysis attempted to assess whether there were differences in the perception of male and female vocalisations in the experiment. A 2 (speaker sex) x 3 (vocalisation type) ANOVA was performed. There was a significant main effect of speaker sex (F[1,43] = 10.232, p = .003, η_p^2 = .192) and vocalisation type (F[2,86] = 17.398, p < .001, η_p^2 = .288) as well as an interaction of speaker sex and vocalisation type (F[2,86] = 26.971, p < .001, η_p^2 = .385). Post-hoc t-test confirmed that performance for male and female vocalisations was significantly different for all vocalisations (all ps ≤ .002), with performance being better for female spontaneous vocalisations, and better for male vowels (**Figure 7**).

Linking performance accuracy to acoustic features

To establish whether and which acoustic cues were particularly salient for sex perception from vocal signals in the context of this study, acoustic properties of the stimuli were linked to the raw accuracy of speaker sex for the individual tokens. Initially, two multiple regression models were run, by speaker sex (one for stimuli produced by male speakers and one for stimuli produced by female speakers), including Fo mean as a predictor of raw accuracy. It was hypothesised that higher Fo in females should enhance accuracy, while higher Fo in males should decrease accuracy (see Honorof & Whalen, 2010). For female vocalisations, Fo mean was not a significant predictor for accuracy (R^2 = .047, β = .218, t[44] = 1.462, p = .151), while changes in Fo mean predicted accuracy for male vocalisation, showing in line with predictions lower accuracy for higher pitched sounds (R^2 = .383, β = -.618, t[29] = -4.171, p < .001). To further explore whether and how Fo is linked to accuracy within each vocalisation, six additional regression analyses were run (2 speaker sex x 3 vocalisations). Fo mean was not found to be a significant predictor of accuracy in any of these models (all ps ≥ .186) with the exception of male crying where against predictions higher pitch was associated with better accuracy (R^2 = .506, β = .721, t[9] = 2.865, p = .021).

To investigate whether acoustic properties other than Fo significantly predicted accuracy, further multiple regression analyses were run, entering all remaining acoustic measures as predictors using the Enter method. These were run with Total Duration, Percentage of Unvoiced Segments, Burst Duration, Fo SD, HNR and Spectral Centre of Gravity as predictors (Fo min, Fo max, Shimmer and Jitter were excluded from the analyses due to avoid excessive collinearity). Models were run

across all vocalisations and for each individual type of vocalisation. In contrast to the previous multiple regression analysis, there were no *a priori* hypotheses regarding different directionalities of effects within speaker sex, therefore no analysis by speaker sex was performed. The regression model including all vocalisations was significant (R^2 = .186, p = .026) with total duration being a significant predictor of accuracy (β = -.257, t[74] = -2.102, p = .039, all other ps ≥ .121), counterintuitively indicating that accuracy was higher for shorter vocalisations. Furthermore, the regression model for Vowels was significant (R^2 = *435*, p = .009) with HNR being a significant predictor (β = -.7, t[24] = -3.285, p = .004, all other ps ≥ .059), indicating higher accuracy for vocalisations that were more harmonic within the voiced portions of the signal. The acoustic predictors did not explain a significant amount of variance for regressions models looking at $Laughter_S$ and $Crying_S$.

2.1.6 Discussion

The current experiment explored whether vocalisations produced under reduced volitional control would affect the extraction of speaker sex from non-verbal vocal signals. Performance was impaired for $Laughter_S$ and $Crying_S$ compared to Vowels, in line with the prediction that reduced volitional control during voice production has a detrimental effect on the extraction of speaker sex. Intriguingly, performance for $Crying_S$ was, however, also significantly better compared to $Laughter_S$. Reaction times supported the prediction with response latencies being significantly longer for spontaneous versus volitional vocalisations, potentially indicating greater task difficulty.

The distinct perceptual and acoustic features of these spontaneous vocalisations may have affected participants' performance in accurately identifying speaker sex: in spontaneous vocalisations, drastic changes in acoustic properties, such as Fo, occur. Cues to speaker sex that are usually encoded within the same acoustic properties are affected in spontaneous vocalisation production. These diagnostic acoustic cues may thus be 'overwritten' or become perceptually less salient (e.g. global modulations of Fo for laughter, potentially resulting in less marked differences between male and female laughter or shifting vocalisations produced by males into the acoustic space traditionally inidicative of vocalisations produced by females) – these changes in diagnostic cues should therefore impair task performance. As an alternative hypothesis, performance could have been affected by perceptual and affective qualities of spontaneous emotional vocalisations: Previous research has shown that emotional content captures a perceiver's attention (Öhman, Flykt & Esteves, 2001; Vuilleumier, 2005). With such emotional content requiring immediate assessment, its processing may be prioritised over the extraction of other types of information, such as cues to identity. In this specific context, cues to speaker characteristics, such as speaker sex, may be less relevant to a perceiver compared to the emotional information conveyed, thus performance on judgements of speaker sex are impaired (Stevenage & Neil, 2014; see also Goggin, Thompson, Strube & Simental, 1991). A third alternative explanation may be found in the differences in the exposure of listeners to volitional vocal signals and spontaneous vocal signals – in everyday experience, listeners are more familiar with extracting speaker characteristics from volitional vocal signals as these are more frequent, thus potentially leading to poorer performance for comparatively unfamiliar spontaneous vocalisations.

Against predictions, performance for $Laughter_S$ was significantly lower compared to $Crying_S$. This may be due to the stimuli selected for $Laughter_S$ being rated significantly higher in arousal and authenticity than those selected for $Crying_S$ in this study. Arousal and volitional control have been shown to be intimately linked (Lavan et al., 2016; McKeown, et al., 2015) and the $Laughter_S$ used in these experiments is therefore arguably less controlled in its production, leading to more extreme changes in the signal and consequently impairments to performance. In line with previous studies that report above-chance accuracy for judgements of speaker sex despite acoustic manipulations of the signal (Honorof & Whalen, 2010; Lass et al., 1995; Mullenix et al., 1995), the current findings confirm that speaker sex is encoded in a range of acoustic features, resulting in a robust percept of speaker sex despite drastic changes introduced to the signal: if one salient acoustic cue, such as Fo in the current stimuli, is modulated and may thus become relatively less salient and diagnostic, other acoustic cues to speaker sex may still remain informative to listeners.

Intriguingly, speaker sex effects were apparent in the data set: When analysing raw accuracy scores per item and split by speaker gender, accuracy for spontaneous vocalisations was higher for vocalisations produced by females, while it was lower for female Vowels, a volitional vocalisation. Furthermore, response biases towards 'female' for $Laughter_S$ and $Crying_S$ and towards 'male' for Vowels were apparent. This may be a further indication that important markers for speaker sex, such as Fo have been modulated drastically in males in spontaneous vocalisations, approximating Fo values frequently encountered in female vocalisations. This finding was confirmed in a regression analysis showing that a higher Fo leads to lower accuracy in speaker sex identification for males but not for females. It should, however, be noted that this may

merely reflect an effect of vocalisation type, showing a trend of accuracy decreasing across these vocalisation (Vowels > $Crying_S$ > $Laughter_S$) while Fo mean increases (Vowels < $Crying_S$ < $Laughter_S$, see Table 1). This finding more generally poses intriguing questions regarding the role of Fo across vocalisations for, for example, mate selection: Based on anatomical features of the male and female vocal tract, lower pitch Fo in vocal signals is generally associated with male speakers while higher Fo is associated with female speakers. A low Fo in males has been shown to be perceived to be attractive and dominant by female listeners, as it arguably signals favourable mate choices (see e.g. Puts et al, 2016). For the spontaneous vocalisations used in this study, the sexual dimorphism is drastically reduced between males and females. Thus, the production of high-Fo spontaneous vocalisations as a male should have detrimental effects on reproduction success. Future studies should determine whether the role of Fo is comparable across vocalisations and the signalling context.

No clear relationship was observed between other acoustic measures and accuracy, within or across vocalisations: Only Fo predicted accuracy in $Crying_S$ as well as across all vocalisations produced by males – although against predictions: higher Fo was related to higher accuracy for $Crying_S$ but lower accuracy across all vocalisations. This was, however, not the case for $Laughter_S$ or Vowels, indicating that vocalisation-specific acoustic effects of may be present. These results should however be treated with caution, due to the limited number of stimuli and lack of variability in the data.

While there are some indications that acoustic features may have affected listeners' performance, their specific role is still unclear. It is furthermore unclear from the current set of results which specific perceptual properties of the spontaneous vocalisations drive this detrimental effect on performance. First, from the current

results it cannot be determined whether changes in performance are due to the acoustic consequences of differences in the degree of control of the voice production only, or whether there may be effects of vocalisation type: Non-verbal emotional vocalisations may be generally less informative for decoding speaker characteristics compared to vowels. They may furthermore not be well attended to in everyday communication in general, as reported for laughter where the frequency of laughter in everyday interactions is dramatically underestimated in self-report questionnaires (Vettin & Todt, 2004). If this were the case, vocalisation type (i.e. laughter versus crying versus vowels) *per se* could have a greater influence on the observed effects instead of other features, such as the spontaneous nature of the vocalisations. Finally, the effects may be a simple effect of increasing arousal (Vowels < $Crying_S$ < $Laughter_S$). Experiment 2 addresses these issues by contrasting volitional and spontaneous laughter stimuli that were better matched in arousal and can be classed under a single type of vocalisation but differ in the degree of emotional content and volitional control.

2.2 Experiment 2

2.2.1 Introduction

In Experiment 2, the perception of speaker sex from $Laughter_S$ (produced during authentic amusement under reduced volitional control), $Laughter_V$ (produced on demand, under full volitional control) and Vowels will be compared as a follow-up study to Experiment 1. For this experiment, $Laughter_V$ and $Laughter_S$ were better matched for arousal (see Section 2.2.3) but showed within-vocalisation differences in volitional control over production. With these stimuli, it could thus be directly assessed whether detrimental effects on performance in the tasks in Experiments 1 resulted from reduced volitional control over the voice, differences in arousal between vocalisations or whether the effects were indicative of general differences in the perceptual processing of different vocalisation types regardless of levels of volitional control. If differences in vocalisation type or arousal modulate performance, performance for $Laughter_S$ and $Laughter_V$ should be lower than for Vowels, as both types of laughter differ from Vowels in arousal and are a different type of vocalisation. If reduced volitional control over production modulates performance, performance for Vowels and $Laughter_V$ should be higher than for $Laughter_S$.

2.2.2 Participants

43 participants (39 female; M_{Age}: 19.2 years; *S*D: 1.1 years; range 19-21 years) were recruited at the Department of Psychology at Royal Holloway, University of London and received course credit for their participation. The sample size was determined based on the sample size of the previous study in which a similar sample size resulted in reliable effects. No participant reported any hearing difficulties. Ethical approval

was obtained from the Departmental Ethics Committee. None of the participants were familiar with the speakers used. Average performance across conditions for each participants was within 2 standard deviations from the mean and therefore all participants were included in the following analyses.

2.2.3 Materials

Materials were the same as in Experiment 1 with the exception that $Crying_S$ was replaced by $Laughter_V$ produced by the same 5 speakers (see Experiment 1). The procedure for the recording and elicitation procedure was as described in McGettigan et al. (2015). In short: For $Laughter_V$, the speakers were instructed to produce natural and positive sounding laughter, without inducing a specific emotional state. Thus, $Laughter_V$ was produced under full volitional control over the voice (and in the absence of amusement), while $Laughter_S$ was produced spontaneously and thus under reduced volitional control, in response to viewing and listening to amusing stimuli. $Laughter_V$ was recorded in the same session as $Laughter_S$, with $Laughter_V$ always being recorded first to avoid carry-over effects. Based on the ratings from the pilot study (see Experiment 1), 25 $Laughter_V$ stimuli (5 per speaker) were selected.

Perceptual features of the stimuli

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]; $t[48] = 5.881$, $p < .001$). $Laughter_S$ and $Laughter_V$ were significantly higher in arousal than Vowels ($Laughter_V$: $t[48] = 12.789$, $p < .001$, Cohen's $d = 3.692$; $Laughter_S$: $t[48] = 13.147$, $p < .001$, Cohen's $d = 3.795$), but more closely matched (compared to the highly

significant differences in arousal between $Laughter_S$ and $Crying_S$ in Experiment 1), albeit still significantly 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*[48]= 2.085, *p* = .042, Cohen's *d* = .602). Note that high correlations between arousal and authenticity ratings are present for laughter. In order to minimise the differences in arousal, the laughs that were rated as most/least authentic could not be included in the stimulus set – resulting in highly significant but not extreme differences in perceived authenticity for $Laughter_S$ and $Laughter_V$. 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[48]= .208, *p* = .836, Cohen's *d* = .06). The overall duration of the stimuli was matched (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).

Acoustic features of the stimuli

Table 2 shows an overview of the means for the acoustic properties of the stimuli. T-tests were performed to assess acoustic differences between vocalisations (corrected for multiple comparisons, α =.006). As has been reported for Experiment 1, the acoustic properties of $Laughter_S$ were significantly different from Vowels for all measures (*p* ≤ .001), except total duration (*t*{48] = .917, *p* = .364, Cohen's *d* = .131) and Fo standard deviation (*t*[48] = 2.706, *p* = .009; Cohen's *d* = .781). $Laughter_V$ was also distinct from Vowels for all acoustic measures (*p*s ≤ .003) with the exception of duration (*t*[48] = 2.456, *p* = .018; Cohen's *d* = .709) spectral centre of gravity (*t*[48] = 2.001, *p* = .051; Cohen's *d* = .577). $Laughter_S$ and $Laughter_V$ were similar to each other across spectral and temporal features (total duration, Fo variability, spectral centre of

gravity, deviation of spectral centre of gravity, percentage of unvoiced segments and burst duration; all *p*s ≥ .028) and differed significantly from each other in measures of fundamental frequency and periodicity (Fo mean, Fo minimum, Fo maximum, Fo range, HNR, Shimmer, Jitter, all *p*s ≤ .004). Thus, while $Laughter_S$ and Vowels are acoustically most different from each other, $Laughter_V$ appears to fall between the two vocalisations.

2.2.4 Methods

The experimental set up was identical to the one used in Experiment 1. Participants were presented with all 75 stimuli (25 per vocalisation; Vowels, $Laughter_S$, $Laughter_V$; see Section **Error! Reference source not found.**) in a fully randomised order. Participants were not pre-informed about the inclusion of spontaneous and volitional laughter in the tasks.

2.2.5 Results

Sex recognition from volitional and spontaneous non-verbal vocalisations

Figure 8 Average *d'* scores per vocalisation for the sex identification task, b) average reaction times per vocalisation for the sex identification task of Experiment 2.

			By Gender				By Speaker									
			Male		Female		CL		CR		DB		SE		SS	
Vocalisation	Acoustic Measure	Unit	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Vowels	Duration	secs	2.41	0.25	2.64	0.27	2.57	0.26	2.51	0.36	2.69	0.23	2.25	0.05	2.71	0.21
	Burst Duration (mean)	secs	0.34	0.12	0.37	0.08	0.40	0.15	0.34	0.08	0.41	0.11	0.29	0.04	0.35	0.05
	Unvoiced Segments	percent	25.48	10.89	20.08	15.37	16.73	5.37	36.74	16.27	11.74	4.68	34.24	6.81	11.77	4.43
	Fo (mean)	Hz	140.12	28.50	250.61	33.08	163.96	19.83	263.30	22.20	230.35	45.52	116.27	3.50	258.19	21.90
	Fo (SD)	Hz	54.10	44.29	94.48	53.35	85.51	40.19	84.94	31.35	77.09	79.35	22.68	18.22	121.42	35.82
	Spectral center of gravity	Hz	786.15	551.51	622.79	167.79	404.20	184.11	618.39	143.85	677.25	233.56	1168.10	534.56	572.74	128.19
	HNR	Hz	13.89	3.77	20.59	3.50	13.76	3.08	21.46	3.82	21.07	4.31	14.02	4.74	19.25	2.50
	Jitter	dB	1.57	0.48	1.08	0.50	1.91	0.48	1.10	0.67	0.77	0.36	1.24	0.11	1.38	0.25
	Shimmer	dB	0.67	0.23	0.53	0.21	0.82	0.19	0.51	0.11	0.45	0.34	0.51	0.14	0.62	0.09
$Laughter_V$	Duration	secs	2.15	0.23	2.44	0.40	2.13	0.21	2.27	0.41	2.40	0.45	2.16	0.26	2.64	0.32
	Burst Duration (mean)	secs	0.09	0.02	0.08	0.03	0.09	0.02	0.06	0.01	0.08	0.02	0.08	0.02	0.11	0.02
	Unvoiced Segments	percent	54.50	11.64	60.00	9.28	56.70	7.32	60.06	7.34	65.65	9.12	52.31	15.46	54.29	9.16
	Fo (mean)	Hz	300.94	61.56	333.99	90.47	318.69	73.69	396.77	92.55	265.24	87.28	283.18	48.06	339.96	39.82
	Fo (SD)	Hz	114.95	38.29	136.41	68.71	107.15	40.00	166.63	60.88	87.56	70.87	122.75	39.32	155.03	56.88
	Spectral center of gravity	Hz	830.41	257.87	898.88	283.08	847.05	237.75	1156.17	134.65	708.90	302.42	813.76	303.97	831.58	193.31
	HNR	Hz	6.12	1.74	5.44	1.99	5.19	1.72	4.13	0.95	6.75	1.92	7.05	1.31	5.45	2.22
	Jitter	dB	3.74	0.60	4.09	0.46	3.94	0.49	4.23	0.25	4.04	0.66	3.54	0.68	4.01	0.44
	Shimmer	dB	1.30	0.22	1.43	0.31	1.39	0.27	1.65	0.14	1.14	0.31	1.20	0.13	1.49	0.23
$Laughter_S$	Duration	secs	2.75	0.14	2.28	0.31	2.76	0.08	2.22	0.27	2.36	0.20	2.74	0.19	2.25	0.47
	Burst Duration (mean)	secs	0.14	0.08	0.10	0.04	0.15	0.11	0.08	0.02	0.14	0.02	0.13	0.05	0.07	0.04
	Unvoiced Segments	percent	57.27	14.32	59.85	15.45	49.99	14.97	61.15	8.19	43.37	4.41	64.56	10.22	75.04	10.89
	Fo (mean)	Hz	417.65	109.96	538.61	94.77	418.19	120.55	559.47	110.00	499.35	56.20	417.10	112.57	557.03	115.24
	Fo (SD)	Hz	115.46	34.70	111.73	39.23	136.49	30.66	105.75	14.56	115.42	32.60	94.42	25.75	114.01	63.58
	Spectral center of gravity	Hz	838.38	260.52	1187.72	377.08	687.21	161.96	1060.85	276.67	1293.35	141.86	989.54	263.35	1208.94	605.29
	HNR	Hz	9.61	1.69	9.19	3.28	9.45	1.29	11.43	3.31	8.56	2.14	9.77	2.17	7.60	3.48
	Jitter	dB	3.13	0.64	3.00	1.36	3.31	0.80	2.21	0.97	2.62	1.08	2.95	0.44	4.16	1.30
	Shimmer	dB	1.16	0.14	1.14	0.30	1.11	0.17	0.96	0.28	1.03	0.19	1.20	0.10	1.44	0.20

Table 2 Table of means and standard deviation of acoustic descriptors of vocalisations used in Experiment 2, 4 and 5.

Results are shown in **Figure 8**. Data were analysed in the same way as in Experiment 1 (see Section **Error! Reference source not found.**). *D'* scores were entered into a one-way repeated measures ANOVA. This showed a main effect of vocalisation on performance ($F[2,84] = 28.93$, $p < .001$, $\eta_p^2 = .41$). Three post-hoc paired t-tests (corrected alpha = .017, Bonferroni correction) confirmed that performance was worse when identifying speaker sex from $Laughter_S$ compared to $Laughter_V$ ($t[43] = 7.77$, $p < .001$, Cohen's $d = 2.37$) and Vowels ($t[43] = 6.05$, $p < .001$, Cohen's $d = 1.8452$), while performance was similar for $Laughter_V$ and Vowels ($t[43] = -.22$, $p = .824$, Cohen's $d = .067$) (**Figure 8**a). A one-way repeated measures ANOVA on reaction times confirmed a main effect of vocalisation ($F[2, 84] = 13.27$, $p = .001$, $\eta_p^2 = .20$). Post-hoc t-tests (Bonferroni-corrected alpha = .017) showed that reaction times were significantly slower for $Laughter_S$ compared to Vowels ($t[43] = -3.643$, $p = .001$) and $Laughter_V$ ($t[43] = -4.29$, $p < .001$, Cohen's $d = 1.308$) (**Figure 8**a), while reaction times did not differ for $Laughter_V$ and Vowels ($t[43] = .08$, $p = .938$, Cohen's $d = .02$) (**Figure 8**b). In line with Experiment 1, *D'* values were generally high for each vocalisation and correspond to average accuracy scores of 94.38% for Vowels, 94.75% for $Laughter_V$ and 86.78% for $Laughter_S$.

In line with Experiment 1, a response bias analysis was conducted but no significantly biases were found (all $ps \geq 274$).

Differences in perception of speaker sex for male and female vocalisations

In line with the analysis of Experiment 1, further analysis attempted to assess differences in the perception of male and female vocalisations. A 2 (speaker sex) x 3 (vocalisation type) ANOVA was performed. There was a significant main effect of

vocalisation type (F[2,86] = 16.132, p < .001, η_p^2 = .282). Neither the main effect of speaker sex (p = .081) nor the interaction of speaker sex and vocalisation type (p = .836) were significant, means are illustrated in **Figure 9**.

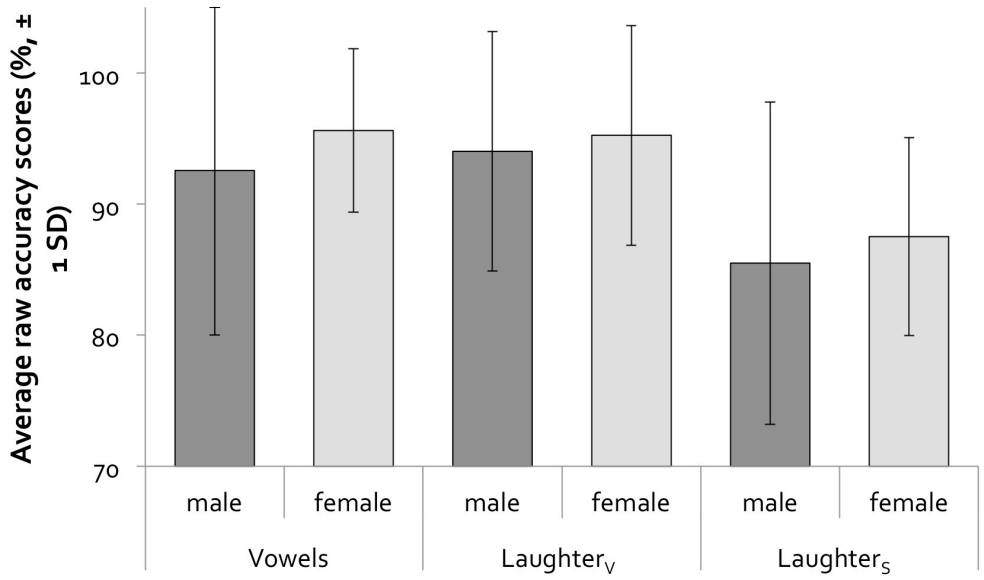

Figure 9 Raw accuracy scores per item split for male and female vocalisations for the speaker sex identification task.

Linking accuracy to acoustic features

To establish whether acoustic cues were particularly salient for sex perception from these vocal signals, acoustic properties of the stimuli were tested as predictors of the raw accuracy of speaker sex for the individual tokens. In the first instance, two multiple regression models were run, by speaker sex (one for stimuli produced by male speakers and one for stimuli produced by female speakers), including Fo mean as a predictor of raw accuracy. Higher Fo in females should increase accuracy, while higher pitch in males should decrease accuracy. For female vocalisations, Fo mean significantly predicted accuracy, unexpectedly showing higher accuracy for lower pitched sounds (R^2 = .42, β = -.648, t[44] = -5.548, p < .001). Variation in Fo mean furthermore predicted accuracy for male vocalisations, (R^2 = .162, β = -.403, t[29] = -

2.311, p = .027), again showing higher accuracy for sounds with lower Fo. To further explore if Fo could be linked to accuracy within each vocalisation, six additional regression analyses were run (2 speaker sex x 3 vocalisations). Fo mean was not found to be a significant predictor of accuracy in any of these models (all ps > .245 regression analyses).

Further multiple regression analyses, including all remaining acoustic measures as predictors using the Enter method, were run with Total Duration, Percentage of Unvoiced Segments, Burst Duration, Fo SD, HNR and Spectral Centre of Gravity as predictors. In line with the analyses of Experiment 1, Fo min, Fo max, Shimmer and Jitter were excluded from the analyses to avoid excessive collinearity. As in Experiment 1, there were no *a priori* hypotheses regarding different directionalities of effects within speaker sex, therefore no separate analysis by speaker sex was performed. None of the models were able to explain a significant amount of variance in accuracy (all ps ≥ .229).

2.2.6 Discussion

By using Laughter_V and Laughter_S, Experiment 2 explored whether the effects observed in Experiment 1 reflected processing differences for different types of vocalisations, or whether they could have resulted from differences in the specific affective and perceptual qualities associated with spontaneous vocalisations. Performance was lower for Laughter_S than for Vowels and Laughter_V, while performance was comparable for Vowels and Laughter_V. This pattern was also reflected in reaction times. The current results thus indicate that reduced volitional control, and not arousal or vocalisation type *per se*, affects sex identification

performance: If arousal or vocalisation type affected the listener's perception of speaker sex, performance for $Laughter_V$ should have been lower than performance for Vowels, as $Laughter_V$ is both higher in arousal and a different type of vocalisation from Vowels.

In contrast to Experiment 1, no systematic differences in accuracy for male versus female vocalisations was found, despite considerable overlap in the stimuli used between the two experiments – this indicates that the overall context created by the stimuli used within an experiment may affect participants' responses. In line with the results of Experiment 1, no clear relationship between acoustic cues and accuracy per item was found: Regression analyses did, however, indicate that, in line with Experiment 1, performance was higher for lower-pitched vocalisations produced by males (across all vocalisations). Additionally, performance was higher for lower Fo sounds for female speakers across all vocalisations – it is unclear why this might be the case but it suggests that there is no direct and linear relationship between changes in Fo and sex judgements from vocal sounds. No other acoustic predictors explained significant proportions of the variance in judgements for any of the conditions. Thus, in the context of highly variable vocalisations, changes in acoustic properties cannot be easily linked to changes in performance.

2.3 General discussion

Experiment 1 and 2 investigated whether the identification of speaker sex from non-verbal vocalisations is affected by vocal flexibility, introduced by using different types of vocalisations (laughter, crying, vowels) produced under different level of volitional control (spontaneous versus volitional vocalisations). Results indicate that

performance is impaired for spontaneous compared to volitional vocalisations, with graded differences being apparent for different types of vocalisations (better performance for $Crying_S$ compared to $Laughter_S$). It remains difficult to dissociate the underlying mechanisms driving the reported effects: On the one hand, perceptual qualities of the stimuli may be driving the effect, with attention being automatically captured by spontaneous (emotional) vocalisations (Öhman et al., 2001; Vuilleumier, 2005). This may then allow for the rapid and prioritised processing of crucial emotional information in favour of the accurate extraction of identity-related information (Stevenage & Neil, 2014; see also Goggin, et al., 1991). On the other hand, spontaneous vocalisations occupy an acoustic space that is relatively distinct from that of volitional vocal signals. Acoustic cues that are diagnostic for speaker sex in neutral vocal signals are drastically modulated during the production of spontaneous vocalisations, rendering them less diagnostic and thus potentially impairing task performance.

The analyses attempting to link acoustic features to the perception of speaker sex partially support this claim, showing that for vocalisations produced by males, performance is highest for vocalisations with a relatively low F0 – notably, however, no evidence for these trends was found within analyses per vocalisation, and as noted in Experiment 1, these results could therefore be attributed to simple effects of acoustic differences across the different categories of vocalisation used in the two experiments. There is therefore overall no clear relationship between acoustic parameters and listeners' performance. Despite a large literature showing clear relationships between changes in acoustic features and perceptual qualities of sounds (e.g. Honorof & Whalen, 2010; Mullenix et al., 1995), the effects in the current study

thus appear not to be primarily driven by linear variations in acoustic parameters. An explanation for this could lie in the choice of acoustic measures and their extraction: Some of the acoustic features (jitter, shimmer, spectral measures) analysed in the current study may not be perceptually meaningful in the context of the vocalisations or task used (see e.g. Kreiman, Gerratt, Garellek, Samlan & Zhang, 2014 for a discussion). Furthermore, while Fo is known to be a salient cue for speaker sex judgements from volitional speech sounds, its role and importance in determining speaker sex is largely unknown for other types of vocalisations. With Fo being modulated in non-verbal emotional vocalisations (volitional or spontaneous, Bryant & Aktipis 2014; Lavan et al, 2016; McGettigan et al. 2015), its importance and salience may be reduced, while other cues (that may not accounted for in the current analyses) may gain perceptual importance.

One limitation of these (and the following) experiments is that no formal ratings of the speaker's emotional state during the production of the laughter and crying were collected: speakers only informally reported genuine feelings of amusement or sadness during and after the recordings sessions (see Materials). The possibility that speakers were not genuinely sad or amused during the production of the 'spontaneous' vocalisations cannot be fully ruled out. If this were the case, the difference between what is labelled here as 'spontaneous' and 'volitional' vocalisations as well as the underlying causes for the effects reported here become less clearly defined. Future studies and stimulus recording sessions should thus obtain online (or post-hoc) data of the emotional state of speaker during the recording sessions. It should also be noted that while the decrease in performance for $Laughter_S$ and $Crying_S$ compared to $Laughter_V$ and Vowels was significant, performance across

vocalisations was overall still high (*d'* values > 2, maximum: 4.11), indicating that identifying speaker sex was relatively easy – despite drastic modulations of Fo. Since there was only relatively little variability within the accuracy scores per item, this may have additionally affected the results of the regression analyses. Given this high performance, future studies should thus increase task difficulty by, for example, presenting stimuli in noise to create more variability in performance. Intriguingly, despite the generally high performance, large individual differences are apparent for both experiments. Assessing and describing the factors underlying these differences will be a challenge for further research into how variability in vocal signals affects speaker sex identification and judgements of speaker characteristics.

3 Speaker discrimination from volitional and spontaneous vocalisations

Experiments 3 and 4 investigate how the perception of another speaker characteristic, that is speaker identity, is affected by vocal flexibility, introduced by spontaneous and volitional vocalisations. Participants performed a speaker discrimination task on pairs of vowels, spontaneous crying and spontaneous laughter produced by 5 unfamiliar speakers. Performance was significantly impaired for pairs of spontaneous laughter and crying compared to vowels. Additionally, listeners failed to generalise identity related information across pairs including different types of nonverbal vocalisations (e.g. pairs of laughter and vowels) with performance indicating at times an inability to discriminate between speakers. Experiment 4 further assessed whether these effects may result from differences in arousal, vocalisation type or volitional control, using spontaneous laughter, volitional laughter and vowels. Performance was similar for vowels and volitional laughter, but impaired for spontaneous laughter. Results are discussed in the light of auditory expertise and prototypical representations of (volitional) vocal signals.

3.1 Experiment 3

3.1.1 Introduction

Listeners are not only able to extract specific speaker characteristics, such as sex and age, from a person's voice. Individual voices differ from each other – the anatomy of a person's vocal tract and idiosyncratic features of voice use result in distinctive vocal outputs that allow listeners to discriminate between and (in the case of familiar speakers) recognise and identify individuals from their voices alone (see Kreiman & Sidtis, 2011; Mathias & von Kriegstein, 2013 for a review; see Section 1.5.3). In terms of identity processing from voices, Van Lancker and Kreiman (1987) propose that speaker discrimination and speaker recognition are separate, dissociable abilities based on a patient study showing selective impairments of both abilities. They suggest that (familiar) speaker recognition is underpinned by the processing of complex large-scale patterns of acoustic cues: Listeners perceive diagnostic and characteristic features in a familiar voice and thus recognise a person without close acoustic analysis of the voice. In contrast to this, (unfamiliar) speaker discrimination is thought to be based on performing close analysis of differences in a wide range of acoustic features between voices (see Kreiman & Sidtis, 2011 for an overview). The acoustic factors underlying listeners' ability to determine speaker identity from voices have not been established yet – while formant frequencies and Fo generally allow listeners to determine speaker sex from voices, no one acoustic feature has been found to be universally salient for speaker identification or discrimination across sets of speakers and listeners – salient acoustic cues have indeed been shown to vary across speakers and across listeners (Kreiman, Gerratt, Precoda & Berke, 1992).

Nonetheless, as is the case with speaker sex identification, listeners are readily able to discriminate between unknown voices with relatively high accuracy in the context of neutral speech stimuli (> 90% for healthy young adult listeners; Van Lancker & Kreiman, 1987; Reich & Duke, 1979; Wester, 2012). There is, however, evidence that variability in speech sounds, introduced by volitional changes to vocal signals, has a detrimental effect on speaker discrimination: Reich & Duke (1979) conducted a study in which listeners were asked to discriminate speakers based on pairs of sentences, with one sentence produced in undisguised voice and another produced in either an undisguised voice or a disguised voice (e.g. hoarse voice, extremely slow speech and nasal speech). Performance in their study was significantly better for pairs of undisguised sentences compared to pairs that included both disguised and undisguised sentences and thus more vocal variability. Similar findings are reported in studies of earwitness accuracy: listeners' ability to identify a voice from a line up decreases when vocal variability, in this case due to (volitional) emotional content, is introduced between study and test (Saslove & Yarmey, 1980; Read & Craik, 1995). It has furthermore been reported that listeners' accuracy in discriminating and recognising speakers is affected when listeners are asked to discriminate between speakers across speech samples from different languages produced by bilingual speakers (Perrachione et al., 2011; Winters et al., 2008). In sum, these results suggest that generalising identity-related cues across variable signals is challenging for listeners.

All of the studies reviewed above have used speech stimuli – which only form one (prominent) part of human vocal communication: other types of vocalisations, such as non-verbal vocal signals additionally permeate everyday interactions. It

should be noted that speech, especially when produced in a language familiar to the listener (see Goggin et al., 1991; Orena et al., 2015; Winters et al., 2008), is uniquely rich in cues to speaker characteristics and identity, including regional accent, lexical content and individual differences in pronunciation. Such speech-specific cues have been shown to be crucial for extraction of speaker characteristics and identity (e.g. Remez, Fellowes & Rubin, 1997) but are largely absent in non-verbal vocalisations. Using speech signals may thus have provided relatively favourable conditions for the extraction of speaker characteristics in previous studies, inflating stimulus discriminability and participant performance. For nonverbal vocal signals, it could be hypothesised that the detrimental effects of variability on speaker discrimination may be even more pronounced for non-verbal vocal signals given the lack of such speech-specific cues. Further, as has been shown in Experiments 1 and 2, the perceptual features of vocal signals produced under reduced volitional control differ from those of vocal signals produced under full volitional control. It could thus additionally be predicted that this should also result in impaired performance on a speaker discrimination task.

For the current study, participants performed a same-different speaker discrimination task on pairs of nonverbal vocalisations, including 3 within-vocalisation conditions (Vowels-Vowels, $Crying_S$-$Crying_S$, $Laughter_S$-$Laughter_S$) and 3 across-vocalisation conditions ($Crying_S$-$Laughter_S$, $Crying_S$-Vowels, and $Laughter_S$-Vowels). Based on the findings of Experiments 1 and 2 and studies exploring how vocal flexibility affects generalisation (e.g Reich & duke, 1979, Saslove & Yarmey, 1980; Read & Craik, 1995), it was predicted that natural flexibility of vocal signals (introduced by manipulating the presence or absence of authentic emotional states

and thus different levels of volitional control over voice production) would affect the perception of person identity in unfamiliar voices. Specifically, it was predicted that a) speaker discrimination performance would be better for within-vocalisation trials compared to across-vocalisation trials, b) this impairment would be more marked for within-pair mismatches in volitional control (i.e. when one vocalisation is produced under full control while the other is not) and c) performance would generally be impaired for pairs of vocalisations produced under reduced volitional control (see Experiments 1 and 2). Specific predictions per condition, based on stepwise decreases in accuracy in the presence of these three factors, are illustrated in Figure 10.

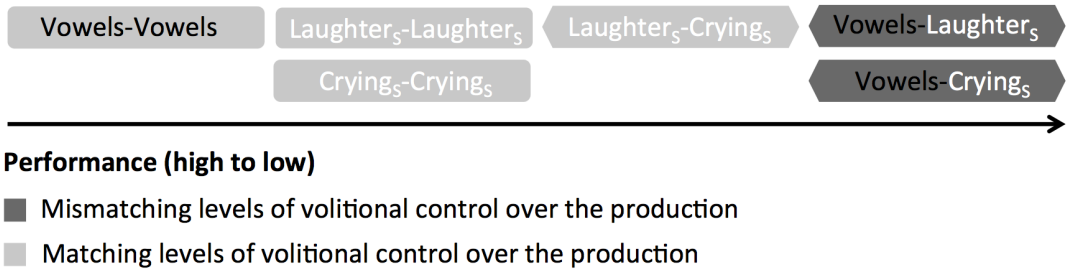

Figure 10 Predicted pattern for performance on the speaker discrimination task (from high performance to low performance). Boxes with rounded edges represent within-vocalisation pairs, hexagons represent across-vocalisation pairs. Black text: vocalisations produced under full volitional control; white text: vocalisations produced under reduced volitional control. Specific predictions follow the pattern Vowels-Vowels (full volitional control, within-vocalisation, matching levels of volitional control) > Crying$_S$-Crying$_S$ (reduced volitional control, within-vocalisation, matching levels of volitional control) = Laughter$_S$-Laughter$_S$ (reduced volitional control, within-vocalisation, matching levels of volitional control) > Crying$_S$-Laughter$_S$ (reduced volitional control, across-vocalisation, matching levels of volitional control) > Crying$_S$-Vowels (reduced volitional control, across-vocalisation, mismatching emotional content) = Laughter$_S$-Vowels (reduced volitional control, across-vocalisation, mismatching levels of volitional control).

3.1.2 Participants

Participants were the same as in Experiment 1 (see Section 2.1.2). Average performance across conditions for each participants was within 2 standard deviations

from the mean and therefore all participants were included in the following analyses.

3.1.3 Materials

Stimuli used were identical to the ones used Experiment 1 (see Section 2.1.3): 25 stimuli of $Laughter_S$, $Crying_S$ and Vowels each, presented in pairs.

3.1.4 Design and Procedure

After hearing all stimuli once in a brief speaker sex identification task (see Experiment 1), participants performed a speaker discrimination task. Participants heard permutations of pairs of $Laughter_S$, $Crying_S$ and Vowels, with the two sounds being presented sequentially with a pause of 0.7 seconds between them. For each of the 6 conditions (3 within-vocalisation conditions [Vowels-Vowels, $Crying_S$-$Crying_S$, $Laughter_S$-$Laughter_S$] and 3 across-vocalisation conditions [$Crying_S$-$Laughter_S$, $Crying_S$-Vowels, and $Laughter_S$-Vowels]), there were 50 trials, with 25 trials including two vocal signals from the same speaker and 25 trials presenting two sounds from different speakers – this yielded 300 trials in total. With the inclusion of across-vocalisation conditions the hypotheses regarding listener's ability to generalise identity information could be tested. With the inclusion of within-vocalisation conditions vocalisation type- and emotion effects could be probed. 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 $Crying_S$-$Laughter_S$ trials, half began with a laughter stimulus and half began with crying. Speaker pairings were fixed across participants. After the presentation of the sounds, participants were prompted 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. Key presses and reaction times were recorded. For this study, reaction times were not analysed since stimuli were presented in pairs, which could have introduced confounds. The task lasted for approximately 35 minutes.

3.1.5 Results

Speaker discrimination from non-verbal vocalisations

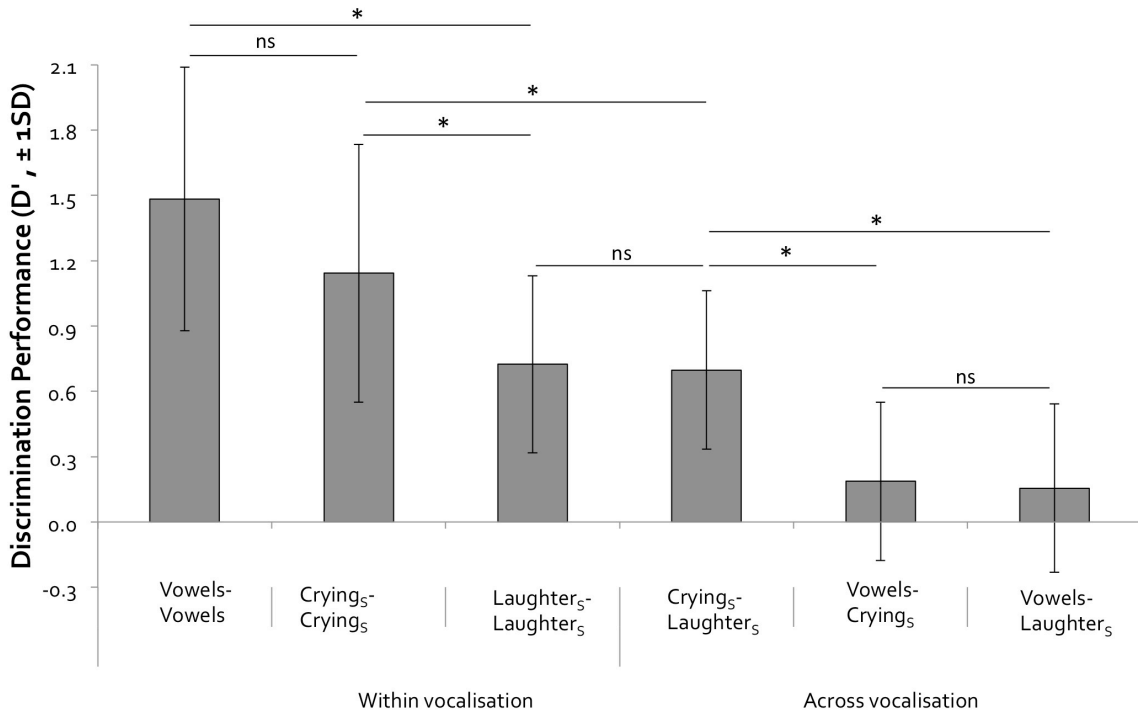

Figure 11 Average d' scores per condition for the speaker discrimination task. Significant comparisons (Bonferroni-corrected, , α = .008) are highlighted with an asterisk.

D' scores were calculated from the raw responses as described in Experiments 1 (see Section 2.1.5). The *d'* values in the current study ranged from 1.48 to 0.15, which corresponds to average accuracy values of 75% - 53% (chance level = 50%). *D'* scores were entered into a one-way repeated measures ANOVA with 6 levels for the different conditions. There was a significant effect of condition on the *d'* scores ($F[5,210] = 74.01$, $p < .001$, $\eta_p^2 = .63$). To further explore condition-specific effects, 8

pairwise post-hoc t-tests (Bonferroni corrected, α = .006) were conducted to assess the prediction of a stepwise decrease in performance, introduced by 1) across-vocalisation judgements, 2) the presence of authentic emotional content and 3) a mismatch in authentic emotional content within a pair (leading to differences in control over voice production and concomitant acoustic changes), see **Figure 11**.

T-tests confirmed these predictions, with the exception of the comparisons of $Crying_S$-$Crying_S$ versus $Laughter_S$-$Laughter_S$ (t[44] = 4.5, p < .001, Cohen's d = 1.356; prediction: ns), and Laughter-Laughter versus $Laughter_S$-$Crying_S$ (t[44] = .616, p = .54, Cohen's d = .185; where a significant difference had been predicted; **Figure 11**). Performance for speaker identification in both $Laughter_S$-Vowels and $Crying_S$-Vowels was close to zero ($d'_{LaughterS\text{-}Vowels}$= .15, $d'_{CryingS\text{-}Vowels}$ = .19), suggesting very low discriminability for these across-vocalisation pairs. One-sample t-tests (against 0) revealed that listeners were able to discriminate between speakers for all conditions except $Laughter_S$-Vowels (α = .008, Bonferroni corrected, t[44] = 2.65, p = .011; all other ps ≤ .001).

To directly assess whether performance for speaker discrimination was higher for within-vocalisation trials compared to across-vocalisation trials, the scores for the three conditions including within-vocalisation trials were averaged and compared to the averaged across-vocalisation trials. A paired-samples t-test showed that, as predicted, participants were better at discriminating speakers for within-vocalisation trials compared to across vocalisation trials (t[44]= 12.455, p < .001, Cohen's d = 3.755).

A response bias analysis was run to further explore the underlying processes for different trial types: C was calculated as described in Experiments 1 and 2. The values for every condition were entered into one sample t-tests (testing against 0), to

determine whether any biases observed were significant. There were trends towards responses being significantly biased towards 'same' for within vocalisation trials (all *ps* < .024*)*. For the across-vocalisation trials no bias was found for Vowels paired with $Laughter_S$ or $Crying_S$ (*ps* > .89) while a trend level bias towards 'same' was found for $Laugher_S$-$Crying_S$ (t[44] = 2.427, p = .02, Cohen's *d* = .749). These results may thus suggest that greater within-pair similarity in vocalisation 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).

Contributions of acoustic properties to speaker recognition performance

Table 3 Absolute difference scores averaged per condition. Heatmaps per acoustic feature were overlayed onto the means (green = lowest value, red = highest values, yellow = intermediate values), with green indicating a relatively smaller average difference within pairs, red highlighting a relatively larger difference.

		Vowels-Vowels	**$Crying_S$-$Crying_S$**	**$Laughter_S$-$Laughter_S$**	**$Crying_S$-$Laughter_S$**	**Vowels-$Crying_S$**	**Vowels-$Laughter_S$**
Total Duration	Mean	0.28	0.30	0.34	0.33	0.36	0.43
	SD	0.21	0.28	0.28	0.28	0.25	0.26
% of Unvoiced Segments	Mean	25.71	29.72	23.27	25.60	27.51	26.03
	SD	17.32	20.15	18.13	18.36	19.32	18.03
Burst Durations	Mean	0.44	0.53	0.42	0.45	0.47	0.42
	SD	0.28	0.35	0.33	0.32	0.33	0.30
Fo Mean	Mean	52.36	108.88	130.61	121.19	250.87	283.77
	SD	44.93	81.54	86.60	90.94	100.98	114.06
Fo SD	Mean	57.53	47.86	41.18	44.82	64.69	63.90
	SD	40.79	36.02	37.66	37.78	50.01	54.12
Spectral Centre of Gravity	Mean	335.77	135.88	325.73	326.98	275.54	371.18
	SD	355.54	181.52	365.93	414.35	334.66	545.66
HNR	Mean	6.04	7.70	6.11	6.53	6.74	6.08
	SD	5.04	4.92	5.61	5.21	5.05	5.03
Jitter	Mean	1.35	1.74	1.69	1.62	1.49	1.50
	SD	0.83	1.23	1.27	1.25	1.05	1.06
Shimmer	Mean	0.45	0.53	0.43	0.45	0.47	0.43
	SD	0.28	0.35	0.33	0.32	0.33	0.30

To further explore whether acoustic factors may underlie these biases, per-trial averages for participant's responses as well as absolute difference scores of acoustic measures for each stimulus pair for each trial across all participants were calculated. Table 3 shows a breakdown of acoustic differences within pairs by condition. No overall pattern of small or large differences in acoustic measures by condition emerges from the heat maps. It is, however, apparent that the range of means across conditions can be small for some measures (e.g. shimmer, total duration) while the standard deviations are large, indicating that acoustic differences within conditions were very variable, at times exceeding differences in mean differences across conditions.

To formally assess whether variability in acoustic differences can predict participants' responses, hierarchical logistic regressions were run using the Enter method, with the binary same/different response as a dependent variable. Participant was included as a covariate in a first block, to account for any potential subject effects. Absolute difference scores for all acoustic measures were included as covariates in the second and final block. These analyses were run across all pairs and within condition. Participant explained a significant amount of variance for most of the models (see Table 4). For all 7 models (6 condition-specific models, 1 overall), the model fit, after partialling out the variance explained by participant, was significant with a range of acoustic parameters predicting responses (see Table 4). It should be noted, however, that Nagelkerke's R^2 is low for all models, with the notable exception of Vowels-Vowels (Nagelkerke's R^2 = .322). This indicates that while trends may be apparent for most conditions in this large data set, the overall model fit is poor and very little

variance is explained by these acoustic measures (and by participant effects) may be perceptually meaningless.

Table 4 Results of the second block of the logistic regression models. Results for the first block (including only participant) are omitted. Significant *p* values are highlighted in bold and significant covariates are highlighted in light grey.

		β	S.E.	Wald	p	Odds
Overall	Participant	-0.01	0	53.67	**< .001**	0.99
***N* = 12993**	Total Duration	0.08	0.07	1.13	0.289	1.08
Block χ^2 = 916.257, *p* < 0.001	% of Unvoiced Segments	0	0	2.03	0.154	1
Nagelkerke R^2= .095	Burst Duration	0.83	0.29	8.2	**0.004**	2.3
	Fo Mean	0	0	642.19	**< .001**	1
	Fo SD	0	0	2.37	0.124	1
	Spectral Centre of Gravity	0	0	63.07	**< .001**	1
	HNR	0.01	0.01	2.08	0.149	1.01
	Jitter	-0.05	0.02	4.96	**0.026**	0.95
	Shimmer	-0.81	0.3	7.35	**0.007**	0.45
Vowels-Vowels	Participant	-0.01	0	6.67	**0.01**	0.99
***N* = 2177**	Total Duration	-0.68	0.24	8.14	**0.004**	0.51
Block χ^2 = 593.163, *p* < 0.001	% of Unvoiced Segments	0	0	0.56	0.454	1
Nagelkerke R^2= .322	Burst Duration	1.19	0.73	2.65	0.103	3.29
	Fo Mean	-0.03	0	307.34	**< .001**	0.97
	Fo SD	0.01	0	20.45	**< .001**	1.01
	Spectral Centre of Gravity	0	0	32.71	**< .001**	1
	HNR	0.07	0.01	21.91	**< .001**	1.07
	Jitter	0.16	0.09	3.33	0.068	1.17
	Shimmer	-2.11	0.77	7.53	**0.006**	0.12
$Crying_S$-$Crying_S$	Participant	0	0	1.04	0.308	1
***N* = 2170**	Total Duration	0.77	0.19	15.76	**< .001**	2.15
Block χ^2 = 172.504, *p* < 0.001	% of Unvoiced Segments	0.01	0	3.4	0.065	1.01
Nagelkerke R^2= .11	Burst Duration	0.9	0.61	2.19	0.139	2.46
	Fo Mean	-0.01	0	102.37	**< .001**	0.99
	Fo SD	0	0	0	0.961	1
	Spectral Centre of Gravity	0	0	17.53	**< .001**	1
	HNR	0.03	0.02	4.75	0.029	1.04
	Jitter	-0.13	0.06	5.51	0.019	0.88
	Shimmer	-0.67	0.62	1.14	0.285	0.51

		β	S.E.	Wald	p	Odds
Laughter$_S$-Laughter$_S$	Participant	-0.01	0	1.75	0.186	1
N = 2160	Total Duration	-0.34	0.17	4	**0.045**	0.71
Block χ^2 = 98.114, p < 0.001	% of Unvoiced Segments	0	0	0.06	0.808	1
Nagelkerke R^2= .061	Burst Duration	-1	0.81	1.5	0.221	0.37
	Fo Mean	0	0	47	**< .001**	1
	Fo SD	0	0	9.66	**0.002**	1
	Spectral Centre of Gravity	0	0	4.84	**0.028**	1
	HNR	0.03	0.01	4.02	**0.045**	1.03
	Jitter	-0.19	0.06	11.57	**0.001**	0.83
	Shimmer	1.06	0.84	1.58	0.209	2.89
Crying$_S$-Laughter$_S$	Participant	-0.01	0	4.04	**0.045**	0.99
N = 2158	Total Duration	0.59	0.17	12.49	**< .001**	1.8
Block χ^2 = 97.266, p = 0.061	% of Unvoiced Segments	0	0	0.02	0.887	1
Nagelkerke R^2= .012	Burst Duration	0.51	0.73	0.49	0.486	1.66
	Fo Mean	-0.01	0	72.49	**< .001**	1
	Fo SD	0	0	0.01	0.925	1
	Spectral Centre of Gravity	0	0	3.14	0.076	1
	HNR	-0.01	0.01	0.69	0.406	0.99
	Jitter	-0.08	0.05	2.46	0.117	0.93
	Shimmer	-0.19	0.74	0.07	0.794	0.83
Vowels-Crying$_S$	Participant	-0.02	0	21.27	**< .001**	0.98
N = 2162	Total Duration	-0.17	0.2	0.72	0.396	0.84
Block χ2 = 39.679, p < 0.001	% of Unvoiced Segments	0	0	0.47	0.493	1
Nagelkerke R^2= .036	Burst Duration	3.41	1.19	8.17	**0.004**	30.24
	Fo Mean	0	0	21.55	**< .001**	1
	Fo SD	0	0	1.05	0.305	1
	Spectral Centre of Gravity	0	0	0.92	0.339	1
	HNR	0.02	0.01	2.53	0.112	1.02
	Jitter	0.04	0.06	0.39	0.534	1.04
	Shimmer	-3.85	1.22	10.02	**0.002**	0.02
Vowels-Laughter$_S$	Participant	-0.02	0	42.71	**< .001**	0.98
N = 2166	Total Duration	0.03	0.19	0.02	0.892	1.03
Block χ^2 = 29.414, p = 0.001	% of Unvoiced Segments	0	0	0.23	0.63	1
Nagelkerke R^2= .045	Burst Duration	1.34	0.89	2.24	0.135	3.8
	Fo Mean	0	0	3.07	0.08	1
	Fo SD	0	0	0.42	0.518	1
	Spectral Centre of Gravity	0	0	3.44	0.064	1
	HNR	0	0.01	0.02	0.886	1

Jitter	-0.27	0.07	15.11	**< .001**	0.77
Shimmer	-0.87	0.91	0.93	0.336	0.42

3.1.6 Discussion

Previous research using only speech vocalisations reported high probabilities of correct responses for speaker discrimination tasks (> 90% for healthy young adults; Van Lancker & Kreiman, 1987; Reich & Duke, 1979; Wester, 2012). The current results, however, indicate that impairments in the ability to discriminate between speakers can emerge when listeners are presented with vocal signals produced under reduced volitional control, when they encounter within-pair mismatches in volitional control and when they are required to perform across-vocalisation judgements. Performance was highest for Vowels-Vowels – a volitionally produced, within-vocalisation conditions, while performance was significantly impaired for $Laughter_S$-$Laughter_S$ compared to $Crying_S$-$Crying_S$ (but was equivalent for the $Laughter_S$-$Laughter_S$ and $Laughter_S$-$Crying_S$ conditions). As in Experiment 1, this could point towards modulation of voice perception through higher arousal – and concomitant acoustic variability (Ruch & Ekman, 2001) – in laughter compared to crying. Similarly, basic differences across vocalisation types and the inherent variability within laughter vocalisations (e.g. noisy breathing, wheezing, vocal bursts, snorts, etc., see Bachorowski & Owren, 2001) could also underlie the lower performance for $Laughter_S$-$Laughter_S$. It should also be noted that male and female speakers were present in the current stimulus. This may have inflated performance, as male and female voices can be easily distinguished from another (in neutral speech – e.g. Owren et al., 2007) allowing participants to base their judgement on basic speaker sex

discrimination and not on holistic identity discrimination – this issue was addressed by using an all-female stimulus set in Experiment 5.

There were response biases towards 'same'-responses for within-vocalisation conditions while listeners more frequently perceived two sounds to come from different speakers in across-vocalisation conditions. In a per-trial analysis, responses were shown to be linked to the degree of within-pair acoustic difference on a range of measures. In the presence of great variability in acoustic differences within condition, however, very little of the variance in discrimination accuracy could be explained by these acoustic difference measures, making it difficult to interpret the role of individual acoustic features in a perceptually meaningful way.

The striking effects of the different combinations of vocalisation pairs on participants' task performance point towards listeners' limited ability to generalise the markers of identity-related information in the presence of natural and meaningful variability (introduced here by differences in volitional control over voice production) across different vocal signals from unfamiliar individuals. For some of the conditions, participants were not able to discriminate between speakers (as indicated by a *d'* score that is very close to or no different from zero). This is in line with previous findings, showing that variability in vocal signals introduced by volitional voice changes impairs performance on speaker recognition, identification and discrimination tasks (Read & Craik, 1995; Reich & Duke, 1979; Saslove & Yarmey, 1980; Winters et al., 2008). There is furthermore a body of research that has shown that by manipulating specific acoustic properties of a vocal signal and thus introducing variability, the processing of identity-related information can be harmed (see Kreiman & Sidtis, 2011 for an overview). The primary aim of studies using acoustic

manipulations of stimuli was to identify sets of salient acoustic features used by listeners to make inferences about a speaker; they can, however, also be interpreted as evidence for a lack of generalisation across variability in vocal signals: For successful generalisation, the effect of manipulations on one parameter would need to 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, Gath & Rosenhouse, 2000; Sell, Suied, Elhilali & Shamma, 2015).

In parallel to the interpretation of the results of Experiment 1, the lower discriminability apparent for vocalisations produced under reduced volitional control could be explained in different ways, which are not mutually exclusive: 1) The acoustic changes in vocal signals present in spontaneous vocalisations may partly override cues to speaker identity; 2) The authentic emotional content for spontaneous vocalisations is automatically processed in preference to speaker identity-related information; 3) Listeners are less familiar with these spontaneous signals[3], as they occur relatively rarely compared to volitional signals, which could mean that the observed impairment in speaker discrimination is an expertise effect. In general, current results indicate that several factors have detrimental effects on the listener's ability to extract indexical speaker properties from vocalisations. As was the case in Experiment 1, which used

[3] It may seem intuitive that the high-intensity spontaneous vocalisations that are being investigated as part of this thesis are produced relatively infrequently in every day life (compared to lower intensity version of these signals, speech or other vocalisations entirely). There is, however, no empirical data on the frequency of spontaneous versus volitional laughter production. A study of laughter production by Vettin and Todt (2004), reports an average of 6 bouts of laughter during 10 minutes of conversation. Self-report studies, do, however, report a lower rate of laughter (e.g. an average of 17 laughs per day has been reported in self-report studies [Martin & Kuiper, 1999], 13.4 laughter bouts per day [Mannell & McMohan, 1982]). Notably, none of these studies discriminate between different laughter types or intensities, nor do they relate laughter production frequency to the production frequency of other vocal signals.

the same stimuli for a speaker sex identification task, it cannot yet be determined which specific properties of the emotional vocalisations are exerting this detrimental effect on performance: the effects could be due to differences in vocalisation type, with emotional vocalisations being processed in a different way to vowel sounds during the decoding of speaker identity. Second, laughter and crying stimuli in this stimulus set differ in the perceived arousal and authenticity of emotional content ratings – these differences in arousal could thus be a confound for the current results. In a second speaker discrimination experiment, these issues were therefore addressed by introducing relevant contrasts in affective properties within the same vocalisation category, through the use of Laughter_S and Laughter_V.

3.2 Experiment 4

3.2.1 Introduction

In parallel to Experiment 2, the speaker discrimination task from Experiment 3 was replicated to further explore whether the effects observed in Experiment 3 were a result of reduced volitional control over the production of the spontaneous vocalisations, due to differences in arousal between conditions, or whether the effects were indicative of general differences in the processing of different vocalisation types regardless of levels of volitional control. In order to address these questions, Vowels, $Laughter_S$ and $Laughter_V$ were used in the current experiment. $Laughter_V$ 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 (see Experiment 2, Section 2.2.3 for a more detailed description of the stimuli). Crucially for this study, arousal was more closely matched for $Laughter_V$ and $Laughter_S$ compared to $Laughter_S$ and $Crying_S$ in the previous experiment. With these stimuli, differences in volitional control could be directly contrasted within the same vocalisation for stimuli that were more closely matched in arousal but that differed in terms of volitional control over production. In line with the predictions of Experiment 3, it was hypothesised that performance for across-vocalisation trials ($Laughter_V$-Vowels, $Laughter_S$-Vowels) would be less accurate than for within-vocalisation trials (Vowels-Vowels, $Laughter_V$-$Laughter_V$, $Laughter_S$-$Laughter_S$, $Laughter_V$-$Laughter_S$), that a mismatch in volitional control for vocalisations within a pair and the presence of vocalisation produced under reduced volitional control would further impair performance. Specific predictions are illustrated in **Figure 12**.

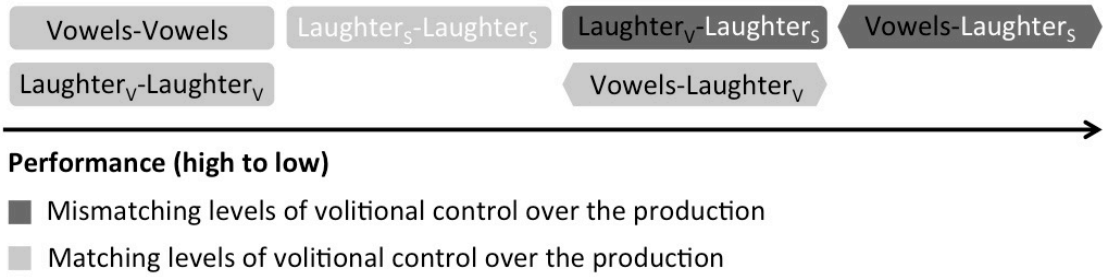

Figure 12 Predicted pattern for performance on the speaker discrimination task (from high performance to low performance). Boxes with rounded edges represent within-vocalisation pairs, hexagons represent across-vocalisation pairs. Black text: vocalisations produced under full volitional control; white text: vocalisations produced under reduced volitional control. Specific predictions follow the pattern Vowels-Vowels (full volitional control, within-vocalisation, matching levels of volitional control) = Laughter$_V$-Laughter$_V$ (full volitional control, within-vocalisation, matching levels of volitional control) > Laughter$_S$-Laughter$_S$ (reduced volitional control, within-vocalisation, matching levels of volitional control) > Laughter$_V$-Laughter$_S$ (reduced volitional control, within-vocalisation, mismatching levels of volitional control) = Laughter$_V$-Vowels (full volitional control, across-vocalisation, mismatching emotional content) > Laughter$_S$-Vowels (reduced volitional control, across-vocalisation, mismatching levels of volitional control).

3.2.2 Participants

Participants were the same as in Experiment 2 (see Section 2.2.2). Average performance across conditions for each participants was within 2 standard deviations from the mean and therefore no participant was excluded in the following statistical analyses.

3.2.3 Materials

Stimuli used were identical to the ones used Experiment 2 (see Section 2.2.3): 25 stimuli of Laughter$_S$, Laughter$_V$ and Vowels (including 5 stimuli from 5 different talkers) were selected.

3.2.4 Design and Procedure

The task was the same as the one used in Experiment 3 (see Section 3.1.4). 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-vocalisation conditions (Vowels-Vowels, Laughter$_V$-Laughter$_V$, Laughter$_S$-Laughter$_S$, Laughter$_V$-Laughter$_S$) and 2 across-vocalisation conditions (Laughter$_V$-Vowels, Laughter$_S$-Vowels). Participants were not pre-informed about the inclusion of Laughter$_S$ and Laughter$_V$ in the tasks. No stimuli were repeated during the task. Further in contrast to Experiment 3, the pairs of speakers was not fixed but randomised across participants in the current experiment.

3.2.5 Results

Speaker discrimination from non-verbal vocalisations

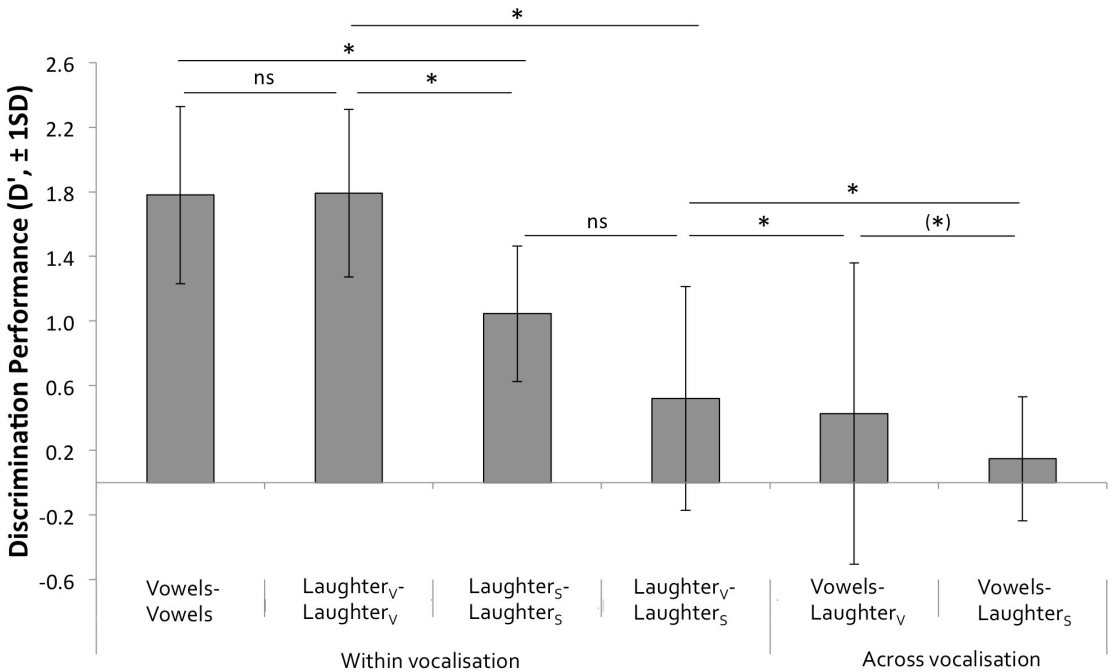

Figure 13 Average *d'* scores per condition for the speaker discrimination task. Significant comparisons (Bonferroni-corrected, see Results for alpha levels) are highlighted with an asterisk; marginally significant results are highlighted with an asterisk in brackets.

D' scores were calculated from the raw responses (for details see Experiment 1 and 2, see Section 2.1.5). *D*' values ranged from 1.78 to 0.15, which corresponds to average accuracy values of 78% - 53% (chance level = 50%). The *d*' scores were entered into a one-way repeated measures ANOVA with 6 levels for condition. Results are shown in Figure 13. There was a significant effect of condition on the *d'* scores ($F[5,220] = 61.12$, $p < .001$, $\eta_p^2 = .59$). Post-hoc t-tests (8 comparisons, corrected alpha = .006) tested for the predicted pattern illustrated in **Figure 12**. Predictions were confirmed for all within-vocalisation judgements with performance $Laughter_S$-$Laughter_S$ being significantly lower than for $Laughter_V$-$Laughter_V$ and Vowels-Vowels (*ps* < .001), while performance for $Laughter_V$-$Laughter_V$ and Vowels-Vowels was similar. Following the predictions, performance for $Laughter_V$-$Laughter_S$ was also significantly lower compared to $Laughter_S$-$Laughter_S$ ($t[42] = 10.706$, $p < .001$, Cohen's $d = 3.304$). As expected, performance for $Laughter_V$-$Laughter_S$ and Vowels-$Laughter_V$ were similar ($t[42] = 10.706$, $p = .535$, Cohen's $d = .193$). There was, however, only a marginally significant difference between Vowels-$Laughter_V$ and Vowels-$Laughter_S$ ($t[42] = 1.842$, $p = .073$, Cohen's $d = .568$). Overall, there was a steep decline in performance across the conditions, with performance being not significantly different from zero for Vowels-$Laughter_S$ (one-sample t-test, against zero: $t[42] = 2.521$, $p = .016$; Bonferroni-corrected $\alpha = .008$; all other *p*s ≤ .004), indicating an inability of participants to discriminate signal from noise in this condition (see Figure 13).

To directly assess whether speaker discrimination was more accurate for within-vocalisation trials compared to across-vocalisation trials, the scores for the four within-vocalisation conditions were averaged and compared to the averaged scores for the two across-vocalisation conditions. Participants performed better at

discriminating speakers for within-vocalisation trials compared to across-vocalisation trials (t[43]= 12.83, p < .001, Cohen's d = 3.959).

A response bias analysis using C as a measure was run to further explore the underlying processes for different trial types (see Experiment 3, Section 3.1.5). C values per condition were entered into a one sample t-test (testing against 0), to determine whether any biases observed were significant. In line with findings from Experiment 3, this showed that for all within-vocalisation conditions, with the exception of $Laughter_V$-$Laughter_S$, there was a significant bias towards responding 'same' (all ps < .001). In contrast to Experiment 3, where either a positive or no bias was found for the across-vocalisation trials, there was, however, a significant bias towards responding 'different' in this experiment for all across-vocalisation conditions as well as $Laughter_V$-$Laughter_S$, (all ps < .001). This suggests that again greater within-pair similarity in vocalisation type affected how responses were chosen for judgements of speaker identity (for similar effects of linguistic similarity on response bias, see Narayan et al., 2016).

Contribution of acoustic properties to speaker discrimination performance

To further explore whether acoustic factors may underlie these biases, per trial responses as well as absolute difference scores of acoustic measures for each stimulus pair for each trial across all participants. Table 5 shows a breakdown of acoustic differences by condition. From the heatmaps, it is apparent that across most acoustic properties, within-pair differences were higher for the across vocalisation conditions – this is in contrast to the results of this particular analysis conducted for Experiment 3, where no trend was apparent. This may tie in with the bias towards being more likely

to perceive speakers as being different for across-vocalisations pairs as well as the overall low performance for these conditions. As in Experiment 3, it is, however, again apparent that standard deviations for acoustic measures are large, indicating that acoustic differences within conditions were very variable, at times exceeding differences in mean differences across conditions.

Table 5 Absolute difference scores averaged per condition. Heatmaps per acoustic feature were overlayed onto the means, with green indicating a relatively average smaller difference within pairs, red highlighting a relatively larger difference (green = lowest value, red = highest values, yellow = intermediate values), with green indicating a relatively smaller average difference within pairs, red highlighting a relatively larger difference.

		Vowels-Vowels	$Laughter_V$-$Laughter_V$	$Laughter_S$-$Laughter_S$	$Laughter_V$-$Laughter_S$	Vowels-$Laughter_V$	Vowels-$Laughter_S$
Total Duration	Mean	0.28	0.59	0.34	0.51	0.48	0.43
	SD	0.21	0.62	0.28	0.55	0.47	0.26
% of Unvoiced Segments	Mean	10.96	13.14	15.88	17.33	35.64	36.79
	SD	10.58	12.75	12.32	12.55	17.51	18.34
Burst Duration	Mean	0.23	0.38	0.26	0.34	0.81	0.58
	SD	0.18	0.37	0.19	0.29	0.30	0.26
Fo Mean	Mean	52.25	97.38	129.19	190.43	128.93	283.74
	SD	44.59	91.01	86.96	127.88	78.88	113.77
Fo SD	Mean	57.05	61.89	40.98	56.78	72.89	60.87
	SD	40.27	44.95	31.15	41.47	47.13	39.91
Spect. Centre of Gravity	Mean	356.77	334.02	368.07	367.58	442.93	541.00
	SD	336.09	276.69	323.25	305.31	297.71	370.13
HNR	Mean	4.89	2.24	3.05	4.23	12.50	8.72
	SD	3.24	1.92	2.16	2.82	5.04	4.78
Jitter	Mean	0.51	0.81	1.14	1.30	2.67	1.78
	SD	0.42	0.98	0.91	0.86	0.72	1.01
Shimmer	Mean	0.23	0.38	0.26	0.34	0.81	0.58
	SD	0.18	0.37	0.19	0.29	0.30	0.26

To formally assess whether acoustic differences can predict participants' responses, hierarchical logistic regressions with the binary same/different responses of all participants as a dependent variable, using the Enter method. Participant was included as a covariate in a first block to account for any potential subject effects. Absolute acoustic difference scores for all acoustic measures were included as

covariates in the second and final block. These analyses were run across all conditions (full model), and within each condition separately. With the exception of the models for $Laughter_S$-$Laughter_V$ (p = .027) and $Laughter_S$-$Laughter_S$ (p = .012, see Table 6), none of the other models showed a significant effect of participant. For all 7 models (6 conditions, 1 overall), the model fit was highly significant with a range of acoustic parameters predicted responses (see Table 6). It should be noted, however, be noted that Nagelkerke's R^2 is low across all models, indicating that while trends may be apparent, the overall model fits are relatively poor and very little variance is explained by these acoustic measures (and by participant effects). This may be linked to each measure being in itself vary variable – see the large standard deviations. It is unclear why relatively little variance is explained for Vowels-Vowels (2.6%) compared to Experiment 3 (32.2%), despite the same stimuli being used for both experiments.

Table 6 Results of the second block of the logistic regression models. Results for the first block (including only participant) are omitted. Significant *p* values are highlighted in bold and significant covariates are highlighted in light grey.

		β	S.E.	Wald	p	Odds
Overall	**Participant**	0	0	1.16	0.281	1
***N* = 12795**	Total Duration	0.37	0.05	63.52	**< .001**	1.44
Block χ^2 = 192.534, p < 0.001	% of Unvoiced Segments	0	0	8.87	**< .001**	1
Nagelkerke R^2= .02	Burst Duration	0.3	0.09	10.61	**0.001**	1.34
	Fo Mean	0	0	6.74	**0.009**	1
	Fo SD	0	0	5.71	**0.017**	1
	Spectral Centre of Gravity	0	0	43.35	**< .001**	1
	HNR	0	0.01	0.3	0.582	1
	Jitter	0.08	0.03	7.79	**0.005**	1.08
Vowels-Vowels	Participant	0	0	0.34	0.56	1
***N* = 2136**	Total Duration	-0.14	0.24	0.32	0.571	0.87
Block χ^2 = 204.15, p < 0.001	% of Unvoiced Segments	0.01	0.01	2.99	0.084	1.01
Nagelkerke R^2= .026	Burst Duration	0.53	0.33	2.49	0.115	1.7
	Fo Mean	-0.02	0	91.25	**< .001**	0.99
	Fo SD	0.01	0	11.59	**0.001**	1.01
	Spectral Centre of Gravity	0	0	36.43	**< .001**	1
	HNR	-0.04	0.02	4.98	**0.026**	0.96
	Jitter	0.28	0.12	5.38	**0.02**	1.33

		β	S.E.	Wald	p	Odds
$Laughter_V$-$Laughter_V$	Participant	0	0	0.06	0.813	1
***N* = 2136**	Total Duration	0.91	0.14	43.56	**< .001**	2.48
Block χ^2 = 106.201, *p* < 0.001	% of Unvoiced Segments	-0.04	0.01	35.12	**< .001**	0.96
Nagelkerke R^2= .02	Burst Duration	-0.2	0.24	0.68	0.41	0.82
	Fo Mean	0	0	1.81	0.179	1
	Fo SD	0	0	0.71	0.4	1
	Spectral Centre of Gravity	0	0	6.3	**0.012**	1
	HNR	-0.06	0.03	2.91	0.088	0.95
	Jitter	0.11	0.1	1.28	0.258	1.11
$Laughter_S$-$Laughter_S$	Participant	0.01	0	7.05	**0.008**	1.01
***N* = 2131**	Total Duration	0.27	0.17	2.52	0.113	1.31
Block χ^2 = 37.038, *p* < 0.001	% of Unvoiced Segments	-0.01	0	7.04	**0.008**	0.99
Nagelkerke R^2= .024	Burst Duration	0.29	0.38	0.57	0.45	1.33
	Fo Mean	0	0	17.23	**< .001**	1
	Fo SD	0	0	0.02	0.881	1
	Spectral Centre of Gravity	0	0	0.34	0.557	1
	HNR	0.03	0.03	1	0.316	1.03
	Jitter	-0.1	0.09	1.2	0.273	0.91
$Laughter_V$-$Laughter_S$	Participant	0.01	0	4.94	**0.026**	1.01
***N* = 2132**	Total Duration	-0.29	0.11	6.95	**0.008**	0.75
Block χ^2 = 19.241, *p* = 0.073	% of Unvoiced Segments	0.01	0.01	2.13	0.144	1.01
Nagelkerke R^2= .012	Burst Duration	0.2	0.25	0.65	0.422	1.22
	Fo Mean	0	0	0.01	0.934	1
	Fo SD	0	0	0.01	0.932	1
	Spectral Centre of Gravity	0	0	3.15	0.076	1
	HNR	0.04	0.03	2.82	0.093	1.04
	Jitter	-0.05	0.08	0.39	0.534	0.95
Vowels-$Laughter_V$	Participant	-0.01	0	1.91	0.167	1
***N* = 2136**	Total Duration	0.24	0.11	4.96	**0.026**	1.27
Block χ^2 = 29.369, *p* = 0.001	% of Unvoiced Segments	0	0	0.02	0.877	1
Nagelkerke R^2= .019	Burst Duration	-0.01	0.22	0	0.979	0.99
	Fo Mean	0	0	9.93	**0.002**	1
	Fo SD	0	0	4.92	**0.027**	1
	Spectral Centre of Gravity	0	0	4.36	**0.037**	1
	HNR	0	0.01	0	0.977	1
	Jitter	-0.1	0.09	1.27	0.261	0.91
Vowels-$Laughter_S$	Participant	0	0	0.04	0.84	1
***N* = 2128**	Total Duration	-0.32	0.18	3.28	0.07	0.73
Block χ^2 = 41.236, *p* < 0.001	% of Unvoiced Segments	0	0	2.2	0.138	1
Nagelkerke R^2= .026	Burst Duration	-0.44	0.27	2.64	0.104	0.64
	Fo Mean	0	0	1.12	0.289	1
	Fo SD	0	0	0.1	0.755	1
	Spectral Centre of Gravity	0	0	1.26	0.263	1
	HNR	0.05	0.01	14.81	**< .001**	1.05
	Jitter	-0.15	0.07	4.97	0.026	0.86

3.2.6 Discussion

By using spontaneous and volitional laughter, Experiment 4 explored whether the effects observed in Experiment 3 were due to processing differences between different categories of vocalisations, differences in arousal or whether they could have resulted from differences in volitional control over the voice. Performance was highest and similar for $Laughter_V$-$Laughter_V$ and Vowels-Vowels – showing that vocalisation type (laughter versus vowels) *per se* does not have an impact on performance for within-vocalisation trials. Furthermore, $Laughter_V$ is higher in arousal than Vowels but closely matched to $Laughter_S$ – if differences in arousal were driving these effects, performance should have been lower for $Laughter_V$ compared to Vowels and comparable to $Laughter_S$. For vocalisations produced under reduced volitional control (but in the absence of an across-vocalisation judgement or 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 mismatch in levels of volitional control) and Vowels-$Laughter_V$ (additional across-vocalisation judgement). Finally, performance was not significantly different from zero for Vowels-$Laughter_S$, where all three detrimental factors (presence of vocalisations produced under reduced volitional control, across-vocalisation judgement and mismatch in degree of volitional control) impaired performance.

Per trial responses were linked to the degree of acoustic difference between stimuli, across a range of measures – although, in line with the results of Experiment 3, in the presence of great variability in acoustic differences within condition, at times exceeding across-condition variability, very little of the variance in same-different judgements could be explained by these measures.

3.3 General discussion

Experiments 3 and 4 strikingly illustrate how vocal flexibility affects speaker discrimination from voices only. While listeners were able to successfully discriminate between speakers for vocal signals that were similar in their properties, they were unable to link two distinct vocalisation types when produced by the same speaker. That is, listeners failed to generalise identity related cues in voices across variable signals. Evidence from Experiments 3 and 4 further suggests an advantage in the processing of speaker characteristics for vocalisations produced under full volitional control. This is in line with findings from Experiments 1 and 2 where a similar effect was observed for the extraction of speaker sex. It should be noted that no formal analysis in perceptual distinctiveness of the individual voices was performed for the current set of speaker discrimination experiments, since individual listeners in the current study were only presented with a subset of all possible speaker and stimulus pairings – adequately assessing the distinctiveness of each speaker/stimulus within the context of a pair would require data from all participants on all possible pairings (see Baumann & Belin, 2010). Future studies should, however, explicitly explore how perceptual distinctiveness of different voices (and different vocalisations) interacts with vocal variability. As in Experiment 1 and 2, large individual differences in task performance are apparent in the data – future research should aim to further explore what the underlying causes and factors for such individual differences in speaker discrimination may be.

Volitional vocalisations form the vast majority of human communication, leading to greater exposure and expertise, while vocalisations produced under reduced

volitional control are not only comparatively infrequent but also diverge from volitional vocalisations in terms of production mechanisms and thus acoustic properties. It is thought that, during unfamiliar voice processing, vocal signals are analysed with reference to voice templates or representations based on population-wide averages (Kreiman & Sidtis, 2011; Latinus et al., 2013). Spontaneous vocalisations diverge from prototypical vocal signals (i.e. speech) in their acoustic, affective as well as perceptual properties and may thus not be well-represented within prototypical voice templates, thus generally impairing the extraction of speaker characteristics for such spontaneous signals (see also Experiment 1 and 2). Further, without robust representations that link such non-prototypical and highly idiosyncratic signals to the (volitional) vocal repertoire of a single person, generalisations of identity-related information across a range of variable vocal signals only seems to be possible to a limited extent, as indicated by poor performance in across-vocalisation judgements in these current studies. Only relatively intimate and personal familiarity with and exposure to a speaker's full vocal inventory may enable listeners to form sufficiently detailed and robust representations of a specific voice, including representations of non-prototypical vocal signals. Such detailed representation may then allow them to reliably extract speaker characteristics from all vocal signals. Whether familiarity with a speaker affords listeners an advantage in the processing of identity-related information in variable vocal signals is thus be explored in the next experiments.

4 Speaker discrimination in familiar and unfamiliar listeners

Experiments 5 and 6 explore the effects of familiarity with a voice on the extraction of identity relation from variable vocal signals. In parallel to Experiment 4, familiar and unfamiliar listeners performed a speaker discrimination task on volitional and spontaneous laughter as well as vowels. While familiarity afforded listeners a consistent advantage for speaker discriminations, there was no interaction between familiarity and condition. Performance was therefore impaired to similar extents for familiar and unfamiliar listener groups. In Experiment 6, speaker discrimination in familiar and unfamiliar listeners was explored based on whispered and voiced vocal signals of varying linguistic complexity. Better performance for familiar listeners was found, while additionally for this study the familiarity advantage was greater for voiced compared to whispered signals. Complex interactions of voicing, group and linguistic complexity were also apparent. Findings are discussed with reference to models of prototype-based voice processing, potential underlying mechanisms and representations of familiar and unfamiliar voice perception and different types and aspects of familiarity.

4.1 Experiment 5

4.1.1 Introduction

Experiments 3 and 4 showed that listeners' ability to discriminate between speakers from a range of vocalisations is drastically affected by the demand to perform across-vocalisation generalisations and further by the variability introduced by reduced volitional control during production. The listeners in the previous two studies were, however, unfamiliar with the voices they heard. Some models of voice processing propose different mechanisms for the processing of familiar and unfamiliar voices – thus predicting differences in performance as a function of familiarity. One such 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, et al., 2013, see Section 1.5.3.1). 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. When determining speaker identity from vocal signals, the prototypical templates used during unfamiliar voice processing may thus be underspecified with regard to individual voices and their flexibility, limiting the ability to form generalised percepts in the presence of dramatically different vocalisations and production states. When processing familiar voices, the detailed and person-specific representations can, however, potentially provide a better fit between representation and stimulus, which may thus facilitate accurate identity perception despite variability in vocal signals. This prediction has been to

some extent confirmed by studies finding familiar talker advantages for speech comprehension, where listeners performed more accurately in reporting the content of speech produced by familiar talkers (e.g. Nygaard & Pisoni, 1998). Evidence from the face perception literature seems to also support this representation–based model: studies report that assessments of identity information from photographs are more accurate from familiar than unfamiliar viewers (Bruce, Henderson, Newman & Burton, 2001; Jenkins, White, van Montfort & Burton, 2011; Ramon & Van Belle, 2016). Whether such familiarity advantages extend to the processing of identity-related information in voices has, however, not been directly tested to date.

Given the findings in Experiment 3 and 4 indicating that unfamiliar listeners largely fail to successfully generalise identity-related information across variable non-verbal vocalisations, the current study explores whether familiarity with a voice affects performance in a speaker discrimination task. The current study also attempted to address some of the remaining confounds present in Experiments 3 and 4: A new set of stimuli was recorded from female lecturers working in the Psychology department at Royal Holloway. Given the presence of male and female speakers in Experiment 3 and 4, performance may have been inflated as male and female voices can be easily distinguished from another (in neutral speech: e.g. Owren et al., 2007; see also Experiments 1 and 2) without having to assess broader identity-related properties of a voice. This issue was addressed through the use of an all-female speaker set in Experiment 5. Further, in Experiment 4, the two types of laughter marginally differed in arousal, as a closer matching on this variable was not possible. In the new set of stimuli for Experiment 5, this possible confound was addressed by

selecting sets of volitional and spontaneous laughter that did not significantly differ in arousal.

In the current experiment, a group of listeners familiar with these speakers (students and other members of the Psychology department) as well as an unfamiliar listener group (students from other departments) were tested in a study replicating the design of Experiment 4. Based on the previous research showing familiarity advantages across visual and auditory signals, better performance on speaker discrimination overall was predicted for the familiar listeners. Based on the hypothesis that familiar listeners should have a well-formed mental representation of the voices (Kreiman & Sidtis, 2011), it was further predicted that familiar listeners should demonstrate a greater ability to generalise identity-related information across vocalisations.

4.1.2 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. Twenty-three (16 female; M_{Age}: 31.7 years; *S*D: 10.1 years; range: 19-65 years) of the participants were familiar with the voices of the speakers (referred to as familiar listeners here) represented in the stimuli set by virtue of having been lectured by these individuals for between 12 and 28 hours in the preceding 2-3 terms (dependent on the time point of testing within the teaching term) as part of their degree course or having worked in the department for more than two years. Twenty-three unfamiliar participants (17 female; M_{Age}: 20.2 years; *S*D: 1.9 years; range: 19-27 years) 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 recognising individuals from their faces and voices. One participant from the unfamiliar group was excluded as their average performance in the speaker discrimination task (measured in *d'*) across all conditions was more than 2 standard deviations below the average performance of the group, thus indicating random responses. Note also that listeners groups differed in age and age range.

4.1.3 Materials

New stimuli were recorded for this experiment. The vocalisation types were identical to the ones used in Experiment 4: $Laughter_S$, $Laughter_V$ and Vowels. The sounds were recorded using the same elicitation procedure described in Experiments 1 and 2 (see Sections 2.1.3 and 2.2.3). Six talkers (all female, ages range from 29 – 42 years) were recorded in a sound-treated recording booth. All speakers were lecturers at the Department of Psychology at Royal Holloway and selected based on their exposure to a subgroup of undergraduate degree students at the department. 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 and were seated at a distance of about 50cm from the microphone to avoid that any movement associated with intense

laughter production 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, normalised for root-mean-square intensity in Praat (Boersma & Weenink, 2010). 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 the validation of the stimulus sets of the previous experiments, 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 then additionally rated the perceived authenticity of laughter on a 7-point Likert scale.

Based on the ratings of this pilot study, 25 stimuli were selected (5 per speaker) per vocalisation. One speaker was excluded at this point of the process to ensure that the stimuli were matched for arousal. 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*[48]= 12.114, *p* < .001) , Cohen's *d* = 3.497). $Laughter_S$ and $Laughter_V$ were significantly higher in arousal than Vowels ($Laughter_V$: *t*[48] = 28.503, p < .001, Cohen's *d* = 8.228; $Laughter_S$: *t*[48] = 28.396, *p* < .001, Cohen's *d* = 8.197), 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*[48]= .929, *p* = .360, Cohen's *d* = .268). All vocalisations were furthermore matched 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 can be found in Table 7.

			All		AJ		BL		CM		PD		LM	
Vocalisation	Acoustic Measure	Unit	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Vowels	Duration	secs	2.04	0.34	1.89	0.19	2.4	0.33	2.09	0.17	2.13	0.25	1.69	0.29
	Burst Duration (mean)	secs	0.33	0.1	0.36	0.02	0.3	0.05	0.25	0.06	0.44	0.13	0.3	0.09
	Unvoiced Segments	%	37.02	13.78	33.95	11.28	52.54	10.3	27.73	9.54	24.02	4.75	46.85	6.31
	Fo (mean)	Hz	223.48	32.26	199.6	10.07	257.39	39.55	227.82	32.26	216.56	33.76	216.02	9.75
	Fo (SD)	Hz	109.52	67.12	59.03	9.97	160.06	89.48	97.95	62.67	81.04	55.38	149.5	48.67
	Spectral center of gravity	Hz	540.22	216.86	529.88	190.04	630.6	288.16	602.35	279.84	529.65	191.74	408.61	104.87
	HNR	Hz	14.95	3.77	17.21	0.96	11.54	3.47	15.71	2.78	18.75	2.36	11.55	2.2
	Jitter	dB	1.56	0.74	1.44	0.68	1.63	0.48	1.31	0.47	1.01	0.41	2.41	0.9
	Shimmer	dB	0.68	0.24	0.58	0.06	0.92	0.21	0.55	0.03	0.41	0.1	0.93	0.15
$Laughter_v$	Duration	secs	1.85	0.54	2.01	0.62	2.01	0.56	1.74	0.43	2.19	0.51	1.32	0.13
	Burst Duration (mean)	secs	0.1	0.02	0.1	0.02	0.09	0.01	0.11	0.02	0.12	0.03	0.08	0.01
	Unvoiced Segments	%	68.21	9.52	77.56	6.61	70.63	4.83	64.98	10.6	60.41	8.28	67.44	9.47
	Fo (mean)	Hz	379.59	60.56	386.79	79.13	374.02	30.65	421.65	77.3	363.12	56.79	352.37	45.02
	Fo (SD)	Hz	116.5	54.41	188.11	47.82	99.64	42.61	98.98	57.21	86.62	17.36	109.17	41.84
	Spectral center of gravity	Hz	1029.08	268.53	938.64	70.12	1225.83	233.06	1344.16	165.28	836.03	70.29	800.72	216.06
	HNR	Hz	6.76	1.69	5.08	1.14	5.46	0.39	7.28	1.37	8.72	1.54	7.29	0.57
	Jitter	dB	3.34	0.64	3.64	0.47	3.48	0.38	2.76	0.79	3.19	0.61	3.65	0.65
	Shimmer	dB	1.18	0.18	1.44	0.18	1.13	0.1	1.11	0.05	1.01	0.13	1.21	0.04
$Laughter_s$	Duration	secs	1.94	0.59	2.28	0.5	1.83	0.68	1.8	0.59	1.72	0.64	2.09	0.6
	Burst Duration (mean)	secs	0.1	0.05	0.08	0.01	0.08	0.01	0.07	0.01	0.14	0.05	0.15	0.06
	Unvoiced Segments	%	68.77	10.69	79.01	4	70.6	8.23	68.22	8.49	53.97	3.78	72.05	9.61
	Fo (mean)	Hz	476.28	159.97	462.16	129.33	502.63	69.1	309.81	48.63	473.26	84.19	633.54	236.45
	Fo (SD)	Hz	128.39	75.58	218.39	62.6	149.97	41.32	38.95	21.73	95.03	25.51	139.63	72.56
	Spectral center of gravity	Hz	1035.61	357.68	939.64	197.71	1230.5	121.66	935.95	209.37	796.09	216.63	1275.86	638.49
	HNR	Hz	7.36	3.03	5.59	0.93	5.03	0.86	6.83	1.85	11.56	3.62	7.76	1.85
	Jitter	dB	3.7	1.1	4.51	0.58	3.35	0.56	4.41	1.15	2.71	0.96	3.53	1.2
	Shimmer	dB	1.24	0.27	1.42	0.16	1.18	0.19	1.35	0.32	0.99	0.26	1.24	0.27

Table 7 Table of means and standard deviation of acoustic descriptors of vocalisations used in Experiment 5 and 7.

4.1.4 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 randomised order and were asked to listen to the 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 vocalisation ($Laughter_S$, $Laughter_V$ and Vowels) and prompted to provide the same judgement after each block. The order of these three blocks was randomised. This task lasted for approximately 10 minutes.

Task 2: Speaker recognition from speech

This task was included to objectively assess the familiarity of participants with the speakers. Participants were informed that they had heard 5 different speakers in the previous task 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 any familiarity. Following this, participants underwent a brief voice

(re)familiarisation task: they were presented with a brief speech sample of each of the five speakers (two consecutive sentences from the rainbow passage [Fairbanks, 1960; mean duration: 6.6 secs, SD = .49 secs]) while the speaker's full name and a photograph was presented on the screen. After this, participants were presented with 6 sentences (Bench, Kowal & Bamford, 1979) 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 (Southern British English, Canadian, North American, Scottish and Northern Irish accents were present in the speaker set). 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 after the offset of the stimulus to make a response. This task lasted for approximately 5 minutes.

Task 3: Speaker discrimination from non-verbal vocalisations

Following Tasks 1 and 2, familiar participants were ask 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) – 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 range from 5.71 to 3.54). Overall, listeners thought they were more familiar with the speaker's speaking voices than with their laughter (t[21]= 4.203, p < .001, Cohen's d = 1.8343). After this the main speaker discrimination task was started. The design and

procedure of this task were identical to the one used in Experiment 4 (see Section 3.1.4). This task lasted for approximately 35 minutes.

4.1.5 Results

Speaker recognition from speech

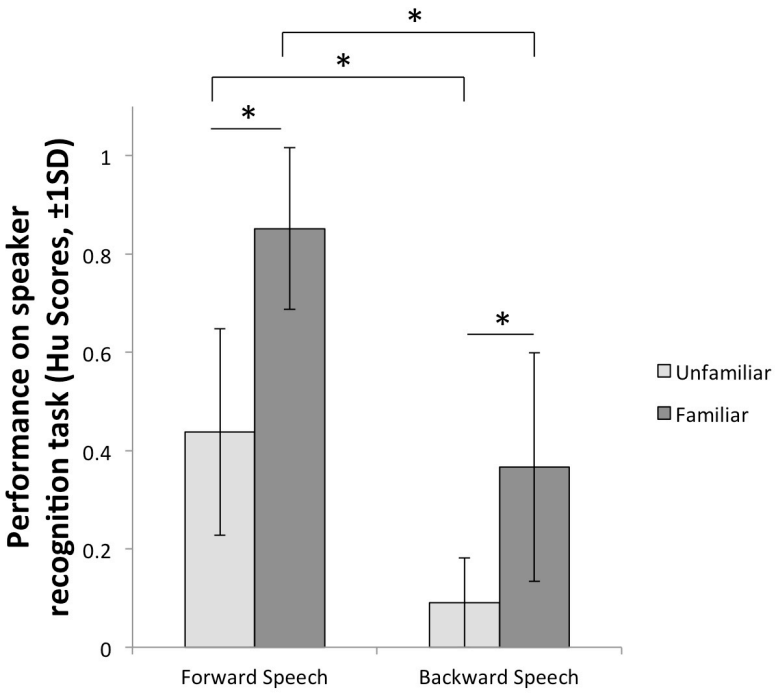

Figure 14 Unbiased hit rates for the speaker recognition task.

One data set for the unfamiliar group was lost due to experimenter error. Results are shown in **Figure 14**. Unbiased hit rates (Hu scores) were calculated using the formula provided by Wagner (1993) and arcsine transformed. These scores were into entered a 2 (familiar, unfamiliar listeners) x 2 (backward speech, forward speech) repeated measures ANOVA. There were significant main effects of listener group (F[1,41] =59.662, p < .001, η_p^2 = .593) and condition (F[1,41] = 143.021, p < .001, η_p^2 = .777) as well as an interaction (F[1,42] = 16.055, p < .001, η_p^2 = .281). Familiar listeners were

significantly better at identifying speakers from both backward and forward speech than unfamiliar listeners (*p*s < .001). The difference between familiar and unfamiliar listener was bigger for forward speech, either indicating that familiar listeners had a bigger advantage for forward speech or that unfamiliar listeners were close to floor (although unfamiliar scores were significantly above zero, as determined by a one sample t-test, $t[20] = 4.584$, $p < .001$, Cohen's $d = 2.05$). In terms of raw accuracy scores, the familiar listeners' performance was high for forward speech (*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* = 58.9%; *SD* = 18.3%) can be explained by the brief familiarisation phase that preceded this task. For backward speech, the performance of unfamiliar listeners was close to chance level (*M* = 26.4%; *SD* = 11.2%), although in line with results based on Hu scores, a one-sample t-test against chance performance revealed significantly higher performance for this group ($t[20] = 2.646$, $p = .015$, Cohen's $d = 1.183$). Familiar listeners' performance was much higher compared to unfamiliar listeners (*M* = 57%; *SD* = 18.4%). Both mean accuracy scores and Hu scores for backward speech thus indicate that familiarity with the voice of the speaker for the familiar group goes beyond identification based on – in this context – speaker-specific cues, such as regional accents.

Speaker discrimination from non-verbal vocalisations

D' scores were computed and entered into a 2 (listener 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 ($F[5,210] = 65.004$, $p < .001$, η_p^2

= .607) but no interaction (F[5,210] = .263, p = .933, η_p^2 = .006). Post-hoc t-tests further explored the effects of condition and listener group. Significant advantages for familiar listeners were expected across all conditions. The predictions for condition effects were identical to those for Experiment 4 (see **Figure 12**). The post-hoc paired t-tests (8 comparisons, corrected alpha = .006) largely replicated the pattern of results per condition shown in Experiment 4 (see **Figure 15**).

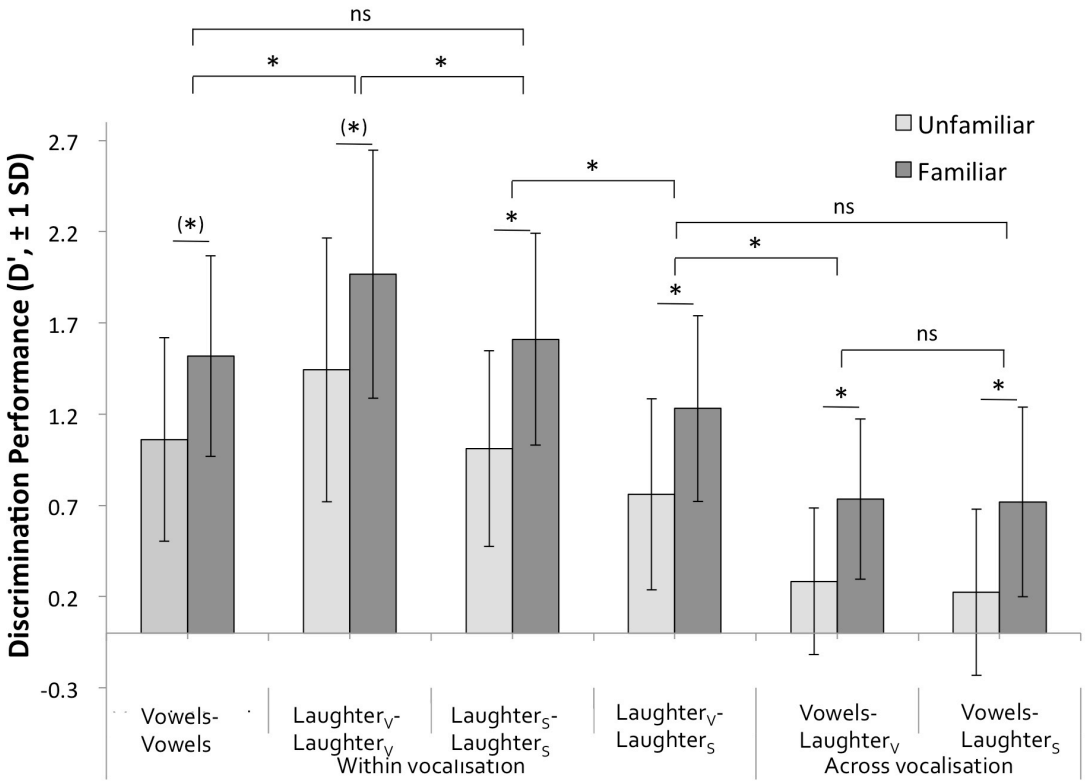

Figure 15 Average d' scores per condition for the speaker discrimination task. Significant comparisons (Bonferroni-corrected) are highlighted with an asterisk; marginally significant results are highlighted with an asterisk in brackets.

In contrast to the findings of Experiment 4, performance for Vowels-Vowels was lower and significantly worse compared to Laughter$_V$-Laughter$_V$ (t[43] = 4.929, p < .001, Cohen's d = 1.502) but was similar to Laughter$_S$-Laughter$_S$ (t[43] = .289, p = .774 Cohen's d = .088). Furthermore, performance for Laughter$_V$-Laughter$_S$ and Vowels-Laughter$_V$ were different (t[43] = 7.385, p < .001, Cohen's d = 2.252), while there was

no difference between Vowels-Laughter$_V$ and Vowels-Laughter$_S$ (t[43] = .567, p = .573, Cohen's d = .173; see **Figure 15**). In line with previous findings from Experiments 3 and 4, participants performed better at discriminating speakers for within-vocalisation trials compared to across-vocalisation trials (t[44]= 13.23, p < .001, Cohen's d = 3.989), with performance dropping to zero for unfamiliar listeners in the two across-vocalisation 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 listener group for each condition. This showed, as predicted, a significant advantage for familiar listeners over unfamiliar listeners for all conditions (all ps ≤ .004), with the exception of marginally significant advantages for Vowels-Vowels (t[21] = 2.679, p = .011, Cohen's d = 1.169) and Laughter$_V$-Laughter$_V$ (t[21] = 2.548, p = .015, Cohen's d = 1.112).

In parallel to Experiment 3 and 4, a response bias analysis was run. Collapsing across listener group, one-sample t-tests showed that for Laughter$_V$-Laughter$_V$ and Laughter$_S$-Laughter$_S$ there was a significant bias towards responding 'same' (ps < .001). For the across-vocalisation trials as well as Laughter$_V$-Laughter$_S$ there was, however, a significant bias towards responding 'different' (all ps < .001). No bias was found Laughter$_V$-Laughter$_S$ (t[43]= .209, p = .836, Cohen's d = .06).

4.1.6 Discussion

Experiment 5 formed a replication of Experiment 4 that additionally explored the effect of familiarity with a speaker on listeners' abilities to generalise identity-related information across diverse vocalisations. The results were very similar to those in Experiment 4, 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 the current set of studies. Performance for Vowels-Vowels was noticeably lower in Experiment 5 compared to Experiment 3 and 4, which could be attributed to the vowel tokens being relatively similar across the speakers used in the current stimulus et (who were all female and with relatively low average pitch) in contrast to the clear differences between male and female speakers in Experiment 4 (e.g. in F0). Otherwise, no striking differences were found between Experiment 4 and 5.

In line with findings from face and speech perception (Bruce et al., 2001; Jenkins et al., 2011; Nygaard & Pisoni, 1998; Ramon & Van Belle, 2016), a consistent advantage for familiar listeners over unfamiliar listeners was found, where the familiar listeners have a greater ability to generalise identity-related information across a range of spontaneous and volitional non-verbal vocalisations in a speaker discrimination task. The current 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 vocalisations to speaker-specific templates that entail idiosyncrasies and are based on a range of vocal outputs. Unfamiliar listeners can only access averaged, prototypical voice templates, which may serve well as a useful heuristic to assess speaker characteristics. The prototypical voice templates are, however, underspecified compared to a familiar listener's speaker-specific representations (see also Experiments 3 and 4). The increased specificity of representations for familiar voices offer 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 experiments (across-vocalisation judgements, the presence of vocalisations produced under reduced volitional control and mismatches in volitional control within a pair) had a similar effect on familiar and unfamiliar listeners. It should be noted that unfamiliar listeners were not able to discriminate between speakers for the across-vocalisation conditions ($Laughter_V$-Vowels and $Laughter_S$-Vowels) as indicated by *d'* scores that are not significantly different from zero, which could potentially mask interactions. Another consideration here could be the nature of the familiarity of the listeners in the study. The current study thus suggests that familiar listeners were able to recognise the five speakers with very high accuracy based on their speech, which serves as an objective measure of familiarity, and it was also shown that perceived familiarity with the speakers' speech and laughter 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 vocalisations, 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. Previous research has suggested that the type of familiarity (e.g. self versus personally familiar close friends, partners versus famous people or work colleagues) will affect how a voice is perceived (see Sidtis &

Kreiman,2012; Sugiura, 2014 and McGettigan, 2015 for discussions). In the context of the current study, it could therefore be argued that listeners may have been able to more successfully generalise across vocalisations if they has been presented with vocal signals from speakers they know in a wider range of contexts. More exposure to the speakers' full vocal inventory, in a way that is more representative of having learned a vocal identity through social interaction, could have linked all vocal signals to such a familiar speaker and allowed for better generalisation. A lack of interaction of condition with group is therefore perhaps not too surprising, given that even the familiar listeners tested were relatively unfamiliar with the vocal signals used in the study (especially with spontaneous laughter). Based on these results, a further experiment explored the performance of a similar group of familiar listeners on speaker discrimination using (voiced and whispered) words, nonwords and vowels – volitionally produced vocal signals that should be more familiar to the listeners.

4.2 Experiment 6

4.2.1 Introduction

Experiment 5 explored how familiarity with a voice affects speaker discrimination from volitional and spontaneous vocalisations. Familiarity with the speakers afforded listeners an overall advantage, although relative performance for spontaneous vocalisations showed an equivalent impairment for familiar and unfamiliar listeners. In the current experiment, the effect of familiarity on identity processing in voices was further explored by contrasting whispered and voiced vocal signals, specifically, vowels, nonwords and words (C-V-C structure). Voiced and whispered speech signals differ in a number of properties from the stimuli used in the previous experiment, such as the presence or absence and modulation of source and filter characteristics, (linguistic) complexity and familiarity with the stimulus type: In contrast to spontaneous vocalisations, where source information is drastically modulated and filter modulations are arguably reduced, filter information is fully preserved in whispered speech while source information is largely absent. Given this lack of diagnostic source information, previous studies using whispered speech have reliably shown that while listeners are still able to extract speaker characteristics from whispered speech, performance is significantly impaired compared to voiced speech – even when familiar with a voice (Abberton & Fourcin, 1978; Bartle & Dellwo, 2015; Orchard & Yarmey, 1995; Pollack, Pickett & Sumby, 1954; Yarmey, Yarmey, Yarmey & Parliament, 2001). For the current experiment, it was thus predicted that the task would be more difficult for whispered signals compared to voiced signals.

With regard to the stimuli used in Experiment 3-5, spontaneous vocalisations have been described as unarticulated vocalisations with only minimal modulation of

the filter characteristics being apparent during their production (see Ruch & Ekman, 2001 for a discussion of laughter). Further, laughter has been described as a 'stereotyped' vocalisation (Provine & Yong, 1993). Previous studies using speech samples have shown that not only increasing the duration of stimuli but also merely increasing the (linguistic) content of speech (e.g. number of phonemes) while controlling for duration improves listeners' accuracy in determining identity from the voice (Bricker & Pruzanksy, 1966; Pollack et al., 1954). The authors of these studies consequently propose that stimuli that include more linguistic content allow the listeners to sample more of the phonetic inventory of a speaker as well as more information about the dynamic use of their vocal apparatus, resulting in more reliable and accurate decoding of speaker identity-related information. To directly test whether this is the case for whispered and voiced speech (and for familiar and unfamiliar listeners), vowels, nonwords and words were used in the current study, thus modulating the linguistic complexity of the stimuli. It was predicted that increasing linguistic complexity, as indexed by more phonemes (vowels: V structure, nonwords/words: C-V-C structure), would improve performance.

With regard to familiarity, it was expected that familiar listeners' performance would be better compared to unfamiliar listeners, replicating the results of listener group in Experiment 5. Crucially, and in contrast to the findings of Experiment 5, an interaction between listener group and voicing was expected because all stimuli in the current study are speech-based. The familiar listeners had been extensively exposed to the speech of all the speakers used in the current experiments and should therefore be highly familiar with this kind of vocal output. It could thus be hypothesised that familiar listeners should have a greater advantage for voiced speech-based vocal

signals over unfamiliar listeners, while they may be similarly affected by the relatively unfamiliar whispered speech signals. The current experiment can thus provide further insights into the extent of the familiarity advantage of the listener groups used for the previous experiments. It also investigates how linguistic complexity, voicing and familiarity (with a stimulus and with a speaker) affect identity-related processing in the voice, factors which most previous studies (but see Bartle & Dellwo, 2015) have only looked at in isolation.

4.2.2 Participants

32 familiar (M_{Age}: 21.13 years; *SD*: 3.15 years, 27 female), 33 unfamiliar listeners (M_{Age}: 21.64 years; *SD*: 4.19 years, 18 female) were recruited at Royal Holloway, University of London. In line with the familiar listener group of Experiment 5, familiar listeners had been exposed to the voices featured in the stimulus sets 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 of Psychology for more than two years (see also Experiment 5). Unfamiliar listeners were recruited from other departments around campus and had had no exposure to the voices used in the study. All participants were native speakers of English, had 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. Two participants from the familiar listener group were excluded from any analyses due to an average performance (measured in *d'*) in the speaker

discrimination task that was more than 2 standard deviations below the group average, leaving 30 familiar listeners.

4.2.3 Materials

Words, nonwords and isolated sustained vowels were recorded in voiced and whispered versions from 5 speakers (all female, ages range from 29 – 42 years). As in Experiment 5, all speakers were lecturers at the Department of Psychology at Royal Holloway and selected based on their exposure to a subgroup of undergraduate degree students at the department. 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. Speakers recorded monosyllabic words with a C-V-C structure with two voiced consonants ('bad', 'big', 'man', 'long'). 15 words were chosen from the MRC psycholinguistics database (Wilson, 1988) based on being relatively frequent as assessed by a Brown verbal frequency ranging between 50 and 140 in the stimulus set. Words were furthermore chosen to minimise regional accents cues: since the sample of speakers included individuals speaking of varieties Scottish, Northern Irish, North American, Canadian and Southern British English, care was taken to, for example, not include words including potentially rhotic segments, such as 'car' or words including diphthongs that may be monothongised in some of the regional accents, such as the vowel in 'game'. Furthermore, 15 nonwords were created based on the same constraints within which the words were selected. All nonwords were monosyllabic C-V-C words, with consonants exclusively being voiced obstruents (e.g. /geb/, /boib/, /roim/). For the vowel category, speakers produced /i/, /a/, and /u/, with

each individual vowel lasting for around 0.6 seconds to match the duration of the words. All stimuli were extracted and normalised for root-mean-square intensity in Praat (Boersma & Weenink, 2010).

4.2.4 Design and Procedure

Participants completed two computer-based tasks for during the testing session: A brief speaker recognition task to objectively assess familiarity in listeners, and a speaker discrimination task. For speaker recognition from speech, the task was identical to the one used in Experiment 5 (see Section 4.1.4): Participants listened to brief sentences presented as forward and backward speech, while identifying individual speakers in a 5-way forced choice paradigm. The speaker discrimination task was similar to the one used in Experiments 3-5: Participants were presented with paired triplets of vocalisations (e.g. '/i/ /a/ /u/' or 'gone bad mean') and indicated in a two-way forced choice paradigm whether they thought these two triplets were produced by the same speaker or by two different speakers. Triplets were chosen to approximate the duration of the stimuli in Experiments 3-5 (~2 seconds). In the current experiment, there were six conditions including two triplets of vowels, nonwords and words each – presented either in voiced or whispered realisations. No across-vocalisation or across-voicing conditions were included. The beginning of each triplet was cued by a tone to avoid confusions with regard to which stimuli belonged to which triplet. A brief silent period of 0.1 seconds was inserted between each sound within a triplet. Triplets were separated by another 0.5 seconds of silence. The order of sounds within triplets was pseudorandomised for the vowel category, ensuring that each triplet included one exemplar per vowel (/i/, /a/, /u/; presented in randomised

order). Stimuli within triplets containing words and nonwords were fully randomised. Following these tasks, familiar participants were asked to report how familiar they thought they were with each lecturer's speaking voice on a scale from 1 (not familiar at all) – 7 (very familiar). These data confirm that familiar listeners indeed perceived themselves to be highly familiar with the speaking voices ($M_{all\ speakers}$ = 5.58; $SD_{all\ speakers}$ = 1.53; means for individual speakers ranging from 6.54 to 4.92).

4.2.5 Results

Speaker recognition from speech

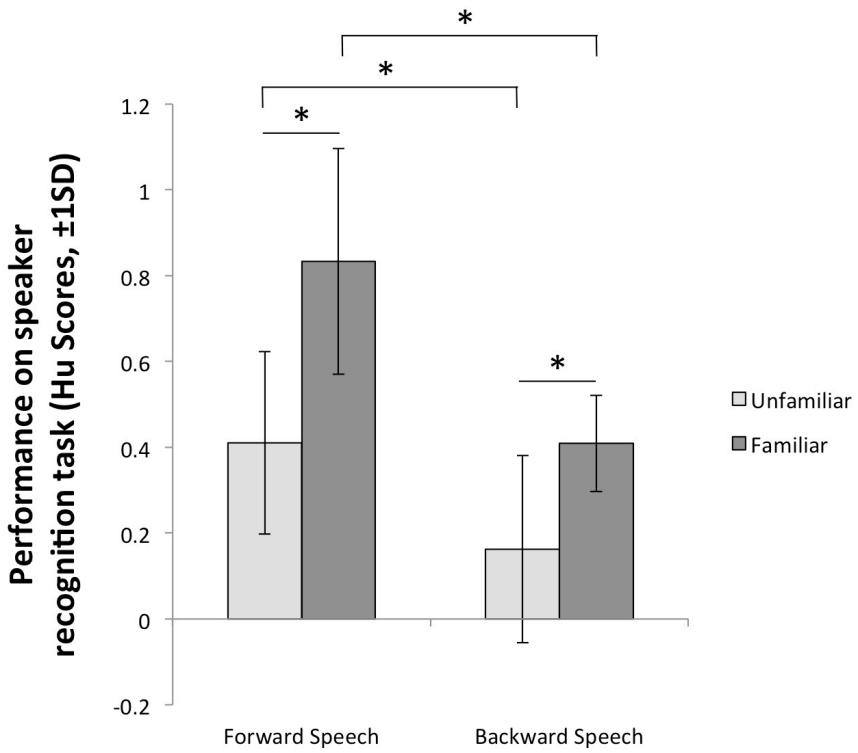

Figure 16 Performance in the speaker recognition task.

Data from 30 familiar and 29 unfamiliar participants were entered into the analysis – 4 data sets of unfamiliar listeners for this task were lost due to experimenter error. Results are shown in **Figure 16**. The same analyses were run as in Experiment 5:

unbiased hit rates (Hu scores) were calculated using the formula provided by Wagner (1993) and arcsine transformed. These scores were entered into a 2 (familiar, unfamiliar listeners) x 2 (backward speech, forward speech) repeated measures ANOVA. There were significant main effects of listener group (F[1,57] =51.492, p < .001, η_p^2 = .484) and condition (F[1,42] = 107.308 p < .001, η_p^2 = .661) as well as an interaction (F[1,57] = 20.151, p < .001, η_p^2 = .268). The results were very similar to those of the same task reported for Experiment 5: Familiar listeners were significantly better at identifying speakers from both backward and forward speech than unfamiliar listeners. Reversing the speech had a bigger effect on familiar listeners, possibly driven by unfamiliar listeners being close to floor (although again unfamiliar scores were significantly above zero, as determined by a one sample t-test, t[28] = 7.839, p < .001, Cohen's d = 2.963). In terms of raw accuracy scores, the familiar listeners' performance was high for forward speech (M = 84.1% SD = 13.6%). Similar to the results of Experiment 5, the clear above-chance performance (i.e. >20% correct) for forward speech for unfamiliar listeners (M = 53.1% SD = 23.5.3%) can be explained by the brief familiarisation phase that preceded this task. For backward speech, the performance of unfamiliar listeners was close to chance level (M = 31% SD = 15.5%) although a one-sample t-test against chance performance revealed significantly above-chance performance for this group (t[28] = 3.839, p = .001, Cohen's d = 1.451). Replicating the results of Experiment 5, familiar listeners' performance was much higher (M = 56.3%; SD = 16.8%).

Speaker discrimination from voiced and whispered vocalisations

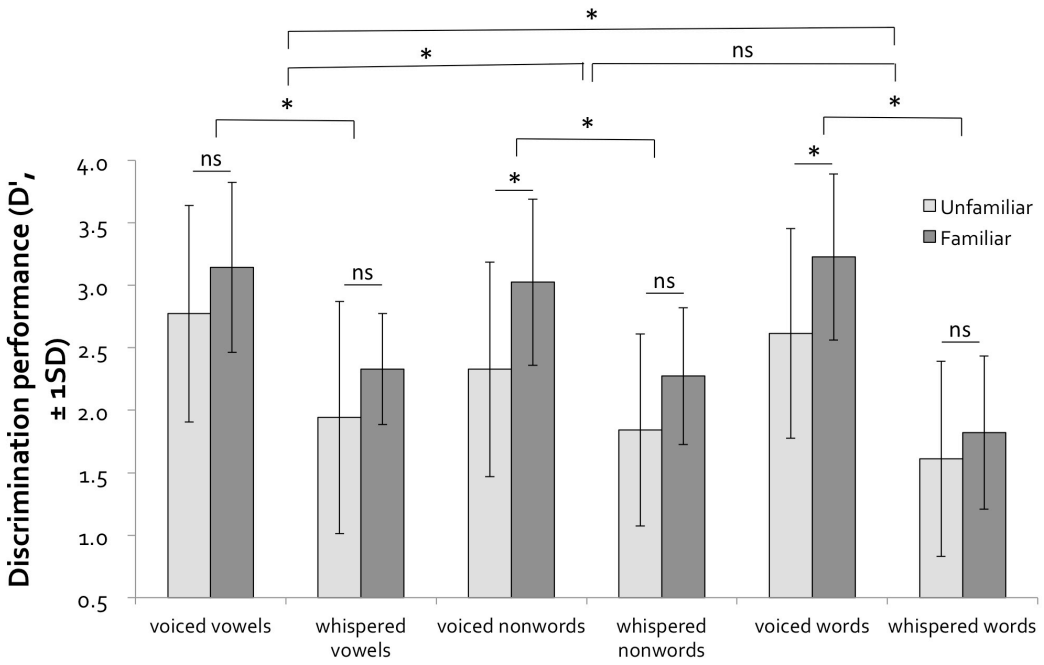

Figure 17 Average d' scores per condition for the speaker discrimination task of Experiment 6.

D' scores were computed and entered into a 2 (listener group) x 2 (voicing) x 3 (linguistic complexity) repeated measures ANOVA. There were significant main effects of listener group (F[1,61]= 8.483, p = .005, η_p^2 = .122), voicing (F[1,61]= 311.291, p < .001, η_p^2 = .836) and linguistic complexity (F[2,122]= 8.475, p = .001, η_p^2 = .122). There was furthermore a significant interaction between voicing and listener group (F[1,60]= 4.684, p = .034, η_p^2 = .071), indicating that the difference in performance between familiar and unfamiliar listeners is overall bigger for voiced compared to whispered conditions. A further interaction was found between voicing and linguistic complexity (F[2,122]= 15.277, p < .001, η_p^2 = .2). Neither the two-way interaction between linguistic complexity and listener group (F[2,122] = 1.435, p = .242) nor the three-way interaction between linguistic complexity, listener group and voicing (F[2,120]= 1.962, p = .145) were significant.

Post-hoc t-tests further explored the effects of condition, voicing and listener group. Paired t-tests (3 comparisons x 2 groups) were run to investigate the effect of voicing: this shows that listeners were able to better discriminate speaker from voiced signals across all listeners and within listener group for all three vocal signals (all *p*s < .001). Post-hoc paired t-tests (3 comparisons, corrected alpha = .017) also explored the effect of linguistic complexity using averaged scores of voiced and whispered conditions by vocal signals. Against predictions, increasing linguistic complexity did not result in better performance. In fact, performance was significantly better for vowels compared to words and nonwords (*p*s ≤ .003), while performance was similar for words and nonwords (*t*[61] = .839, *p* = .388, Cohen's *d* = .215). Further post-hoc independent-samples t-tests (6 comparisons, corrected alpha = .008) were run to explore the effect of group for each condition. While there were trends apparent across all conditions (*p*s ≤ .065; although whispered words, *t*[61] = 1.174, *p* = .245, Cohen's *d* = .301), indicating better performance for familiar compared to unfamiliar listeners, performance differed significantly only for voiced nonwords (*t*[61] = 3.561, *p* = .001) and voiced words (*t*[61] = 3.18, *p* = .002, Cohen's *d* = .912). Means are displayed in **Figure 17** and highlight the complex pattern on interactions between the factors of interest.

In parallel to the previous speaker discrimination experiments, a response bias analysis was run. Collapsing across listener group, one-sample t-tests showed there was a significant bias towards responding 'same' for voiced nonwords *t*[61] = 3.012, *p* = .004, Cohen's *d* = .771). Significant biases towards responding "different" were found for voice and whispered vowels and whispered nonwords (all ps < .001). No biases were found for whispered and voiced words (*p*s > .037). While this complex pattern of

response biases does not correspond to the findings from the previous experiments, conditions and vocalisations differed, which may explain these divergent results.

4.2.6 Discussion

The current study investigated how familiarity, voicing and linguistic complexity affect listeners' ability to discriminate between speakers. In line with the results of Experiment 5 and a large literature showing familiarity advantages during the processing of vocal signals, a main effect of familiarity was found, with familiar listeners performing better at the task than unfamiliar listeners (although this was statistically significant for only 2 conditions). The current study further replicates findings from work (Abberton & Fourcin, 1978; Bartle & Dellwo, 2015; Orchard & Yarmey, 1995; Pollack et al., 1954; Yarmey et al., 2001), showing that the extraction of identity related information from voices is more difficult for whispered speech than for voiced speech, since crucial source information is absent in whispered speech. It should be noted that despite the absence of source information in whispered speech, performance across all whispered conditions in the discrimination task was still relatively high ($d' > 1.8$). This suggests that even with only filter information being available, enough cues to speaker identity are still encoded in the vocal signals to allow listeners to accurately discriminate between speakers – a finding consistent with previous studies on whispered speech (Abberton & Fourcin, 1978; Bartle & Dellwo, 2015; Orchard & Yarmey, 1995; Pollack et al., 1954; Yarmey et al., 2001).

Intriguingly and in contrast to the findings of Experiment 5, there was an interaction between voicing and listener group – with the difference in performance between familiar and unfamiliar listeners being overall bigger for voiced compared to

whispered conditions. Bartle and Dellwo (2015) conducted a study on whispered and voiced speech samples with a group of naïve listeners and an expert group of phonetically trained listeners. These authors first report an overall advantage for phoneticians. They also report an interaction between listener group and stimulus type: they found that the difference in performance between naïve and trained listeners was smaller for voiced speech compared to whispered speech. The interaction reported by Bartle and Dellwo (2015) shows the opposite pattern compared to the one in the current study: Trained listeners were less affected by whispered speech compared to untrained listeners. This does not necessarily conflict with the current findings. While overall, both formally trained listeners (Bartle & Dellwo, 2015) and untrained familiar listeners' performances (current study) were better than those of the naïve listener groups, these two groups of expert listeners are not directly comparable. Phoneticians are likely to use different strategies to pick up on secondary cues to speaker identity to perform the task based on their specific training, while untrained (albeit familiar) listeners may not adapt the same listening strategies. As hypothesised, the familiar listeners in this study had mainly been exposed to speech signals from the speakers prior to the study, having been taught by them in a lecture setting – this increased exposure to voiced speech (as opposed to whispered speech) may thus have resulted in the bigger advantage for the voiced conditions for familiar compared to unfamiliar listeners. Familiar listeners were, however, unlikely to have heard the speakers whisper for prolonged periods of time and are thus relatively unfamiliar with this specific type of vocal signal produced by the familiar speakers (see also the discussion of Experiment 5). With no explicit formal training to compensate for the lack of acoustic information (see Bartle & Dellwo,

2015) and/or low familiarity with the whispered speech of their lecturers, the advantage over unfamiliar listeners is thus less pronounced for whispered speech. This finding thus implies that the familiarity advantage in untrained listeners is closely linked to the vocal signals listeners to which they have been most frequently exposed, with only limited compensation being apparent for a lack of source cues.

Another factor explored in the current study was the effect of linguistic complexity on speaker discrimination. Based on studies showing that speech samples including more (linguistic) information result in better performance (Bricker & Pruzansky, 1966; Pollack et al., 1954; Schweinberger et al., 1997), it was hypothesised that increasing the linguistic complexity of the stimuli would enhance performance. This was not the case, as performance for vowels was significantly better compared to words and nonwords. On the one hand, it could be argued that, in the presence of only limited linguistic context (V versus C-V-C structures), there is a benefit in having reliable, steady-state filter and (for voiced signals) source information as is the case for vowels in contrast to more rapid changes in signals involving consonants – although this would directly conflict with previous findings (Brick & Pruzanksy, 1966; Pollack et al., 1954). On the other hand, this effect could be an artefact of the task design: For vowels, participants directly compared two triplets of the same vowels (/i/, /a/ and /u/ in randomised order), while for words and nonwords, participants compared different item sets of words (i.e. it was possible that none of the words from the first triplet occurred in the second triplet). Therefore, comparing word/nonword triplets would require a higher level of abstraction or generalisation, potentially decreasing accuracy. It is, however, particularly surprising that performance for words (both whispered and voiced) was similar, or even lower, compared to nonwords and

vowels. The words used in the study were frequent and should thus be highly familiar as a stimulus – at least in contrast to novel nonwords. The prototype theory would have predicted that through this familiarity with the words, well-formed specific representations for these words should be present, while no well-formed representations should be present for the novel nonwords. This should potentially be even more strongly the case for the familiar listeners, if they had previously heard the speakers utter the same words (although this cannot be verified for the current sample of participants). Stimulus effects or unexpected interactions with, for example, intelligibility may thus have led to the results reported here.

Overall, this study confirms the findings of Experiment 5, showing that familiarity with a voice affords listeners an advantage during the extraction of identity-related information from voices for a range of vocal signals. In the context of untrained familiar listeners, the magnitude of this advantage seems to be directly linked to the vocal signals to which these listeners have previously been exposed - in this case, mainly voiced speech. The current study did not find any conclusive evidence to show that increasing the linguistic complexity of a signal, providing greater semantic and phonological content, increases performance for speaker discrimination, thus failing to replicate findings from previous work (e.g. Bricker & Pruzansky, 1966). From the current study, it thus remains unclear whether sampling more varied spoken signals allows listeners to extract speaker-related information more successfully.

4.3 General discussion

The last two studies have shown that accurately attributing divergent vocal signals to a single individual is challenging even in the context of being familiar with a person's voice. Without prior exposure to the full vocal inventory of a speaker, (untrained) listeners cannot fully compensate for the absence (e.g. missing Fo in whispered speech) or drastic modulation (e.g. modulations of source and filter characteristics in spontaneous vocalisations) of one or more diagnostic cues to a speakers' identity. Experiments 5 and 6 further highlight that familiar listeners may need to be highly familiar with a vocal signal from these speakers in order to reliably extract speaker characteristics: When processing relatively infrequent and thus unfamiliar vocalisations, such as spontaneous laughter in Experiment 5, familiar listeners still have an advantage over unfamiliar listeners but performance decreases drastically. Bigger advantages for familiar listeners can be seen for highly familiar vocal signals, such as voiced speech signals in Experiment 6, that listeners are likely to have encountered before by these speakers. Along a similar line of argument, studies further show that the type of familiarity (famous person, professional relationship, friendship or romantic relationship, etc.) will on the one hand affect which vocal signals listeners have been exposed to – on the other hand, additional features, such personal significance (e.g. famous person versus partner), will be present or absent in such different types of familiarity. Both factors may influence listener's judgements through expertise in processing highly familiar stimuli or through socio-emotional associations formed with a voice (Sidtis & Kreiman, 2012; Suguira, 2014; McGettigan, 2015).

In line with a prototype-based approach to voice perception, the familiarity advantage observed during speaker discrimination may be based on the retrieval and matching of the incoming vocal signal to underlying representations (prototypical representations for unfamiliar listeners versus speaker-specific representations of voices for familiar listeners, see 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. Listeners may encode voices based on abstract representations of the vocal tract, that is 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 (see also Experiments 3 and 4), are furthermore likely to be formed and shaped based on the specific long-term exposure to vocal outputs of each familiar person. Thus, more frequently encountered vocalisations from a familiar speaker will have a more robust representation, while representations for infrequently encountered vocal signals will be less well formed. 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 Experiments 5 and 6 have 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, and this tradition has meant that direct comparisons of familiar and unfamiliar listeners within the same task are rare in the literature. Yet in order to probe the underlying representations of voices, it is important to consider how familiarity affects voice perception in multiple contexts, not just in overt recognition. For example, there is evidence that listeners can report a sense of familiarity with a known voice in the absence of overt recognition (see Hanley, Smith & Hadfield, 1998), thus directly providing evidence that familiarity is not reducible to the task of person naming. The data of Experiment 5 and 6 suggest that familiarity with voices can affect performance on a speaker discrimination task. In the context of the current studies, familiar listeners' prior exposure to the voices may have led to the development of speaker-specific expertise that may be linked to the refinement of prototypical representation, which may interact with different aspects of voice processing across a range of tasks.

One limitation of the current experiments is that while familiarity with the voices used was assessed based on self-report and speaker recognition based on sentences, no explicit speaker recognition task was run to shed further light on whether listeners were actually able to recognise the voices based on the stimuli they

heard in the speaker discrimination task and whether there were speaker-specific, vocalisation-specific or listener-specific biases. Future studies will need to formally assess speaker recognition from a wider range of vocalisations. Future work will furthermore need to determine the strategies used to perform the task in familiar listeners: From the current studies, it remains to be determined whether familiar listeners recognised one or both of the stimuli presented within a pair and based their decision on (partial) recognition, or whether other strategies were used. It would be interesting to explore if the listener's performance differed across trials where a speaker was recognised for one vocal signal in a pair, compared with trials lacking recognition. It could further be explored how discrimination performance on such trials interacts with recognition accuracy (i.e. whether performance is affected by a sense of recognition, regardless of whether that impression is correct).

5 The neural underpinnings of voice identity processing in familiar and unfamiliar listeners

Experiment 7 aims to investigate how familiarity with a speaker and variability in vocal signals modulate the neural networks associated with voice identity processing. Familiar and unfamiliar listeners performed a one-back speaker discrimination task in the scanner while listening to short sentences, vowels, volitional laughter and spontaneous laughter. Results suggest that activation in auditory cortices (overlapping with the TVAs, Pernet et al., 2015) is modulated by vocalisation type and speaker identity. Activation in frontal lobes is modulated by familiarity with a voice. Findings are discussed in the context of differences in task demands and stimulus properties.

5.1.1 Introduction

Neuroimaging studies have attempted to probe the neural underpinnings of the processing of person-related information. The majority of these studies have focussed on faces in the visual domain, establishing that faces are preferentially processed in fusiform gyrus (FG, also known as the fusiform face area) by contrasting pictures of human faces with other similarly complex objects, such as houses (Kanwisher, McDermott & Chun, 1997). In parallel to this, it was proposed that voices are also processed in a specific brain area (e.g. Belin et al., 2004, 2011; Ellis, Jones, Mosdell, 1997): The temporal voice areas (TVAs) have consequently been described as candidate regions for voice-selective processing, with sections of bilateral STG and STS preferentially responding to human vocal sounds compared to other non-human sounds (Pernet et al., 2015; see also Belin et al., 2011). Researchers have thus attempted to further on the one hand describe and map out the specific functions of (subsections of) these temporal voice areas. On the other hand, they have attempted to link speaker identity processing to responses in the TVAs. For explicit speaker identity processing, anterior regions of STG which is part of the TVAs, have been implicated to be involved in speaker identity processing (see Mathias & von Kriegstein, 2013 for a review): Studies have shown that anterior STG/STS activation is modulated by speaker identity in unfamiliar voices (Belin & Zatorre, 2003; Formisano, De Martino, Bonte, & Goebel, 2008; Imaizumi et al., 1997; von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003) and by hearing familiar voices (learned voices: Andics et al., 2010; Latinus, Crabbe, & Belin, 2011; personally familiar voices: Nakamura et al., 2001; von Kriegstein & Giraud, 2006; von Kriegstein, Kleinschmidt, Sterzer & Giraud, 2006). Furthermore, the temporal pole has specifically been highlighted by authors has a hub

for the (amodal) processing of person identity (Belin et al., 2011; Olson, Plotzker & Ezzyat, 2007; see Perrodin, Kayser, Abel, Logothetis & Petkov, 2015 for a recent review), which is in line with the evidence from the literature of other identity related stimuli (e.g. faces or names; see Gorno-Tempini et al., 1998; Pourtois, Schwarz, Seghier, Lazeyras & Vuilleumier, 2005).

Speaker identity processing may, however, not be strictly limited to the processing of voices but may also include higher-order and abstracted processing of the vocal signals, outside of the TVAs. Some studies have indeed associated additional regions outside of temporal cortices with the processing of familiar and unfamiliar voices: Parietal regions, including the precuneus, have been shown to be sensitive to voice identity processing in general (Rämä et al., 2004) while also showing differential responses to familiar and unfamiliar voices (von Kriegstein et al., 2006; Shah et al., 2001; see Van Lancker, Cummings, Kreiman & Dobkin, 1988 and Van Lancker, Kreiman & Cummings, 1989 for patient studies). It should be noted that parietal activation has mostly been found in fMRI studies using both face and voice cues, as opposed to just audio or video stimuli (e.g. von Kriegstein et al., 2006; Andics et al., 2010): Latinus et al. (2013) do not report any parietal activation in a purely auditory study. It has thus been proposed that parietal cortex may serve as a crossmodal hub for identity processing (e.g. Campanella & Belin, 2007; Latinus et al., 2013). Additionally, studies using a range of designs and task have reported frontal regions, such as right medial frontal gyrus (MFG) and superior frontal gyrus (SFG), dorsolateral prefrontal cortex (DLPFC) as well as other inferior regions of the frontal cortex (Latinus, Crabbe & Belin, 2011; Stevens, 2004; von Kriegstein & Giraud, 2004) to be selectively activated for voice identity processing over linguistic processing as

well as for familiar versus unfamiliar voice processing. This has been at times described as a result of differences in task difficulty between conditions and tasks, resulting in differential responses in areas linked to executive functions. Additionally, the fusiform gyrus, implicated in selectively processing faces, has been linked to familiar voice processing by some authors, proposing that cross-modal processing takes places where the memory is a person's face is automatically co-activated when recognizing a familiar person from their voices (e.g. Giraud & von Kriegstein, 2004; von Kriegstein et al., 2006; von Kriegstein et al., 2007).

In sum, the processing of speaker identity has been strongly linked with activation in right anterior STS by studies using a range of tasks and manipulations. Other sites in parietal and frontal regions as well as fusiform have also been linked to (familiar) voice processing. As is the case with behavioural studies of voice processing, these neuroimaging studies have almost exclusively used relatively uniform vocal stimuli, such as short sentences or vowel sounds. The current study explored the processing of speaker identity related information in unfamiliar and familiar listeners, using variable vocal signals (sentences, vowels as well as spontaneous and volitional laughter). Based on previous research, a sensitivity to the presence of different speakers in anterior temporal cortices was predicted, with activation in (right) anterior STS and temporal poles being modulated by listener familiarity, despite the variability in vocal signals.

5.1.2 Participants

21 unfamiliar (14 female, mean age: 21.1 years, *SD*: 3.2 years) and 19 familiar (17 female, mean age: 20.3 years, SD: 1.5 years) participants were scanned. One

unfamiliar listener was excluded from all analyses due to an abnormality that was found in their brain. None of the participants reported any history of neurological incidents. All participants were right-handed, native speakers of English and reported healthy hearing. The study was approved by the Ethics committee of the Department of Psychology at Royal Holloway, University of London and participants were paid £15 for their participation.

5.1.3 Materials

Stimuli recorded from 6 female speakers (including the 5 speakers recorded for Experiment 5 plus one additional speaker) were used in this experiment: Vowels, $Laughter_S$, $Laughter_V$ and brief sentences (BKB sentences; Bench, Kowal & Bamford, 1979) were included in the stimulus set, with 8 tokens per vocalisation per speaker being selected. Brief sentences were included as a control condition – familiar participants have been shown to very reliably recognise the speakers used in this study from these sentences and should thus most reliably show processing differences between familiar and unfamiliar listeners (see Speaker Recognition tasks in Experiment 5 and 6, as well as the results reported below). For details recording the recording procedure, see Experiment 5. Based on ratings from a pilot study (see Experiment 5), stimuli were selected. This resulted in a stimulus set including 192 sounds. Vowels included /a/ and /i/ only, and each sentences trial included two sentences in order to approximate the duration of the other vocalisations. Stimulus durations across all vocalisations ranged from 1.05 – 2.6 seconds (Mean: 2.14 seconds; *SD* = .43 seconds). Mean durations across vocalisations were matched as closely as possible ($Vowels_{Mean}$ = 2.1 seconds, *SD* = .33 seconds; $Laughter_{S\ Mean}$ = 2.06 seconds,

SD = .42 seconds; $Laughter_{V\ Mean}$ = 1.88 seconds, *SD* = .51 seconds; $Sentences_{Mean}$ = 2.49 seconds, *SD* = .1 seconds). $Laughter_S$ was selected to be significantly higher in authenticity than $Laughter_V$ ($Laughter_{S\ Mean}$ = 5.39, *SD* = .44; $Laughter_{V\ Mean}$ = 3.2, *SD* = .75; t[94] = 17.166, *p* < .001, Cohen's *d* = 3.541). While an as close as possible match for arousal was attempted, $Laughter_S$ was nonetheless higher in arousal ($Laughter_{S\ Mean}$ = 5.13, *SD* = .48; $Laughter_{V\ Mean}$ = 4.53, *SD* = .42; *t*[94] = 6.887, *p* < .001, Cohen's *d* = 1.421).

5.1.4 Practice Task

Before entering the scanner, participants performed a behavioural practice of a variation of the task they would perform in the scanner. For this, participants were presented once with all 192 sounds. Participants then completed a covert (i.e. no overt responses necessary) same/different speaker 1-back task. 40 occasional on-screen prompts asking participants to judge whether the previous two sounds were produced by the same or two different speakers. Prompts were spaced out through the task to gather responses from 4 trials per comparison (2 same, 2 different) of the possible 10 vocalisation pairings (Sentences-Sentences, Vowels-Vowels, $Laughter_V$-$Laughter_{V,}$ $Laughter_S$-$Laughter_{S,}$ Sentences-Vowels, Sentences-$Laughter_{V,}$ Sentences-$Laughter_{S,}$ Vowels-$Laughter_V$. Vowels-$Laughter_{S,}$ $Laughter_V$-$Laughter_S$). Sounds were pseudorandomised to result in an approximately equal number of 'same' and 'different' speaker trials to at least partially account for task difficulty. One data set was lost due to experimenter error. Independent samples t-tests show that in line with the previous studies (Experiments 5 and 6), even in this very small data set the familiar listeners' overall performance measured in percent correct was higher compared to

that of the unfamiliar group when averaged across all conditions (t[37]=2.305, p = .025, Cohen's d = .7578). The small number of trials precluded any further statistical analysis of the data.

5.1.5 fMRI image acquisition

Functional images were acquired in a 3T MR scanner (Magnetom Tim Trio, Siemens Medical Solutions, Erlangen, Germany). The auditory stimuli were presented via MRI-compatible insert earphones (Sensimetrics Corporation, Malden, MA, EUA) via a SONY STR-DH820 digital AV control centre (Sony, Basingstoke, UK) in MATLAB (version 2013b, Mathworks, Inc., Natick, MA) using the Psychophysics Toolbox extension (http://psychtoolbox.org/). Visual information was presented on a screen via a back projector, which participants viewed through a mirror placed on top of the head coil. Two functional runs were collected, presenting all stimuli twice throughout the experiment. The two functional runs were split into 2 miniblocks each (lasting around 9 minutes), with the order of miniblocks being counterbalanced across participants using a Latin square. For each miniblock, 120 echo-planar whole-brain volumes (TR = 4.3 seconds, TA = 1.6 seconds, TE = 30ms, flip-angle = 78 degrees, 24 slices, 3mm x 3mm x 3mm in plane resolution with an inter-slice gap of 0.75mm) were acquired in ascending order. Data were acquired using sparse acquisition, allowing for auditory stimuli to be presented in silence (Hall et al., 1999). The onset of the auditory stimuli for each trial was timed, so that the mid-point of the sound occurred always at the same point of the trial, 1.3 seconds after trial onset, with the varying duration of the sounds thus providing natural jitter. Each miniblock included 96 sound trials, 12 rest trials (fixation cross presented on the screen, participants did not engage in any

specific task or behaviour), 6 vigilance trials (including two volumes each) plus 3 dummy scans at the start and end of each block. The in-scanner vigilance task was identical to the practice task: participants performed a same/different speaker one-back task with occasional prompts to monitor whether participants were paying attention to the sounds presented to them. Each run included 6 prompts, thus requiring participants to make 12 overt judgments in total throughout the experiment. On average, participants responded to the prompt 82% of the time (SD = 17%), with the lowest percentage per participant being 55%. One factor that may have contributed to the at times relatively low response rate may have been that participants only had two seconds to press a button, thus despite paying attention, participants' responses to these infrequent prompts may have been delayed and this not logged. Further, response rate may have been this low for some participants as a result of not paying attention to the task. A response rate of over 50% was, nonetheless considered to be sufficient evidence of participants being alert, thus no participants were excluded on this basis.

Within the four miniblocks, two fixed sets of stimuli were presented twice per participant, thus keeping the content of each miniblock independent from the other miniblock and keeping the content furthermore constant across all participants. The two subsets of stimuli presented within each miniblock were matched for arousal, duration and authenticity across sets (all *p*s ≥ .54). The order of trials was pseudorandomised within each miniblock: Randomisations were balanced for same/different judgements in order to control to some extent potential differences in task difficulty across different sequential pair types (see response bias analyses in Experiments 3-5). Rest trials were presented in two blocks of 6 trials at jittered time

points within each miniblock. Further, a high-resolution T1-weighted anatomical image was acquired (HiRes MP-RAGE, 160 sagittal slices, voxel size = 1 mm^3) after the functional runs. The total time in the scanner was around 55 minutes.

5.1.6 Data analysis

Data were preprocessed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/). Scans were realigned to the anterior commissure, spatially normalised to 2 mm^2 isotropic voxels using the parameters derived from the segmentation of each participant's T1-weighted image, and smoothed with a Gaussian kernel of 8 mm full-width at half maximum. At the single-subject level, event onsets from 25 conditions (4 vocalisation types x 6 speakers + vigilance task) and 6 movement regressors of no interest were modelled as instantaneous and convolved with the canonical hemodynamic response function. Rest trials were not modelled and were thus used as an implicit baseline. First-level models were masked with subject-specific binarised grey matter masks, created from the subject-specific segmented grey matter (smoothed with a Gaussian kernel of 8 mm full-width at half maximum to match the smoothing kernel of the functional data). A 2 (listener group) x 4 (vocalisation type) x 6 (speaker) ANOVA was conducted at the group level in GLM flex fast2 (http://mrtools.mgh.harvard.edu/) using first-level contrast images of each individual condition compared with the implicit baseline. GLM Flex fast2 was used as it uses partitioned error terms and can be used to run full-factorial models with more than 2 within-subject factors, while in-built ANOVA tools in SPM only allow for a pooled error term across all within-subject factors. All results of the functional runs in the experiment are reported at an uncorrected voxel height

threshold of $p < .001$, with FWE (family-wise error) cluster extent correction (number of voxels is dependent on the number and size of cluster, as implemented in bspmview [http://www.bobspunt.com/bspmview/]). The anatomical locations of significant clusters (at least 8 mm apart) were labelled using the Anatomy Toolbox (version 18; Eickhoff et al., 2005).

5.1.7 Results

Second-level 2 x 4 x 6 ANOVA: Effects of listener group, speaker and vocalisation type

The main effect of speaker gave rise to clusters in the temporal lobes (**Figure 18**a). The main effect of listener group gave rise to a cluster of activation in right superior and medial frontal gyri (**Figure 18**b). For the main effect of vocalisation type, clusters in bilateral STG, bilateral inferior frontal gyrus (pars triangularis, IFG) and bilateral inferior occipital gyrus among others were found (**Figure 18**c). For the interactions between group and vocalisation type, activations in right IFG, right insula lobe and left STG were found (**Figure 18**d). Auditory areas in the temporal lobes, extending anteriorly into the bilateral temporal poles, were found for the interaction between vocalisation type and speaker (**Figure 18**e). A small cluster in left fusiform gyrus was found for the interaction between listener group and speaker (**Figure 18**f). Notably, effects of speaker and vocalisation as well as the interaction between speaker and vocalisation overlap with each other and also with bilateral TVAs (Pernet et al., 2015; see **Figure 19**). For a full list of peak and sub-peak voxels for all significant effects, see **Table 7**.

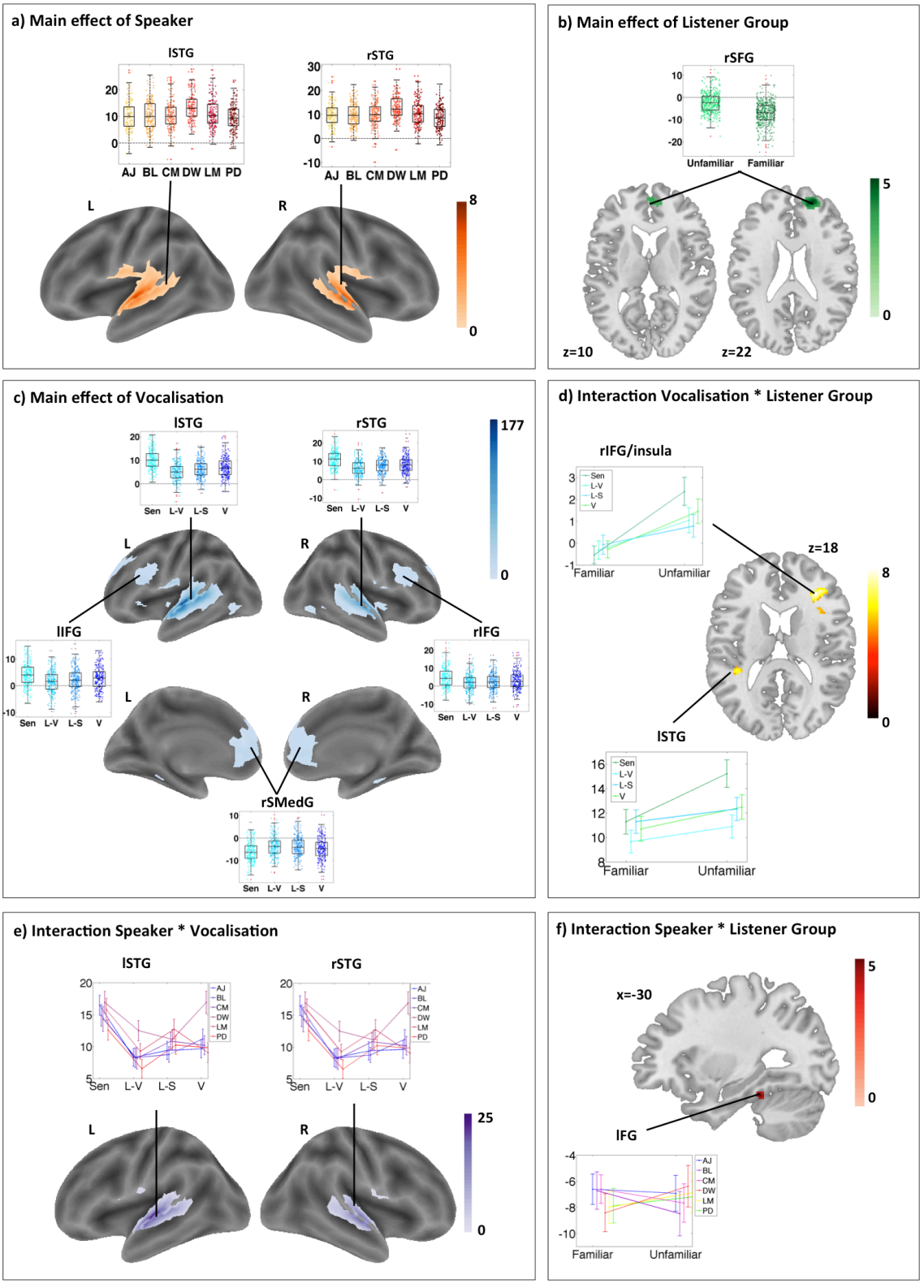

Figure 18 Results of the univariate analysis (peak threshold of p = 0.001 with FWE cluster correction). Parameter estimates are displayed in the y-axis of each line plot. Data points in boxplots represent mean parameter estimates per condition based on the first-level models. Sen = sentences, L-V = volitional laughter, L-S = spontaneous laughter, V = Vowels.

Table 7 Results of the univariate analysis at peak threshold of p = 0.001 and FWE cluster correction. Local maxima separated by more than 20 mm are listed.

				MNI coordinates		
Main Effects and Interactions	**Region Name**	**No of Voxels**	**F/t**	**x**	**y**	**z**
Vocalisation Type	L Middle Temporal Gyrus	5599	177.78	-58	-18	0
	L Temporal Pole		85.87	-52	8	-12
	L Fusiform Gyrus		14.78	-34	-32	-10
	R Superior Temporal Gyrus	4093	145.61	60	-4	-4
	R Middle Temporal Gyrus		64.93	54	-34	6
	R Temporal Pole		54.30	50	12	-16
	L Superior Medial Gyrus	5955	26.23	-6	52	8
	L Superior Frontal Gyrus		22.04	-18	46	42
	R Mid Orbital Gyrus		19.00	12	42	2
	L IFG (p. Triangularis)	1200	25.26	-44	16	24
	R IFG (p. Triangularis)	1298	21.92	50	24	26
	R Parahippocampal Gyrus	310	14.85	36	-36	-10
	L Inferior Temporal Gyrus	42	14.40	-54	-16	-24
	L Inferior Occipital Gyrus	188	11.67	-44	-72	-4
	L Cerebellum (VII)	125	11.49	-14	-76	-34
	R Inferior Occipital Gyrus	258	10.26	44	-68	-8
	R Inferior Occipital Gyrus		8.60	34	-86	-6
	Left Hippocampus	55	10.20	-14	-28	0
	L Precuneus	56	9.27	-12	-54	16
	R Insula Lobe	24	8.37	28	16	-14
	R Precuneus	33	8.22	16	-50	18
	R IFG (p. Orbitalis)	33	7.93	30	28	0
Speaker	R Superior Temporal Gyrus	3340	88.01	60	-4	-4
	R Temporal Pole		12.07	50	12	-14
	L Superior Temporal Gyrus	3570	76.84	-60	-14	4
	L Temporal Pole		10.81	-52	8	-12
Group	R Superior Frontal Gyrus	299	5.21	20	54	22
	R Superior Medial Gyrus	299	3.62	2	56	10
Group * Vocalisation	L IFG (p. Triangularis)	194	9.54	32	34	16
	R Insula Lobe	32	7.88	36	14	18
	L Superior Temporal Gyrus	44	7.59	-40	-38	16
Group * Vocalisation	L Fusiform Gyrus	25	5.25	-30	-36	-22
Speaker * Vocalisation	L Middle Temporal Gyrus	3066	25.08	-60	-14	0
	L Temporal Pole	3066	6.41	-52	8	-12
	R Superior Temporal Gyrus	2651	22.71	62	-8	-4
	R Temporal Pole	2651	5.13	52	10	-12

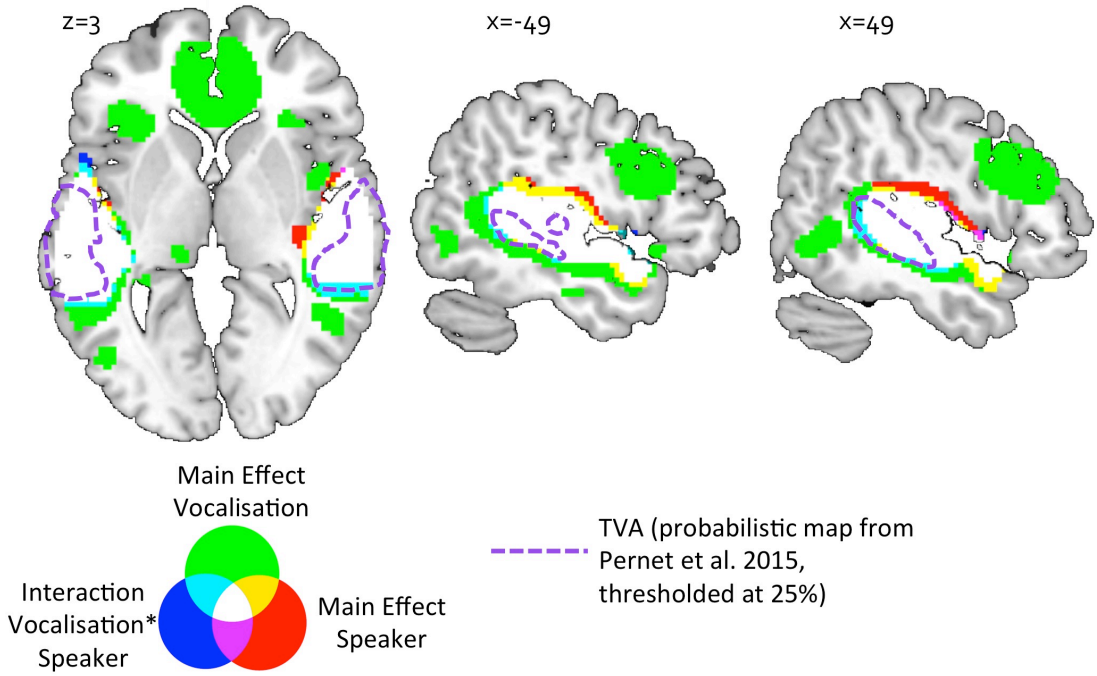

Figure 19 Overlays of activations for main effect of speaker, main effect of vocalisation and interaction between speaker*vocalisation

Group effects by vocalisation

Table 8 Results of the listener group effects by vocalisation, thresholded at p = .001 with a cluster extend of k=211 vowels. Local maxima separated by more than 20 mm are listed.

				MNI coordinates		
T-Test	**Region Name**	**No of Voxels**	**F/t**	**x**	**y**	**z**
Sentences	R Superior Frontal Gyrus	213	-5.96	20	54	22
unfamiliar > familiar	R Superior Frontal Gyrus	251	-5.10	16	34	40
	L Superior Frontal Gyrus	607	-4.43	-14	50	24
	L Middle Frontal Gyrus	607	-4.28	-22	20	40
	L ACC	607	-3.80	-2	20	36
Vowels	-					
unfamiliar > familiar						
Laughter$_V$	R Superior Frontal Gyrus	267	4.37	22	54	22
unfamiliar > familiar	R Superior Medial Gyrus	267	3.65	4	54	6
Laughter$_S$	-					
unfamiliar > familiar						

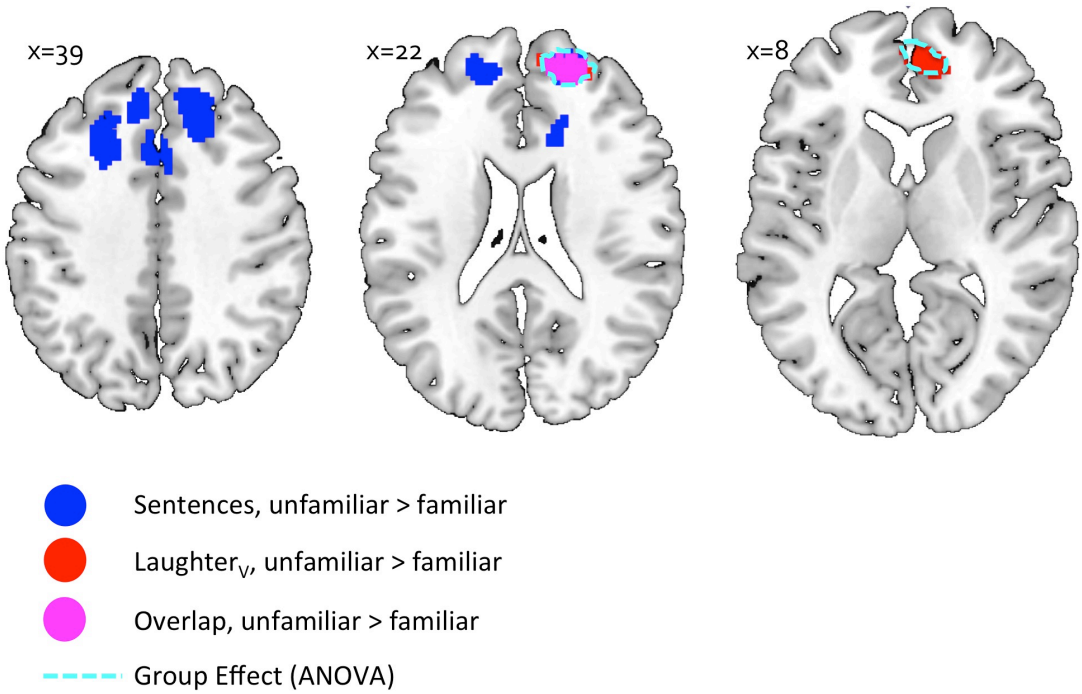

Figure 20 Overlays of activations for listener group effects by vocalisation.

For the vocalisations used in this study, speaker recognition (which may aid speaker discrimination) may have been differentially difficult across vocalisations. This may thus have resulted in distinct patterns of neural activation dependent on vocalisation type. To explore whether neural activity differed between listeners groups for specific vocalisations, four two-sample t-tests, one for each vocalisation type, were thus performed in SPM8. Results were thresholded at peak threshold of p = .001 with a cluster extend threshold of k = 211 voxels (the equivalent of the FWE corrected cluster threshold for the main effect of group in the ANOVA reported above). Clusters are illustrated in **Figure 20** (see **Table 8** for details of local peaks and sub-peaks). Notably, all vocalisations showed increased activation on SFG for unfamiliar listeners compared to familiar listeners (not shown), although only the clusters for Sentences and Laughter$_V$ survived FWE cluster correction. No activations showing increased activation for familiar over unfamiliar listeners were found.

5.1.8 Discussion

The current study set out to explore differences in voice processing in familiar and unfamiliar listeners in the context of variable vocal signals. As expected, speaker identity modulated activation in bilateral temporal cortices. It should, however, be noted that the main effects of speaker, vocalisation and the interaction of these two factors all resulted in activations of bilateral temporal cortices that overlap a) with each other and b) with bilateral TVAs (Pernet et al., 2015). The activation found for the main effect of speaker can thus not be exclusively attributed to speaker identity processing but may related to more general processes that are part of voice perception.

The overlap with TVAs was found for comparisons between conditions that were exclusively vocal in nature. TVAs are localised using a range of human vocal sounds (including speech and emotional vocalisations) contrasted with other environmental sounds (e.g. bells, animal calls, engine sounds, see Belin & Grodbras, 2010; Beling, Zatorre, Ahad & Pike, 2000; Belin, Zatorre & Ahad, 2002; Bestelemeyer, Belin & Grosbras, 2011; Pernet et al., 2015). This approach is problematic since 1) the auditory categories are far more diverse in the environmental sound category compared to the vocal category and 2) vocal and environmental sounds have not been fully matched to each other in their acoustic properties. It has indeed been suggested that activation of the TVAs may not represent voice-selective processing but may rather be a result of differences in acoustic properties or acoustic complexity of auditory signals in general – Leaver and Rauschecker (2010) show that a large part of the variance in the data underlying TVA activations can be accounted for by simple acoustic properties of the stimuli. The current results support this notion, suggesting

that TVAs in fact generally respond to differences in auditory signal properties, be that acoustic features or low-level auditory object processing and thus may not be voice-selective. This is to some extent different to face perception in FG, where increased activation for (upright) faces can be observed even when contrasted with inverted faces (perfect match for low-level visual features, but processed differently to upright faces; Kanwisher, Tong & Nakayama, 1998). Although note that the FG may also respond preferentially to visual objects a person has is highly familiar with (see evidence for FG activation for viewing, for example, cars in car experts; Gauthier, Skudlarski, Gore & Anderson, 2000). The 'selective' activation in both FG (and TVAs) may thus at least partly reflect an expertise response since faces (and voices) are frequently encountered and highly salient in social encounters.

Notably, for the effect of vocalisation parameter estimates were highest for sentences compared to all other vocalisations. This pattern holds within the interaction of speaker and vocalisation as well: while different voices show different profiles across the vocalisations, parameter estimates are overall highest for sentences. This may reflect the fact that in contrast to $Laughter_V$, $Laughter_S$ and Vowels, sentences also include meaningful linguistic content that is processed in STG/STS (Scott, Blank, Rosen & Wise, 2000). Arguably, this thus allows for more elaborate computations to be performed, hence potentially a BOLD response that is more extensive and greater in magnitude (but see Section 1.5.3.1 for a different interpretation of an increase in BOLD response in STG with regard to typical and less typical voice by Latinus et al., 2013).

For the main effect of listener group, activations were found in the right superior frontal gyrus, extending into the right superior medial gyrus. Previous studies

on voice processing also report increased activation in right superior frontal cortices for unfamiliar listeners compared to familiar listeners (e.g. Stevens, 2004; Giraud & von Kriegstein, 2004). Specifically, Giraud and von Kriegstein (2004) report similar findings for listeners that were presented with familiar and unfamiliar voices. The authors argue these differences may be a result of varying task demands for the processing of familiar and unfamiliar voices. In the current study, vocalisation-specific t-tests for listener group show that the pattern of activation is similar for all vocalisations: while differences between unfamiliar and familiar listeners appear to be largest for sentences, followed by $Laughter_V$, all vocalisations show (subthreshold) clusters in prefrontal cortices – but, crucially, not in anterior temporal lobes which have been associated with speaker identity processing. The activations for the listener group effect in Sentences include anterior cingulate cortex (ACC) in addition to bilateral dorsolateral prefrontal cortex (DLPFC). This frontal network has been strongly associated with cognitive control, with DLPFC being linked to the implementation of cognitive control and ACC being linked to performance monitoring (MacDonald, Cohen, Stenger & Carter, 2000). These analyses are thus still in line with an interpretation that these effects reflect differences in task difficulty or engagement with the stimuli between listeners groups rather than stimulus driven differences in, for example, the representation of speaker identity.

For the main effect of vocalisation, a range of brain region outside of sensory cortices were activated, including bilateral IFG as well as a large cluster in superior medial frontal gyrus. While IFG has been strongly associated with speech processing – which may be reflected in stronger activation for sentences that contain linguistic meaning compared to the other vocal signals (Hagoort, 2005), these regions have also

been associated with the frontal-opercular executive processing network (Geranmayeh, Brownsett & Wise, 2014). This network has been shown to be upregulated in the context of higher task demands, reflecting effortful processing. This may reflect the differential difficulty of extracting speaker identity from these different vocalisations: it is interesting to note that superior medial gyrus is the only region that was activated *less* for sentences compared to the other vocal signals. Arguably, discriminating speaker from full sentences is easier compared to nonverbal signals (see for example, performance of > 90% accuracy for sentences, Van Lancker & Kreiman, 1987; Reich & Duke, 1989; Wester, 2012) and task demands are therefore lower for speech signals in the context of this study. Left IFG was also found for the interaction between vocalisation type and listener group – this interaction is driven by stronger activation for sentences in unfamiliar listeners compared to the other vocalisation, while this pattern is not as apparent in familiar listeners. This cluster in IFG may similarly reflect processes relating to task demands: since speaker information is relatively easily processed from sentences, it may be far more difficult for the other vocalisations, especially in unfamiliar voices. These differences may therefore modulate attention and effort in this listener group. A small cluster in left FG was found for the interaction between speaker and listener group. FG has been associated with familiar voice processing: Some authors have argued that argue when processing speaker identity-related information from vocal signals, representations of this speaker's face are automatically co-activated, aiding amodal person recognition (Giraud & von Kriegstein, 2004; von Kriegstein et al., 2006; von Kriegstein et al. 2007). The current pattern of activations (see plot in **Figure 18**f) does not allow for a simple interpretation of this result – familiar listeners were far more familiar with the

speakers' faces but the BOLD response does not clearly show greater activation for familiar compared unfamiliar listeners. Unfamiliar listeners were, however, also briefly presented with pictures of the speakers during the speaker recognition task that participants performed before being scanned, which in principle could have lead to this complex pattern of activations.

The results of this fMRI study thus suggest that listener groups did process the vocal sounds and speaker identity differently, although most effects may be related to differences in task demands with the task being in general more difficult for unfamiliar listeners. Against predictions, no group differences were found in anterior temporal cortices that have been reported to the activated during the processing of identity related information in familiar (and depending on the task) unfamiliar listeners in previous studies. Previous studies contrasting familiar and unfamiliar voice processing have either asked participants to perform two different tasks, for example linguistic processing versus identity processing to probe identity processing, have included familiar and unfamiliar voices for the same listener group (Giraud & von Kriegstein, 2004; Nakamura et al., 2001; von Kriegstein, et al., 2006; von Kriegstein et al., 2007) or have trained one listener group on a set of voices initially unfamiliar to them and have used artificial morphed voices as stimuli (Latinus et al., 2011; Andics et al., 2010). In contrast to this, the current study used the same stimuli to control for acoustic differences across the familiar and unfamiliar listener groups, which further allowed to ensure that familiar listeners had been exposed to the voices in a relatively naturalistic context instead of artificial training environments (or using manipulated stimuli). These are factors that may have influenced the results: since both participant groups were engaged in the same task, group-based differences in magnitude in the BOLD

response may have been masked by, for example, additional task demands for unfamiliar listeners (see Giraud & von Kriegstein, 2004). The nature of familiarity of the current listener groups combined with the naturalistic but variable vocal signals (see Experiments 5 and 6) may furthermore explain the lack of replication of previous findings. Additionally, as has been shown in previous behavioural experiments (Experiments 4 and 5), listeners struggle to reliably link different vocalisations to a single speaker, even when they are familiar with the speaker. Thus, variability within the current stimuli may have further affected results (although note that even for sentences no group differences in temporal lobes was found). Future analyses and studies should attempt to use more sensitive, multivariate approaches (see Formasino et al., 2008 or Evans & Davis, 2015 for methods) to identity pattern-based representations of vocalisations and speakers. With such approaches neural representations of speaker within and across vocalisation type could be identified, compared and contrasted. This would thus allow insights into how familiarity and variability in vocal signals affect the fidelity (or indeed presence or absence) of speaker specific representations on a neural level.

6 General discussion and future directions

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 studies to date have almost exclusively used speech stimuli produced in a volitional, highly controlled manner and in neutral, modal voice. This thesis has started to address the gap in the literature by examining voice perception in the context of sex identification and speaker discrimination tasks across a range of verbal and nonverbal vocalisations representative of vocal flexibility (exemplified here by the degree of volitional control over their production [Experiments 1-5 and 7] or the presence and absence of voicing [Experiment 6]).

The results of these experiments show that while listeners can display relatively high accuracy in extracting speaker characteristics from volitional vocalisations, they have a more limited ability to do this for spontaneous vocal signals: performance was impaired for sex identification and speaker discrimination tasks when performed on spontaneous laughter and spontaneous crying, while performance for vowels and volitional laughter was relatively unaffected. Similarly to spontaneous vocalisations, the absence of voicing during whispered speech significantly impaired listeners' performance for speaker discrimination. Thus, the extraction of speaker characteristics from vocal signals that diverge from voiced volitional vocal signals seems to be more challenging for listeners. Further, for speaker discrimination, when listeners were required to make judgements across vocalisation types, performance became highly unreliable – listeners were for example unable to link the laughter produced by a speaker to the vowels produced by the same speaker. While in general an advantage was observed for familiar listeners in the processing of

indexical speaker properties, they were nonetheless affected by vocal flexibility. An fMRI study showed that when discriminating speakers based on varied vocal signals, unfamiliar listeners more strongly recruit brain regions and networks implicated in executive function, such as the frontal-opercular network, compared to familiar listeners. This indicates that task difficulty and processing demands differ between the two groups. Intriguingly, and in contrast to previous studies (Andics et al. 2010; Formisano et al. 2008; Latinus et al., 2013), the current analyses did not show any group differences in voice-sensitive cortical regions (i.e. auditory cortex and the superior temporal lobes) that could be linked to differential perceptual processing of familiar and unfamiliar voices.

6.1 Aspects of familiarity affecting the perception of speaker characteristics

Throughout this thesis, different types of familiarity or expertise with a stimulus have been considered as potential explanations for advantages or impairments in listener's performance. Previous research has already shown that familiarity with a stimulus type affords listeners processing advantages: musicians are better than non-musicians at identifying individuals from their musical performances (Koren & Gingras, 2014 for harpsichord performances), individuals familiar with a television show are able to identify the original theme tune in a set of pitch-shifted theme tunes (Schellenberg & Trehub, 2003). There are further countless examples of anecdotal evidence of farmers recognising their animals by their calls or car enthusiasts identifying cars from engine sounds (while non-experts fail to do so). Listeners have also been shown to be more accurate at discriminating and recognising speakers when presented with speech in a

language they are familiar with, versus an unfamiliar language Perrachione et al., 2009; Perrachione et al., 2011). Thus expertise, through prolonged exposure and engagement with a stimulus appear to affect performance for identity processing. In the context of the stimuli used in this thesis, spontaneous vocalisations can be considered to occur relatively rarely compared to spoken signals. This relative lack of familiarity or expertise in the processing these vocal signals in listeners may thus underlie the impairment in the extraction of speaker characteristics from spontaneous (or whispered) vocal signals reported in this thesis. Future work would need to formally assess how frequently spontaneous vocalisations are heard and whether performance in the extraction of speaker characteristics is also impaired for other less frequently encountered (volitional) vocalisations, such as singing. Learning studies could further assess whether, if trained on either volitional or spontaneous vocalisations that are matched for familiarity, subsequent increases in the discrimination or recognition of speakers (compared with no-training control conditions) are similar or different across vocalisations. Similar increases in performance would suggest that an expertise effect is present, while differential increases dependent on the training condition would suggest that other factors might be affecting performance.

Another aspect of familiarity that may affect participants' responses pertains to the context in which familiarity is acquired. Familiarity with a speaker is a specific type of expertise, and familiarity with a person can take many different forms – celebrities, colleagues, close friends, relatives and partners are all in some way familiar to a listener. In contrast to familiarity with a stimulus type, these different types of familiarity with a person are not only marked by differential exposure with a

speaker's full vocal inventory. With different types of interpersonal relationships, socio-emotional attributes become associated with hearing a familiar voice (see McGettigan, 2015; Sidtis & Kreiman, 2012; Suguira, 2014 for discussions): Hearing the voice of a partner after a prolonged absence will have a different effect on the listener than hearing the highly familiar voice of Meryl Streep while watching "Mamma Mia!" again after 3 years. How these socio-emotional aspects of personal familiarity may affect voice processing remains a largely open question. The challenge of future research will be to more thoroughly address and describe the effects of different types of familiarity on task performance. Specifically, familiarity in the sense of mere expertise through exposure and engagement needs to be delineated and disentangled from familiarity additionally involving the formation of socio-emotional associations with a stimulus.

6.2 Identifying the underlying mechanisms of speaker processing in the context of vocal flexibility

In this thesis, the extraction of speaker characteristics has been shown to be significantly impaired for spontaneous vocalisations such as authentic laughter and crying. It has been hypothesized that acoustic features of the signal may underpin this effect: Acoustic cues to speaker identity or speaker sex may be absent within these vocalisations, since during production, authentic emotional content may be encoded preferentially to cues to speaker identity. Alternatively, it was hypothesized that spontaneous vocalisations are not only produced relatively infrequently but also (possibly as a result) also occupy an acoustic space that is only rarely encountered and is thus relatively unfamiliar to listeners, impairing speaker identity processing. From

the current studies it is unclear which hypothesis is more likely to explain the findings reported above. Future studies could resolve this issue, for example, by shifting spontaneous vocalisations into the acoustic space of volitional vocalisations (e.g. lowering the fundamental frequency) and *vice versa* in studies of speaker perception. It could be predicted that if the impairment of performance is due to being relatively unfamiliar with the acoustic space for spontaneous vocal signals, performance should be better for shifted spontaneous vocalisations, while it should decrease for volitional vocalisations that have been shifted into the acoustic space of spontaneous vocalisations. If, however, cues to speaker identity are absent in spontaneous vocalisations, shifting, for example, the pitch of spontaneous vocalisations down should not improve performance. From a perception point of view, it was furthermore hypothesized that, while all necessary cues may be encoded in spontaneous and volitional vocalisations alike, the processing of the authentic emotional content present in spontaneous vocalisations may be prioritized over identity-related information (Goggin et al., 1991). Future work should therefore directly investigate modulations of attention for spontaneous and volitional vocal signals to explore potential differences between these types of affective information.

In Experiment 3 and 4 it was attempted to explain patterns of behavioural responses through within-pair acoustic dissimilarity of vocal signals. It was hypothesized that the more acoustically dissimilar pairs of vocal signals from a single speaker are, the more difficult it should be for listeners to assign them to a single speaker (in the absence of familiarity with any of the speakers). This hypothesis was not clearly confirmed since measures of acoustic dissimilarity only predicted a negligible amount of variance in logistic regression models. Future work should

further explore whether acoustic or perceptual descriptions of the stimuli (using multidimensional scaling approaches, see Baumann & Belin, 2008) can shed further light on how variability in vocal signals affects performance.

6.3 Interactions between affect and identity: implications for the pathways in Belin et al.'s (2004) model of voice processing

Belin et al.'s (2004) model of voice processing proposes the hierarchical processing of human vocal signals along three partially independent pathways (see Section 1.5). Interactions during the processing of vocal signals have already been shown for speech and identity pathways, with, for example, familiar speakers being more intelligible and speakers being more easily identified when speaking a language with which the listeners are familiar (Nygaard & Pisoni, 1998; Pisoni, 1993; Perrachione et al., 2009; Perrachione et al., 2011; see Section 4.3 for an overview). Interactions between emotion and identity (and speech) processing have already been shown in the face perception literature (Schweinberger & Soukup, 1998). The findings presented in this thesis provide novel empirical evidence for striking interactions between affect and identity processing pathways for vocal stimuli – crucially, with specifically *spontaneous* emotional information impairing identity processing. Future research will need to determine whether further interactions, for example, between speech and emotion pathways may be present: Highly emotional speech may, for example, be less intelligible. Studies of speech prosody have already reported generally lower emotion recognition rates than studies of non-verbal vocalisations

(see Section 1.5.2) – further work should directly test whether the presence of speech during emotional vocal displays impairs emotion recognition.

6.4 Individual differences

For most experiments in this thesis, individual differences in listener performance were apparent. While individual differences in face recognition have been widely assessed (e.g. Dennett, McKone, Edwards & Susilo, 2012; Hedley, Brewer & Young, 2011; Wang, Li, Fang, Tian & Liu, 2012), very little research exists on individual differences in voice processing. Recently, a validated test of voice memory, assessing the ability of listeners to memorise and recognise a set of unfamiliar voices has reported substantial variability in listeners' ability to recognise voices, ranging from severely impaired performance for a phonagnosic patient to highly proficient listeners scoring far above average (Aglieri, Watson, Pernet, Latinus, Garrido & Belin, 2016). While this research into individual difference in voice identity processing has thus established that there are individual differences that appear to be relatively specific to voice processing, future work will need to determine what the specific physiological and psychological factors are that may underlie such differences in performance across a number of tasks in healthy listeners as well as in special populations. It also remains to be determined whether individual differences in the processing of identity information are modality-specific or amodal (see Lewis, Lefevre & Young, 2016 for potential methodological approaches).

6.5 Looking across different subfields

While a relatively large body of work on identity processing from vocal signals exists, reporting a wealth of findings, studies are spread across several research traditions (applied earwitness research versus experimental/psychoacoustic approaches) with only limited synthesis of findings occurring: studies have already shown that stimulus type, task type, retention interval and type of exposure to a voice at test or prior to the test have all been shown to have an impact, at times substantial, on task performance (e.g. Bricker & Pruzanksy, 1966; Kerstholt, Jansen, Van Amelsvoort & Broeders, 2004; Orchard & Yarmey, 1995; Schweinberger et al., 1997; Yarmey &Matthys, 1992; Yarmey, Yarmey, Yarmey, 1994). Methods, tasks and stimuli differ, however, vastly between research traditions, making it at times difficult to evaluate how an effect reported for applied earwitness studies (using one shot line-up approaches, long, relatively uncontrolled stimulus materials, see Yarmey, 1995 for an overview) would translate to more controlled experimental approaches (and *vice versa*). Even within experimental approaches to voice processing, task type and familiarity are closely linked: For speaker recognition and speaker identification tasks familiar voices are used, while for speaker discrimination mainly unfamiliar voices are used. Future studies should aim at synthesizing the divergent literature into a common framework. They will thus be able continue to create a more comprehensive picture of voice processing by addressing novel questions through the use of a wider range of tasks and stimuli, testing existing frameworks and if appropriate proposing new models of voice processing.

6.6 Conclusion

Overall this thesis highlights that while vocal signals can encode a wealth of cues to identity our full 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. The presence, absence or modulation of salient stimulus properties, such as limited source information in whispered speech and drastically modulated acoustic signals for spontaneous vocalisations, poses additional challenges for the extraction of speaker characteristics for familiar and listeners alike. The findings of this thesis thus put into perspective our ability to extract speaker characteristics from vocal signals, calling for a more nuanced as well as more comprehensive approach to voice processing that accounts for vocal flexibility as well as stimulus and listener characteristics.

7 References

Abberton, E., & Fourcin, A. J. (1978). Intonation and speaker identification. *Language and Speech*, *21*(4), 305-318.

Ackermann, H., Hage, S. R., & Ziegler, W. (2014). Brain mechanisms of acoustic communication in humans and nonhuman primates: An evolutionary perspective. *Behavioral and Brain Sciences*, *37*(06), 529-546.

Aglieri, V., Watson, R., Pernet, C., Latinus, M., Garrido, L., & Belin, P. (2016). The Glasgow Voice Memory Test: Assessing the ability to memorize and recognize unfamiliar voices. *Behavior research methods*, 1-14.

Andics, A., McQueen, J. M., Petersson, K. M., Gál, V., Rudas, G., & Vidnyánszky, Z. (2010). Neural mechanisms for voice recognition. *Neuroimage*, *52*(4), 1528-1540.

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

Bachorowski, J. A. (1999). Vocal expression and perception of emotion. *Current directions in psychological science*, *8*(2), 53-57.

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., & Owren, M. J. (2001). Not all laughs are alike: Voiced but not unvoiced laughter readily elicits positive affect. *Psychological Science*, *12*(3), 252-257.

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.

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614.

Barrett, A. M., Crucian, G. P., Raymer, A. M., & Heilman, K. M. (1999). Spared comprehension of emotional prosody in a patient with global aphasia. *Cognitive and Behavioral Neurology*, *12*(2), 117-120.

Barrett, L. F., & Kensinger, E. A. (2010). Context is routinely encoded during emotion perception. *Psychological Science*, *21*(4), 595-599.

Bartle, A., & Dellwo, V. (2015). Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech. *International Journal of Speech, Language & the Law*, *22*(2), 229-24.

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

Belin, P., & Grosbras, M. H. (2010). Before speech: cerebral voice processing in infants. *Neuron*, *65*(6), 733-735.

Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport*, *14*(16), 2105-2109.

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.

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309-312.

Belin, P., Zatorre, R. J., & Ahad, P. (2002). Human temporal-lobe response to vocal sounds. *Cognitive Brain Research*, *13*(1), 17-26.

Bench, J., Kowal, Å., & Bamford, J. (1979). The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British Journal of Audiology*, *13*(3), 108-112.

Bestelmeyer, P. E., Belin, P., & Grosbras, M. H. (2011). Right temporal TMS impairs voice detection. *Current Biology*, *21*(20), R838-R839.

Boersma, P., & Wennink, D. (2010). Praat: Doing phonetics by computer [Software]. Available from http://www.praat.org.

Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America*, *40*(6), 1441-1449.

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.

Bryant, G. A., & Barrett, H. C. (2008). Vocal emotion recognition across disparate cultures. *Journal of Cognition and Culture*, *8*(1), 135-148.

Bryant, G. A., Fessler, D. M., Fusaroli, R., Clint, E., Aarøe, L., Apicella, C. L., ... & De Smet, D. (2016). Detecting affiliation in colaughter across 24 societies. *Proceedings of the National Academy of Sciences*, *113*(17), 4682-4687.

Brück, C., Kreifelts, B., & Wildgruber, D. (2011). Emotional voices in context: a neurobiological model of multimodal affective information processing. *Physics of Life Reviews*, *8*(4), 383-403.

Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication*, *46*(3), 252-267.

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

Cappa, S. F., Guidotti, M., Papagno, C., & Vignolo, L. A. (1987). Speechlessness with occasional vocalisations after bilateral opercular lesions: a case study. *Aphasiology*, *1*(1), 35-39.

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

Chartrand, J. P., Peretz, I., & Belin, P. (2008). Auditory recognition expertise and domain specificity. *Brain Research*, *1220*, 191-198.

Coleman, R. O. (1971). Male and female voice quality and its relationship to vowel formant frequencies. *Journal of Speech, Language, and Hearing Research*, *14*(3), 565-577.

Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F. (2008). Audio-visual integration of emotion expression. *Brain Research*, *1242*, 126-135.

Darwin, C. (1872). The expression of the emotions in man and animals. *London, UK: John Marry*.

Davila-Ross, M., Owren, M. J., & Zimmermann, E. (2010). The evolution of laughter in great apes and humans. *Communicative & Integrative Biology*, *3*(2), 191-194.

De Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition & Emotion*, *14*(3), 289-311.

Dennett, H. W., McKone, E., Edwards, M., & Susilo, T. (2012). Face aftereffects predict individual differences in face recognition ability. *Psychological Science*, *23*(11), 1279-1287.

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

Eibl-Eibesfeldt, I. (1972). Similarities and differences between cultures in expressive movements. In Hinde, R. (Ed.), *Non-verbal communication*, Cambridge: Cambridge University Press. 297—314.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *Neuroimage*, *25*(4), 1325-1335.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3-4), 169-200.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*(2), 124.

Ellis, H. D., Jones, D. M., & Mosdell, N. (1997). Intra-and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology*, *88*(1), 143-156.

Evans, S., & Davis, M. H. (2015). Hierarchical organization of auditory and motor representations in speech perception: evidence from searchlight similarity analysis. *Cerebral Cortex*, *25*(12), 4772-4788.

Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., Andre, E., Busso, C., et al. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, 7(2), 190–202.

Fairbanks, G. (1960). The rainbow passage. *Voice and articulation drillbook*, 2nd edition. New York: Harper & Row. 124-139.

Fant, G. (1960). Acoustic theory of speech production. The Hague: Mouton.

Fitch, W. T. (2010). *The evolution of language*. Cambridge University Press.

Formisano, E., De Martino, F., Bonte, M., & Goebel, R. (2008). "Who" Is Saying" What"? Brain-Based Decoding of Human Voice and Speech. *Science*, *322*(5903), 970-973.

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J. R., ... & Duchaine, B. (2009). Developmental phonagnosia: a selective deficit of vocal identity recognition. *Neuropsychologia*, *47*(1), 123-131.

Gauthier, I., Skudlarski, P., Gore, J. C., & Anderson, A. W. (2000). Expertise for cars and birds recruits brain areas involved in face recognition. *Nature neuroscience*, *3*(2), 191-197.

Geranmayeh, F., Brownsett, S. L., & Wise, R. J. (2014). Task-induced brain activity in aphasic stroke patients: what is driving recovery? *Brain*, *137*(10), 2632-2648.

Gingras, B., Lagrandeur-Ponce, T., Giordano, B. L., & McAdams, S. (2011). Perceiving musical individuality: performer identification is dependent on performer expertise and expressiveness, but not on listener expertise. *Perception*, *40*(10), 1206-1220.

Gobl, C., & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, *40*(1), 189-212.

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

Gonzalez, J., & Oliver, J. C. (2005). Gender and speaker identification as a function of the number of channels in spectrally reduced speech. *The Journal of the Acoustical Society of America*, *118*(1), 461-470.

Gorno-Tempini, M. L., Price, C. J., Josephs, O., Vandenberghe, R., Cappa, S. F., Kapur, N., ... & Tempini, M. L. (1998). The neural systems sustaining face and proper-name processing. *Brain*, *121*(11), 2103-2118..

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*, *9*(9), 416-423

Hall, D. A., Haggard, M. P., Akeroyd, M. A., Palmer, A. R., Summerfield, A. Q., Elliott, M. R., ... & Bowtell, R. W. (1999). Sparse temporal sampling in auditory fMRI. *Human brain mapping*, *7*(3), 213-223.

Hailstone, J. C., Crutch, S. J., Vestergaard, M. D., Patterson, R. D., & Warren, J. D. (2010). Progressive associative phonagnosia: a neuropsychological analysis. *Neuropsychologia*, *48*(4), 1104-1114.

Hedley, D., Brewer, N., & Young, R. (2011). Face recognition performance of individuals with Asperger syndrome on the Cambridge Face Memory Test. *Autism Research*, *4*(6), 449-455.

Heilman, K. M., Scholes, R., & Watson, R. T. (1975). Auditory affective agnosia. Disturbed comprehension of affective speech. *Journal of Neurology, Neurosurgery & Psychiatry*, *38*(1), 69-72.

Heim, E., Knapp, P. H., Vachon, L., Globus, G. G., & Nemetz, S. J. (1968). Emotion, breathing and speech. *Journal of Psychosomatic Research*, *12*(4), 261-274.

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, *4*(4), 131-138.

Homa, D., Sterling, S., & Trepel, L. (1981). Limitations of exemplar-based generalisation and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(6), 418-439.

Honorof, D. N., & Whalen, D. H. (2010). Identification of speaker sex from one vowel across a range of fundamental frequencies. *The Journal of the Acoustical Society of America*, *128*(5), 3095-3104.

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.

Imaizumi, S., Mori, K., Kiritani, S., Kawashima, R., Sugiura, M., Fukuda, H., Itoh, K., Kato, T., Nakamura, A., Hatano, K. and Kojima, S., 1997. Vocal identification of speaker and emotion activates different brain regions. *Neuroreport*, *8*(12), 2809-2812.

Izard, C. E. (1992). Basic emotions, relations among emotions, and emotion-cognition relations. Psychological Review, 99, 561-565.

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

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

Jürgens, U. (2009). The neural control of vocalisation in mammals: a review. *Journal of Voice*, *23*(1), 1-10.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialised for face perception. *The Journal of Neuroscience*, *17*(11), 4302-4311.

Kanwisher, N., Tong, F., & Nakayama, K. (1998). The effect of face inversion on the human fusiform face area. *Cognition*, *68*(1), B1-B11.

Kerstholt, J. H., Jansen, N. J., Van Amelsvoort, A. G., & Broeders, A. P. A. (2004). Earwitnesses: Effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology*, *18*(3), 327-336.

Kishon-Rabin, L., Amir, O., Vexler, Y., & Zaltz, Y. (2001). Pitch discrimination: Are professional musicians better than non-musicians? *Journal of Basic and Clinical Physiology and Pharmacology*, *12*(2), 125-144.

Koren, R., & Gingras, B. (2014). Perceiving individuality in harpsichord performance. *Individuality in music performance*, 84.

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

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

Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens*, *1*(1).

Kreiman, J., Gerratt, B. R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception. *Journal of Speech, Language, and Hearing Research*, *35*(3), 512-520.

Ladefoged, P., & Disner, S. F. (2012). *Vowels and consonants*. Chichester: John Wiley & Sons.

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

Latinus, M., Crabbe, F., & Belin, P. (2011). Learning-induced changes in the cerebral processing of voice identity. *Cerebral Cortex*, *21*(12), 2820-2828.

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.

Laukka, P., Juslin, P., & Bresin, R. (2005). A dimensional approach to vocal expression of emotion. *Cognition & Emotion*, *19*(5), 633-653.

Lavan, N., Lima, C. F., Harvey, H., Scott, S. K., & McGettigan, C. (2015). I thought that I heard you laughing: Contextual facial expressions modulate the perception of authentic laughter and crying. *Cognition and Emotion*, *29*(5), 935-944.

Lavan, N., & McGettigan, C. (2016). Increased discriminability of authenticity from multimodal laughter is driven by auditory information. *The Quarterly Journal of Experimental Psychology*, 1-30.

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Laugh like you mean it: authenticity modulates acoustic, physiological and perceptual properties of laughter. *Journal of Nonverbal Behavior*, *40*(2), 133-149.

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.

Leaver, A. M., & Rauschecker, J. P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *The Journal of Neuroscience*, *30*(22), 7604-7612.

Leopold, D. A., O'Toole, A. J., Vetter, T., & Blanz, V. (2001). Prototype-referenced shape encoding revealed by high-level aftereffects. *Nature Neuroscience*, *4*(1), 89-94.

Leopold, D. A., Rhodes, G., Müller, K. M., & Jeffery, L. (2005). The dynamics of visual adaptation to faces. *Proceedings of the Royal Society of London B: Biological Sciences*, *272*(1566), 897-904.

Lewis, G. J., Lefevre, C. E., & Young, A. W. (2016). Functional architecture of visual emotion recognition ability: A latent variable approach. *Journal of Experimental Psychology: General*, *145*(5), 589.

Liberman, A. M., Safford-Harris, K., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358.

Lloyd, E. L. (1938). The respiratory mechanism in laughter. *The Journal of General Psychology*, *19*(1), 179-189.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, *288*(5472), 1835-1838.

MacLarnon, A. M., & Hewitt, G. P. (1999). The evolution of human speech: The role of enhanced breathing control. *American Journal of Physical Anthropology*, *109*(3), 341-363.

MacSweeney, M., Campbell, R., Calvert, G. A., McGuire, P. K., David, A. S., Suckling, J., ... & Brammer, M. J. (2001). Dispersed activation in the left temporal cortex for speech-reading in congenitally deaf people. *Proceedings of the Royal Society of London B: Biological Sciences*, *268*(1466), 451-457.

Mannell, R. C., & McMahon, L. (1982). Humor as play: Its relationship to psychological well-being during the course of a day. *Leisure Sciences*, *5*(2), 143-155.

Martin, R. A., & Kuiper, N. A. (1999). Daily occurrence of laughter: Relationships with age, gender, and Type A personality. *Humor*, *12*(4), 355-384.

Mathias, S. R., & von Kriegstein, K. (2014). How do we recognise who is speaking. *Frontiers in Biosciences* (*6)*, 92-109.

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'Hello'? Personality impressions from brief novel voices. *PloS one*, *9*(3), e90779.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1-86.

McGettigan, C. (2014). The social life of voices: studying the neural bases for the expression and perception of the self and others during spoken communication. *Frontiers in Human Neuroscience*, *9*, 129-129.

McGettigan, C., & Scott, S. K. (2014). Voluntary and involuntary processes affect the production of verbal and non-verbal signals by the human voice. *Behavioral and Brain Sciences*, *37*(06), 564-565.

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 mentalising and sensorimotor systems in the evaluation of emotional authenticity. *Cerebral Cortex*, *25*, 246-257.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746-748.

McKeown, G., Sneddon, I., & Curran, W. (2015). Gender differences in the perceptions of genuine and simulated laughter and amused facial expressions. *Emotion Review*, *7*(1), 30-38.

Mullennix, J. W., Johnson, K. A., Topcu-Durgun, M., & Farnsworth, L. M. (1995). The perceptual representation of voice gender. *The Journal of the Acoustical Society of America*, *98*(6), 3080-3095.

Nakamura, K., Kawashima, R., Sugiura, M., Kato, T., Nakamura, A., Hatano, K., ... & Kojima, S. (2001). Neural substrates for recognition of familiar voices: a PET study. *Neuropsychologia*, *39*(10), 1047-1054.

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), 466-478.

Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The enigmatic temporal pole: a review of findings on social and emotional processing. *Brain*, *130*(7), 1718-1731.

Orchard, T. L., & Yarmey, A. D. (1995). The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology*, *9*(3), 249-260.

Orena, A. J., Theodore, R. M., & Polka, L. (2015). Language exposure facilitates talker learning prior to language comprehension, even in adults. *Cognition*, *143*, 36-40.

Osherson, D. N., & Smith, E. E. (1981). On the adequacy of prototype theory as a theory of concepts. *Cognition*, *9*(1), 35-58.

Owren, M. J., Amoss, R. T., & Rendall, D. (2011). Two organizing principles of vocal production: Implications for nonhuman and human primates. *American Journal of Primatology*, *73*(6), 530-544.

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.

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

Paulmann, S., & Uskul, A. K. (2014). Cross-cultural emotional prosody recognition: Evidence from Chinese and British listeners. *Cognition & Emotion*, *28*(2), 230-244.

Pell, M. D., Monetta, L., Paulmann, S., & Kotz, S. A. (2009). Recognizing emotions in a foreign language. *Journal of Nonverbal Behavior*, *33*(2), 107-120.

Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E., Watson, R. H., Fleming, D., Crabbe, F., Valdes-Soda, M., & Belin, P. (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. *Neuroimage*, *119*, 164-174.

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

Perrachione, T. K., Pierrehumbert, J. B., & Wong, P. (2009). Differential neural contributions to native-and foreign-language talker identification. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1950.

Perrodin, C., Kayser, C., Abel, T. J., Logothetis, N. K., & Petkov, C. I. (2015). Who is that? Brain networks and mechanisms for identifying individuals. *Trends in Cognitive Sciences*, *19*(12), 783-796.

Pisanski, K., Cartei, V., McGettigan, C., Raine, J., & Reby, D. (2016). Voice modulation: a window into the origins of human vocal control? *Trends in Cognitive Sciences*, *20*(4), 304-318.

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication*, *13*(1), 109-125.

Pollack, I., Pickett, J. M., & Sumby, W. H. (1954). On the identification of speakers by voice. *The Journal of the Acoustical Society of America*, *26*(3), 403-406.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3p1), 353.

Pourtois, G., Schwartz, S., Seghier, M. L., Lazeyras, F., & Vuilleumier, P. (2005). View-independent coding of face identity in frontal and temporal cortices is modulated by familiarity: an event-related fMRI study. *Neuroimage*, *24*(4), 1214-1224.

Provine, R. R. (2016). Laughter as an approach to vocal evolution: The bipedal theory. *Psychonomic Bulletin & Review*, 1-7.

Provine, R. R., & Yong, Y. L. (1991). Laughter: A stereotyped human vocalisation. *Ethology*, *89*(2), 115-124.

Puts, D. A., Hill, A. K., Bailey, D. H., Walker, R. S., Rendall, D. Wheatley, J. R., Welling, L. L. M., Dawood. K., Cárdenas, R., Burriss, R. P., Jablonski, N. G., Shriver, M. D., Weiss, D., Lameira, A. R., Apicella, C. L., Owren, M. J., Barelli, C., Glenn, M. E., & Ramos-Fernandez, G. (2016). Evidence for sexual selection on male vocal fundamental frequency in humans and other anthropoid primates. *Proceedings of the Royal Society of London B: Biological Sciences, 283,* 20152830.

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.

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

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), 651-666.

Rhodes, G., & Jeffery, L. (2006). Adaptive norm-based coding of facial identity. *Vision research*, *46*(18), 2977-2987.

Rohrer, J. D., Warren, J. D., & Rossor, M. N. (2009). Abnormal laughter-like vocalisations replacing speech in primary progressive aphasia. *Journal of the neurological sciences*, *284*(1), 120-123.

Ruch, W., & Ekman, P. (2001). The expressive pattern of laughter. In A. W. Kaszniak (Ed.), *Emotion, qualia, and consciousness*. Tokyo: Word Scientific Publisher. 426-443.

Russell, J. A. (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39, 1161–1178.

Rämä, P., Poremba, A., Sala, J. B., Yee, L., Malloy, M., Mishkin, M., & Courtney, S. M. (2004). Dissociable functional cortical topographies for working memory maintenance of voice identity and location. *Cerebral Cortex*, *14*(7), 768-780.

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

Sauter, D. A., & Scott, S. K. (2007). More than one kind of happiness: Can we recognize vocal expressions of different positive states?. *Motivation and Emotion*, *31*(3), 192-199.

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010b). Perceptual cues in nonverbal vocal expressions of emotion. *The Quarterly Journal of Experimental Psychology*, *63*(11), 2251-2272.

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010a). Cross-cultural recognition of basic emotions through nonverbal emotional vocalisations.*Proceedings of the National Academy of Sciences*, *107*(6), 2408-2412.

Scheiner, E., Hammerschmidt, K., Jürgens, U., & Zwirner, P. (2006). Vocal expression of emotions in normally hearing and hearing-impaired infants. *Journal of Voice*, *20*(4), 585-604.

Scherer, K. R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1), 227-256.

Scherer, K. R., Banse, R., & Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-cultural Psychology*, *32*(1), 76-92.

Scherer, K. R., Johnstone, T. & Klasmeyer, G. (2003). Vocal expression of emotion. In R. J. Davidson, K. R. Scherer, H. Goldsmith (Eds.). *Handbook of the Affective Sciences* (433–456). New York and Oxford: Oxford University Press.

Schweinberger, S. R., & Soukup, G. R. (1998). Asymmetric relationships among perceptions of facial identity, emotion, and facial speech. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(6), 1748.

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.

Scott, S. K., & Johnsrude, I. S. (2003). The neuroanatomical and functional organization of speech perception. *Trends in Neurosciences*, *26*(2), 100-107.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*(12), 2400-2406.

Scott, S. K., Sauter, D., & McGettigan, C. (2010). Brain mechanisms for processing perceived emotional vocalisations in humans. *Handbook of behavioral neuroscience*, *19*, 187-197.

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.

Shah, N. J., Marshall, J. C., Zafiris, O., Schwab, A., Zilles, K., Markowitsch, H. J., & Fink, G. R. (2001). The neural correlates of person familiarity. *Brain*, *124*(4), 804-815.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303-304.

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.

Skuk, V. G., & Schweinberger, S. R. (2014). Influences of fundamental frequency, formant frequencies, aperiodicity, and spectrum level on the perception of voice gender. *Journal of Speech, Language, and Hearing Research*, *57*(1), 285-296.

Spiegel, M. F., & Watson, C. S. (1981). Factors in the discrimination of tonal patterns. III. Frequency discrimination with components of well-learned patterns. *The Journal of the Acoustical Society of America*, *69*(1), 223-230.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137-149.

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

Stevens, A. A. (2004). Dissociating the cortical basis of memory for voices, words and tones. *Cognitive Brain Research*, *18*(2), 162-171.

Sugiura M. (2014). Neuroimaging studies on recognition of personally familiar people. Frontiers in Bioscience (19), 672–686.

Szameitat, D. P., Alter, K., Szameitat, A. J., Darwin, C. J., Wildgruber, D., Dietrich, S., & Sterr, A. (2009a). Differentiation of emotions in laughter at the behavioral level. *Emotion*, *9*(3), 397–405.

Szameitat, D. P., Alter, K., Szameitat, A. J., Wildgruber, D., Sterr, A., & Darwin, C. J. (2009b). Acoustic profiles of distinct emotional expressions in laughter. *The Journal of the Acoustical Society of America*, *126*(1), 354-366.

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America*, *85*(4), 1699-1707.

Titze, I. R. (1994). Mechanical stress in phonation. *Journal of Voice*, *8*(2), 99-105.

Van Lancker, D. R., & Canter, G. J. (1982). Impairment of voice and face recognition in patients with hemispheric damage. *Brain and Cognition*, *1*(2), 185-195.

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. R., Kreiman, J., & Cummings, J. (1989). Voice perception deficits: Neuroanatomical correlates of phonagnosia. *Journal of Clinical and Experimental Neuropsychology*, *11*(5), 665-674.

Van Lancker, D., & Cummings, J. L. (1999). Expletives: Neurolinguistic and neurobehavioral perspectives on swearing. *Brain Research Reviews*, *31*(1), 83-104.

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.

Von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research*, *17*(1), 48-55.

Von Kriegstein, K., & Giraud, A. L. (2006). Implicit multisensory associations influence voice recognition. *PLoS Biology*, *4*(10), e326.

Von Kriegstein, K., Kleinschmidt, A., & Giraud, A. L. (2006). Voice recognition and cross-modal responses to familiar speakers' voices in prosopagnosia. *Cerebral Cortex*, *16*(9), 1314-1322.

Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention. *Trends in Cognitive Sciences*, *9*(12), 585-594.

Wagner, H. L. (1993). On measuring performance in category judgment studies of nonverbal behavior. *Journal of Nonverbal Behavior*, *17*(1), 3-28.

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual differences in holistic processing predict face recognition ability. *Psychological Science*, *23*(2), 169-177.

Wattendorf, E., Westermann, B., Fiedler, K., Kaza, E., Lotze, M., & Celio, M. R. (2013). Exploration of the neural correlates of ticklish laughter by functional magnetic resonance imaging. *Cerebral Cortex*, *23*(6), 1280-1289.

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

Wild, B., Rodden, F. A., Grodd, W., & Ruch, W. (2003). Neural correlates of laughter and humour. *Brain*, *126*(10), 2121-2138.

Wilson, M. (1988). MRC Psycholinguistic Database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, *20*(1), 6-10.

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. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law*, *1*(4), 792-816.

Yarmey, A. D., & Matthys, E. (1992). Voice identification of an abductor. *Applied Cognitive Psychology*, *6*(5), 367-377.

Yarmey, A. D., Yarmey, A. L., & Yarmey, M. J. (1994). Face and voice identifications in showups and lineups. *Applied Cognitive Psychology*, *8*(5), 453-464.

Yarmey, A. D., Yarmey, A. L., Yarmey, M. J., & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology*, *15*(3), 283-299.