How Talker Identity Relates to Language Processing

Sarah C. Creel* and Micah R. Bregman
University of California, San Diego

Abstract

Speech carries both linguistic content – phonemes, words, sentences – and talker information, sometimes called 'indexical information'. While talker variability materially affects language processing, it has historically been regarded as a curiosity rather than a central influence, possibly because talker variability does not fit with a conception of speech sounds as abstract categories. Despite this relegation to the periphery, a long history of research suggests that phoneme perception and talker perception are interrelated. The current review argues that speech perception itself may arise from phylogenetically earlier vocal recognition, and discusses evidence that many cues to talker identity are also cues to speech-sound identity. Rather than brushing talker differences aside, explicit examination of the role of talker variability and talker identity in language processing can illuminate our understanding of the origins of spoken language, and the nature of language representations themselves.

Spoken language contains a great amount of communicative information. Speech can refer to things, but it also identifies or classifies the person speaking. This dual function links language to other types of vocal communication systems, and means that recognizing speech and recognizing speakers are intertwined. Imagine someone says the word *cat*. The vocal pitch of 'cat' is 200 Hz, and the delay between opening of the vocal tract and the beginning of vocal-cord vibration is 80 ms. The listener has to process all of this acoustic information to identify this word as /kæt/. Many would agree that the vocal pitch is relatively unimportant, while the 80-ms voice onset time is crucial because it distinguishes the phoneme /k/ from the very similar phoneme /g/. But what happens to the 'unimportant' information? On a modular view, information not linked to phoneme identity is discarded by the speech-processing system (e.g. Liberman and Mattingly 1985). Nonetheless, details such as vocal pitch may influence comprehension because they indicate the speaker's identity, and can refine the listener's expectations about what particular speakers are likely to talk about.

Differing theories of language processing have different accounts of talker identification. On one view, two separate, independent systems process speech and talker identification (Belin et al. 2004; González and McLennan 2007). Implicit in the two-systems claim is the assumption that acoustic cues to word identity and talker identity are completely independent of each other. On a different view of language processing, both talker identification and language processing arose from evolutionarily earlier capacities to recognize individuals from vocal cues (e.g. Owren and Cardillo 2006), and are thus intertwined. The one-system claim suggests that talker identification and language processing utilize the same set of memory representations, and that humans *learn* (during language acquisition) what elements of sound suggest a difference in meaning, a difference in talker, or both.

In fact, it has been known for decades that speech-sound characteristics and talker characteristics are *not* independent (e.g. Bricker and Pruzansky 1966; Ladefoged and

Broadbent 1957). Returning to the *cat* example, it turns out that knowing the speaker's vocal pitch in *cat* is important for identifying the vowel as /æ/. Listeners interpret formants (louder frequency regions present in vowels, that differ from vowel to vowel) differently depending on the voice's pitch (the *fundamental frequency*, or f0) (see e.g. Hillenbrand et al. 1995; Miller 1953). Further, listeners can use formants alone to identify talker gender (Fellowes et al. 1997), suggesting that phoneme representations and talker-specific details are closely linked.

The following text outlines what language scientists should know about the role of talker variability in language processing. This information can help us understand the evolutionary roots of human language – how similar is it to other animals' vocal communications? It also speaks to the modularity of speech processing – if speech processing is truly modular, talker information should not impinge. However, as we will see, talker information profoundly affects speech processing.

In the first two sections below, we outline vocal recognition in animals and human listeners. The third section considers how talker-specific sound properties impact speech recognition. The fourth section briefly outlines how talker-varying acoustic properties can constrain sentence comprehension. Finally, we summarize and discuss outstanding questions in this area.

1. Individual Recognition in Other Species

To understand where language may come from, it is important to understand vocal communication systems in other species. Discussions of language evolution often focus on the diversity of meaning that can be communicated through syntactic recombination of words. Non-human animals are far more limited in transmitting meaning by recombining sound units (see e.g. Zuberbühler 2002). How this ability arose in humans is an important scientific question. Sometimes overlooked, however, is a more straightforward parallel between speech and animal vocalizations: communicating the vocalizer's identity and physical characteristics. We focus primarily on birdsong, an animal model frequently used as an analog of human speech. Vocal learning in humans and birds is a clear case of convergent evolution, since our common ancestor approximately 300 million years ago is shared with many non-vocal-learning animals (including reptiles). We also discuss acoustic cues to identity in non-human primates, whose communication systems may be homologous to human speech.

Birdsong is a natural acoustic communication signal used by about 4,500 species of oscine birds (songbirds) to attract mates, claim territory and recognize individual species members (reviewed by Kroodsma and Miller 1996). The songbird auditory system is highly specialized for the perception of song, and physiological study of auditory brain regions has revealed neurons tuned to represent song (reviewed by Bolhuis and Gahr 2006). Birdsong is an important animal model of vocal learning, and is the best-studied non-human analog to speech, primarily because its development parallels the development of human speech perception and production (Doupe and Kuhl 1999). Unlike our nearest primate relatives, who do not learn their vocalizations (though a few mammals do, e.g. whales, dolphins and elephants; Janik and Slater 1997), songbirds are prolific vocal learners; many species learn new songs throughout life. If isolated or deafened during development, songbirds do not produce normal song (Konishi 1963; Marler and Tamura 1964; reviewed by Brenowitz et al. 1997), and continued auditory feedback is required for normal song to be maintained (Nordeen and Nordeen 1992).
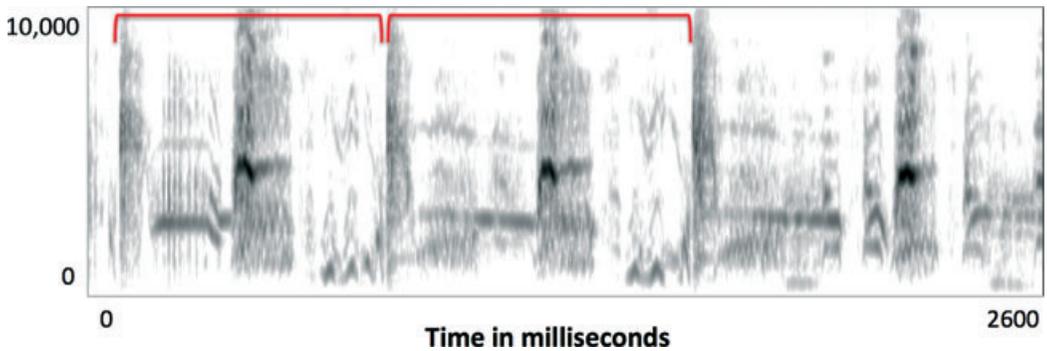
Fig 1. Starling song (courtesy of T. Gentner). Starlings may recognize each other by the motifs they produce in their songs. Red brackets mark two instances of the same motif.

Both birdsong and speech are used by species members to acquire information about the individual producing the vocalization. This information may include cues to an individual's sex, health and reproductive state, and can often be used to precisely identify another individual (Falls 1982). In speech, both characteristics of the vocal apparatus and utterance content (one's choice of words) can conceivably be used to recognize talkers. Similarly, across bird species, vocal–tract characteristics (Weary and Krebs 1992) and song content (Beecher et al. 1994; Gentner and Hulse 2000; see Figure 1) have been implicated in individual recognition in upward of 130 different bird species (Stoddard 1996). Different species use different acoustic features for recognition. For example, song sparrows can use song content both to discriminate between the songs of neighbors and strangers, and to recognize the songs of individual neighbors (Beecher et al. 1991); they are thought to memorize the song types of individual singers to perform this discrimination. European starlings are lifelong vocal learners who sing complex songs that often last longer than 1 minute. These songs comprise repeated units called motifs (Figure 1), and laboratory studies suggest that starlings learn to associate particular motifs with individual singers (Gentner and Hulse 2000).

Individual birds' songs are often highly distinct, and this variability is important for successful individual recognition. Because of memory limitations, birds are likely to concentrate their resources on learning to identify critical individuals (mates, parents). Several avian species are monogamous (at least through a single breeding season), making mate recognition an important function of song. Studies in several species have shown that birds can recognize their mate's song even in the context of the song of many other individuals (Lind et al. 1996; Miller 1979b). Similarly, female zebra finches, which may not recognize large numbers of individuals, can recognize their fathers even after two months of separation (Miller 1979a).

Primate vocalizations are also of interest, because primates are our closest living relatives. Mechanisms of vocal production, described more thoroughly in Section 2, are similar among primate species (including humans). Many primate species produce calls with individually distinctive acoustic features, including rhesus macaques, Barbary macaques, squirrel monkeys, vervet monkeys and others (Cheney and Seyfarth 1980; Hammerschmidt and Todt 1995; Rendall et al. 1996; Rendall et al. 1998). Rendall et al. (1998) mathematically analyzed the acoustics of macaque calls, and determined that at least one call type, coos (Figure 2), provided reliable acoustic cues for discriminating among female Rhesus macaques (*Macaca mulatta*) with 79.4% accuracy.
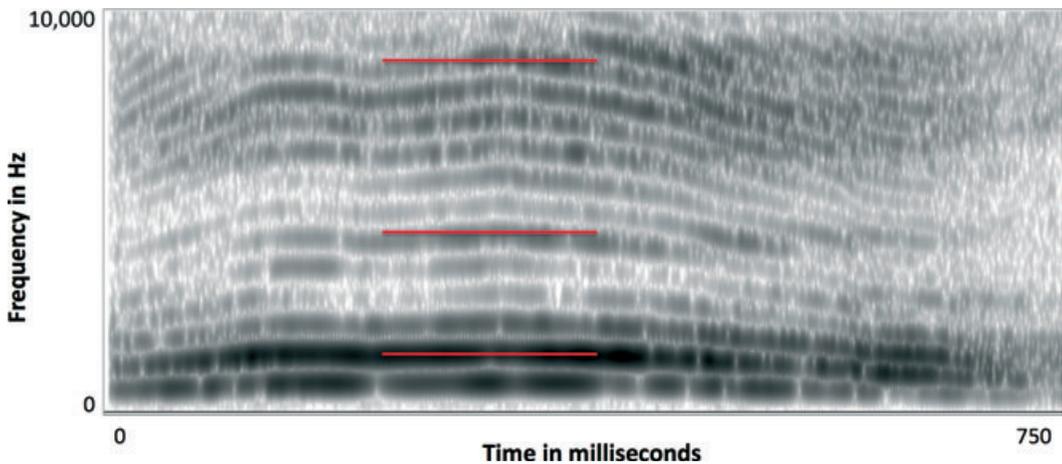
Fig 2. Macaque coo (courtesy of A. Ghazanfar). Small horizontal bands with a slight inverted-U contour are individual harmonics (integer frequency multiples) of the fundamental frequency; thicker, dark horizontal bands, highlighted with red lines, are formants. Macaques may use formant dispersion (the distance between formants) to identify each other.

In 'playback' experiments, researchers assess whether animals actually use these acoustic identity cues by presenting recordings of other animals' vocalizations from audio speakers in the animal's habitat, and measuring their behavioral response. Reliable call discrimination is evident if animals respond differently to different animals' vocalizations. Such experiments suggest that several primate species distinguish individuals by their calls, including forest monkeys (e.g. Waser 1977), rhesus macaques (Rendall et al. 1996), and baboons (Bergman et al. 2003). Baboons recognize kin relationships of other baboons (Cheney and Seyfarth 1999). Primates may use vocal cues to assess other attributes as well (age, sex, body size; reviewed in Ey et al. 2007). Among several primate species, common predictive cues include: mean (average) f0, f0 range, f0 variability, dispersion of power throughout the spectrum, mean gap length (gap = section of a call with low amplitude), frequency in the spectrum with highest amplitude, and call duration (Ey et al.). F0 has often been suggested as a useful predictor of body size; however in primates (including humans) the correlation seems weak once age is accounted for (e.g. Lass and Brown 1978; Perry et al. 2001; though see Krauss et al. 2002). Formant frequency dispersion (pitch distance between vocal-tract resonances; Fitch 1997; Ghazanfar et al. 2007) provides better information, at least in macaques. As we will see below, formants may also be important for identifying human voices (e.g. Fellowes et al. 1997).

Vocal communication systems are enormously diverse among species. Despite this diversity, vocalizations allow listeners to acquire information about the producer. This is often overlooked in human language, where we focus on how speech communicates arbitrary meaning, rather than information about individuals. Human vocalizations may be homologous (evolutionarily related) to primate vocalizations in terms of the sounds produced and recognized (formants, not motifs), but human vocalizations are *analogous* to bird song in that they are flexibly learned. We now examine vocal cues used by human listeners to identify talkers. As with primates, acoustic cues underlying individual recognition of humans are not fully understood, although there has been substantial progress.
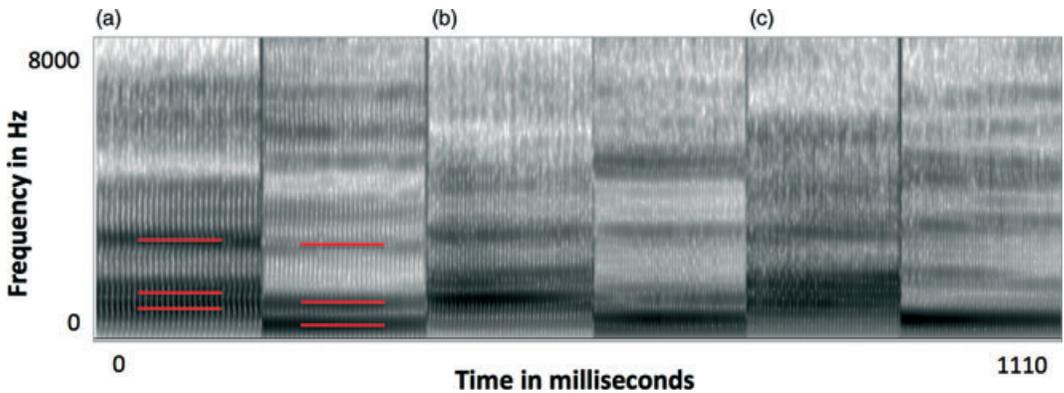
Fig 3. Speakers from Hillenbrand et al. (1995) producing the vowels /a/ and /u/. (a) Male speaker; (b) female speaker; (c) female child. Dark horizontal bars, highlighted with red lines for the first two vowels, are formants. Note the differences in formants between vowels, and between the three speakers. Measurements and sound files available at Hillenbrand's web site: <http://homepages.wmich.edu/~hillenbr/voweldata.html>.

2. What Acoustic Attributes Vary Between Talkers?

As the reader likely observes regularly, many acoustic characteristics vary amongst talkers. These can be divided into 'source' characteristics – properties arising from the vibration of the glottis (the vocal cords, or, more accurately, vocal *folds*) – and 'filter' characteristics – sounds created by dynamically changing the shape and size of the vocal tract, particularly the mouth. Source and filter are largely independent. The source can change while the filter stays put: the vowel 'Aaaah!' denoting satisfaction maintains filter characteristics (the mouth shape which produces /a/) but changes source characteristics (the pitch drops). The filter can also change without the source, such as the 'Lalalalala' sound of someone blocking out a speech act they do not want to hear: the source (pitch) stays the same, while the filter changes rapidly (the tongue tip moving alternately up and down). Filter characteristics for vowels are described in terms of *formants* – frequency regions of the source sound that are amplified for a given vowel (Figure 3; also evident in macaque coos, Figure 2). Formants are numbered in increasing pitch (F1 is lowest, then F2, etc.). F0 is lower than all formants. Ladefoged's classic *A Course in Phonetics* (2006) is an excellent introduction to these topics.

Numerous source properties differ across talkers, including f0, f0 variability, and different types of vocal-fold vibration such as creaky voice and breathy voice. All of these characteristics indicate differences in meaning in some languages, and in certain prosodic or phonetic contexts in other languages (see Gordon and Ladefoged 2001 for more details). While the filter produces most speech-sound contrasts, the source can also produce contrasts. This is counter to a common view (which, as we shall see below, is incorrect) that talker identity resides in the source, and word identity in the filter. Other source characteristics include jitter (micro-variations in f0) and shimmer (micro-variations in loudness), which, in normal speakers, decrease as vocal loudness increases (Brockmann et al. 2008).

Filter properties also differ across talkers. That is, different talkers produce different acoustic realizations of many phonemes (vowels are addressed here, but consonants are also affected; for example, Allen et al. 2003; Newman et al. 2001). Variability exists between accents (Bradlow and Bent 2008; Clopper et al. 2005), genders (Simpson 2009),
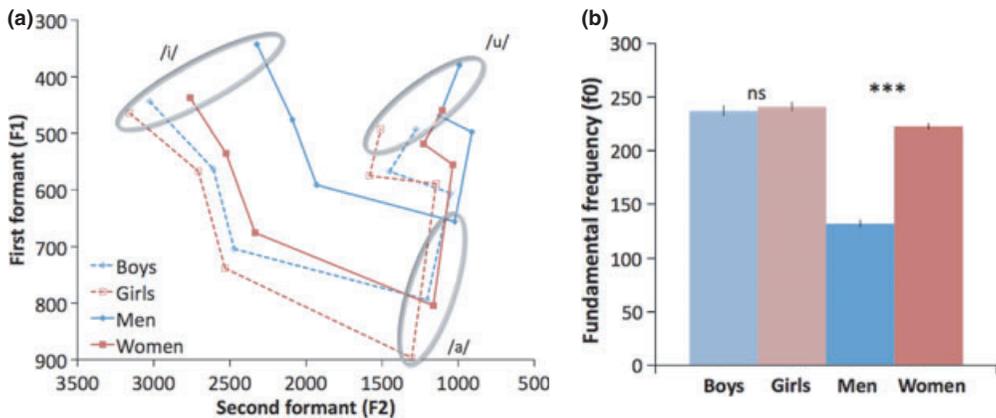
Fig 4. For male and female children and adults in Hillenbrand et al. (1995), F2 vs. F1 (a), and f0 (b). Children do not differ in f0 by gender, but differ in F1 and F2. Children's ages are reported as between 10 and 12, which is likely (but not with certainty) prepubescent. Adults differ in all three. ***$p < 0.0001$.

and individual speakers (Fellowes et al. 1997). Gendered differences may be partly socially conditioned. In children, vocal-tract properties differentiate around puberty (Fitch and Giedd 1999). However, before puberty, female children show higher formant frequencies than males (Figure 4a) even though f0 does not differ yet (Figure 4b; Perry et al. 2001). This suggests that learned 'gender dialects' (Johnson 2006) may generate differences in male vs. female speech patterns. Finally, though not emphasized here, longer-time-scale properties like f0 variability (prosody) and timing characteristics may also distinguish talkers.

Which of these numerous acoustic properties do listeners use to identify voices? One approach to this question presents several voices (usually 20–40) to listeners, who rate voice similarity or try to identify voices (e.g. Baumann and Belin 2010; Bricker and Pruzansky 1966; Goldinger 1996; Kreiman et al. 1992; Murry and Singh 1980; Singh and Murry 1978). If two voices are confused often, or rated as very similar, what attributes do they share? If rarely confused, or rated very different, how do they differ?

Bricker and Pruzansky (1966) explored whether listeners use acoustic attributes independent of speech sounds to identify talkers. This would allow talker recognition and speech recognition to proceed independently, as in two-system models. They presented pairs of vocalizations (e.g. /i/ from Talker 1 and /i/ from Talker 14) and asked listeners to say whether the same person or different people produced them. Among other findings, voices that were confused on one vowel (/i/) were not necessarily confused given another vowel (/u/). This suggests that vowel information and talker-identifying information are not processed independently.

Nonetheless, researchers have continued to search for talker-relevant acoustic properties. The most reliable cue, emerging across all studies, is f0. Others include formant frequencies (Baumann and Belin 2010; Murry and Singh 1980), hoarseness (Murry and Singh; Singh and Murry 1978), vowel duration (Murry and Singh; Singh and Murry), and shimmer (Kreiman et al. 1992). However, other than f0, there is little consistency in the cues that are uncovered.

A second approach to studying talker identification assesses recognition accuracy when certain cues are removed or distorted. If recognition is impaired, then that cue is probably important to recognition. Van Lancker et al. (1985a; and Van Lancker et al. 1985b)

studied famous-voice recognition under different types of distortion (forward or backward playback, altered duration). Each type of distortion impaired recognition of some voices but not others. Remez et al. (1997; Fellowes et al. 1997) removed f0 by presenting sine-wave analogs of the talkers' formants, and found that gender and individual identity could be identified without f0. Fellowes et al. suggested that formant frequencies alone were important to identifying gender, while timing characteristics were also important for identifying individuals.

This is admittedly an inconsistent picture of talker identification cues. Van Lancker et al. (1985a) suggest that 'the critical parameter(s) [for recognition] are not the same for all voices … any of these cues may serve as a 'primary' cue at one time but not another' (33). That is, many cues potentially distinguish talkers, but certain cues only distinguish a small fraction of talkers. Less common cues are unlikely to show up when only a few voices are studied. Studies may also reflect inconsistency among individual listeners (Kreiman et al. 1992). The best current account is that listeners can exploit myriad acoustic cues to talker identity, and at least some cues also function in speech-sound identification. The next section discusses how overlap between talker and speech information affects speech processing.

3. How does Talker Variability Affect Speech Processing?

Given that talkers vary in attributes that distinguish speech sounds, it is not surprising that some acoustic correlates of talker identity affect speech perception. Talker information can both positively and negatively affect speech-sound identification. The positive or negative outcome depends on the type of experimental task. When the talker changes rapidly during an experiment, speech recognition becomes more difficult. However, when talker information is consistent – for instance, when a talker who said a word once says it again – recognition may be facilitated. These will be referred to as *talker interference* effects and *talker specificity* effects respectively.

3.1. TALKER INTERFERENCE EFFECTS

Rapid talker variation interferes with speech-sound identification. Mullennix and Pisoni (1990), Nusbaum and Morin (1992), and Green et al. (1997) demonstrated this using interference paradigms. In each case, listeners classified a speech sound (as, say, /b/ or /p/) in a set of words where the talker varied unpredictably, or where the talker varied predictably (it stayed the same or was correlated with the speech-sound dimension). If listeners could pay attention just to the phonemes, then unpredictable talker variation should not distract them. However, listeners were slower and less accurate in identifying speech sounds when talkers varied unpredictably, suggesting that they could not attend just to phonemic information. Memory for word-lists (*dog*, *shoe*, *cup*, *bike*…) also shows talker interference. Listeners recall fewer words from multiple-talker lists than from single-talker lists (Martin et al. 1989). Those authors suggest that cognitive processing space for encoding words was taken up by adjusting to a different talker on each word. Interestingly, this effect reverses when time between words is increased (Goldinger et al. 1991), suggesting that more time allows listeners to use talker differences to more richly encode words in memory.

The reason for listeners' difficulty in perception when talker changes rapidly seems to be the perceptual or attentional adjustment required, often called *talker normalization*. In an early demonstration, Ladefoged and Broadbent (1957) altered a lead-in phrase, 'Please

say what this word is', shifting formants either higher or lower than the original. They presented identical word recordings (such as 'bit') after different versions of the lead-in. Listeners' perceptions of the vowel in 'bit' changed depending on the lead-in's formants. This suggests that the lead-in calibrated listeners to that speaker's vowel space, consistent with Joos (1948) early account of talker normalization.

Effects of normalization are widely acknowledged, but the underlying causes are still debated. One idea is that it represents not just an adjustment to incoming information, but activation of *expectations*. Johnson (2005; Johnson et al. 1999) describes studies where the talker's apparent gender influences the perception of speech sounds (see Niedzielski 1999; Staum Casasanto 2008; for dialect effects). This suggests that listeners are normalizing to a memory representation of one group of talkers versus another. Magnuson and Nusbaum (2007) even found that listeners who were told they would hear two talkers in a vowel-identification task performed more slowly than those told they would hear one. Thus, normalization may involve the expectation that one must reorient auditory attention.

3.2. TALKER–SPECIFICITY EFFECTS

Distinct from talker interference effects, which presumably result from normalization, are *talker-specificity effects*: processing is better when talker information remains constant from one presentation to the next. For instance, listeners understand familiar talkers better than unfamiliar talkers in noise (Nygaard et al. 1994). This suggests that familiarity with talker-specific acoustics facilitates word recognition. Palmeri et al. (1993) presented listeners with a series of words and non-words, each spoken by a particular talker. In a later phase, they presented a word list consisting of words from the first list (50%), and new words (50%). Listeners now made old word/new word judgments. Crucially, old words were spoken either by the original talker, or a new talker. Listeners were more accurate at responding 'old' when the original talker spoke the word again. (See Schacter and Church 1992; Church and Schacter 1994 for related work.)

At what level is talker-specific information represented? Creel et al. (2008), in a visual-world eye-tracking experiment, found evidence for storage of talker-specific instances of whole words. Other authors have found effects at the sublexical level: Eisner and McQueen (2005; see also Kraljic and Samuel 2005) find talker-specific adaptation to boundary shifts between fricative categories (/f/ vs. /s/) which generalize to words not heard with the shifted phoneme. This suggests that *phoneme* representations, not (just) word representations, are affected by talker-specific information (though see Kraljic and Samuel 2006).

3.3. LEARNING AND DEVELOPMENTAL EFFECTS

A final important aspect of talker variability is how it affects language learning. Infants seem to form word representations that are too acoustically specific (e.g. Houston and Jusczyk 2000). Infant recognition is assessed by how long they look at a sound-producing source: if there is a difference in the time spent looking to a familiar sound vs. an unfamiliar sound, we can conclude that infants recognize the familiar sound. By this measure, 7.5-month-olds cannot distinguish a familiarized word from an unfamiliarized one when it is spoken by a new, dissimilar talker (Houston and Jusczyk 2000), though by 10.5 months they can (for parallels in vocal emotion: Singh 2008; Singh et al. 2008; in accent: Schmale and Seidl 2009; Schmale et al. 2010). Conversely, infants are more accurate at identifying words when they are 'taught' (exposed to) words in a variety of voices (Rost and McMurray 2009, 2010) or vocal emotions (Singh 2008). Similar benefits of

multi-talker learning are evident in L2 speech-sound acquisition (Lively et al. 1993; Logan et al. 1991), perceptual attunement to non-natively accented speech (Bradlow and Bent 2008), and learning of unfamiliar dialects (Clopper and Pisoni 2004).

These studies suggest that new listeners have difficulty generalizing across talker variability, which improves with exposure to a wider range of talkers. One explanation for this effect is that listeners must figure out the right dimensions for identification (F1 indicates meaning differences in my language, while creaky voice does not): with more exposure, learners increasingly tune out cues irrelevant to word identity. This predicts that adults would be less adept than children at identifying voices, because adults have tuned out much talker variability. However, evidence suggests that adults are *better* than children at discriminating or identifying talkers (Bartholomeus 1973; Mann et al. 1979), inconsistent with adults globally tuning out talker information.

An alternative explanation is that learners improve at tuning *in* to different dimensions of speech variability, contingent on the listening situation (identifying words, vs. identifying talkers). Goldinger (1996, 1998) suggests that word and talker information reside in the same exemplar-style representations (Hintzman 1986; Nosofsky 1989) – mental recordings of every speech instance the listener has ever experienced. Storing them all would allow clusters of information – speech sounds, talkers – to emerge naturally. Different clusterings could be focused on by directing attention to particular acoustic attributes (see Nosofsky 1989; Werker and Curtin 2005). For instance, attention to creaky voice might be higher when trying to recognize a talker vs. a word.

4. Knowledge-based Influences of Talker Information on Speech Processing

Finally, talker information is useful at the sentence or discourse level. Speech provides rich information about the talker's attributes (Ladefoged and Broadbent 1957; femininity, Ko et al. 2009; sexual orientation, Pierrehumbert et al. 2004; physical size, Krauss et al. 2002). This information can be used to predict what the talker may say. Some researchers have argued that activating social knowledge along with speech subtly alters the speech's meaning (Geiselman and Bellezza 1976, 1977; Geiselman and Crawley 1983). Geiselman and Bellezza (1977), taking talker gender as a test case, found that listeners rated printed sentences as more 'potent' when men had spoken them previously than when women had. One hopes this effect would be moderated somewhat 30 years later. Dimming these hopes, Ko et al. (2009) found that judgments of a talker's perceived competence (at carrying out a hypothetical job the talker had applied for) were strongly related to vocally based judgments of femininity.

Adults rapidly invoke talker-related knowledge in processing sentences. Using electroencephalography, Van Berkum et al. (2008) found that listeners' predictions of upcoming words in spoken sentences differed by speaker. Previous work (see Kutas and Hillyard 1980, 1984) shows that when listeners hear a sentence like 'I like coffee with ___', and then the sentence ends with an unexpected word (e.g. 'dog'), the brain generates a particular electrophysiological response – an N400 – which does not appear for listeners who hear an expected completion ('sugar'). Van Berkum et al. looked for the N400 in a sentence such as 'I like to drink wine'. Listeners who heard an adult speaking showed a smaller N400 to 'wine' than listeners who heard a child, suggesting that listeners interpret spoken material contingent on vocally perceived social attributes of the talker.

There is also a developmental aspect to learning social correlates of talker variability. Neonates (DeCasper and Fifer 1980) and even fetuses (Kisilevsky et al. 2003) respond differently to their mother's voice than to a stranger's voice, suggesting prenatal encoding

of talker characteristics. Kinzler et al. (2007) showed that infants preferred to look at faces associated with their native language, and 5-year-olds preferred to befriend native-language or native-accent speakers over non-native ones (see also Hirschfeld and Gelman 1997). These studies suggest that children use acoustic familiarity to socially evaluate talkers. Going beyond effects of familiarity, Creel (2010) found evidence for specific, high-level mappings of voice attributes to individuals: preschool-aged children who knew two talkers' favorite colors used talker information early in an instruction sentence ('Can you help me find the square?') to visually fixate shapes of the talker's preferred color, prior to hearing the shape name itself. Crucially, they only did so when the talker asked for herself, not when she asked for the other talker. This suggests that, at least in simple cases, children use talker information to constrain sentence interpretation like adults. It is less clear whether children would be able to learn or utilize more subtle or complex information.

5. General Discussion

In this review, we have explored talker identification and how it relates to language processing. We discussed its similarities to within-species individual recognition in songbirds and primates, and outlined attempts to characterize acoustic cues to human talker identity. We then described two levels at which talker-related acoustic cues affect language processing: at the phonological level, and at the sentential level.

We close by outlining several interesting questions regarding talker variability and speech processing. First, how closely is language related to individual recognition in primates? Among primates, only we humans learn our vocalizations. Did language emerge from non-human primate vocalizations? We cannot answer this definitively. However, one useful observation concerns bird vocalizations, which, though acoustically very different from primate vocalizations, may provide an evolutionary parallel: not all vocalizing bird species learn their vocalizations (e.g. chickens; Konishi 1963). This implies that recognizing non-learned vocalizations is a necessary evolutionary precursor to more flexible vocal learning (Dubbeldam 1998). That is, recognition of non-learned vocalizations in primates may have formed an evolutionary platform for human vocal learning – humans are the 'songbirds' of primates.

Another question is what, other than phoneme-varying cues, distinguishes talkers. Recall that some sound properties which vary idiosyncratically by talker in one language vary phonemically in others (Gordon and Ladefoged 2001). This suggests that many, if not all cues, can function dually to identify phonemes and talkers. Distinguishing talkers may depend on attunement to cues that differentiate the talkers in one's own range of experience (see Perrachione et al. 2010, for supporting evidence). Overall, the degree of overlap between talker-identifying cues and speech-relevant cues, and listeners' apparent inability to process speech information and talker information separately, suggests that language is not a module which is impenetrable to acoustic variability.

A final question, also related to modularity, is how talker and speech information are stored – as the output of two separate systems, part of a single set of representations (e.g. collections of all experienced speech), or both? Three pieces of evidence are relevant. First, as discussed above, many talker-varying acoustic properties also identify speech sounds, suggesting that the same representations may be needed for both talker and speech-sound identification. Second, studies on talker-specificity effects show that when listeners learn to recognize a particular talker's unusual phoneme (e.g. /s/ is ambiguous between /f/ and /s/), they adjust recognition of other words that contain this phoneme

(e.g. Eisner and McQueen 2005; Kraljic and Samuel 2005). That is, talker-specific properties affect recognition of speech sounds. If talker information were stored separately from speech-sound information, this should not happen – a distorted phoneme would be useful for identifying the talker *as that individual*, but would not aid recognition of speech from that talker. Further, listeners would adapt across-the-board to the shifted phoneme, not to a specific talker's productions of that phoneme (Eisner and McQueen; Kraljic and Samuel 2005; though see Kraljic and Samuel 2006).

Finally, as noted in the Introduction, some neuroimaging evidence suggests two separate systems, with speech sounds analyzed in the left brain hemisphere and talker identity in the right (Belin et al. 2004; González and McLennan 2007). Other studies suggest left hemisphere contributions to talker recognition (Perrachione et al. 2009) and environmental sound recognition (Saygin et al. 2003). Why the discrepancies? Results across studies may differ in the temporal scale of the acoustic information being used to identify speech sounds vs. talkers (for instance, voice onset time vs. prosody). This is because the left hemisphere seems biased toward rapid temporal events while the right processes slower events (Zatorre and Belin 2001), meaning that right–hemisphere activation may reflect use of slower temporal scales rather than specialization for talker identification. That is, listeners distinguish speech sounds by short-time-scale properties, but tend to distinguish voices – especially unfamiliar ones – by slower-temporal-scale properties. The information to be identified (talker vs. speech) and expertise or familiarity with the voices (see Perrachione et al. 2009) may affect the temporal scale used, with distinctions among more familiar voices being made from more temporally fine-grained information.

In sum, much research indicates that talker-specific acoustic patterns are related to and impact language processing. However, there is much to learn about how listeners represent talker-identifying information, and what the real-world ramifications are for language processing. We hope the reader is intrigued and tantalized by the prospect of new developments in this area.

Short Biographies

Sarah C. Creel is an Assistant Professor of Cognitive Science at the University of California, San Diego. Her research centers around the specificity of auditory memory and how auditory representations are accessed in real time processing. She has published research on language processing, vocabulary acquisition, and music cognition, in journals including *Cognition*, *Journal of Memory and Language*, and the *Journal of Experimental Psychology: Learning, Memory and Cognition*. Before arriving at UCSD, she completed a dual degree in music (BA) and psychology (BS) at the University of South Carolina-Columbia in 1999, and a PhD in Brain and Cognitive Sciences at the University of Rochester in 2005, where she was a Sproull Fellow and an NSF Graduate Research Fellow. Two more years were spent as a postdoctoral fellow at the University of Pennsylvania. Her current research with preschoolers and adults focuses on the use of talker information in comprehension, influences of talker and accent variability on word learning, and non-speech auditory analogs to language learning.

Micah R. Bregman is a PhD student in the Department of Cognitive Science at the University of California, San Diego where he is a member of the laboratories of Dr Timothy Gentner and Dr Sarah Creel. His research examines the perception and memory of complex acoustic signals in birds and humans. Prior to attending UCSD, he attended Pomona College, where he earned a Bachelors degree in Mathematics and

Music. After graduation, Micah received a Fulbright Fellowship to study music perception and cognition at the University of Jyväskylä in Finland. His current research interests include how birds and humans use pitch in the perception of conspecific communication.

Note

* Correspondence address: Sarah C. Creel, Department of Cognitive Science, University of California, 9500 Gilman Drive Mail Code 0515, La Jolla, CA 92093-0515, USA. E-mail: creel@cogsci.ucsd.edu

Works Cited

Allen, J. Sean, J. L. Miller, and D. DeSteno. 2003. Individual talker differences in voice-onset-time. Journal of the Acoustical Society of America 113. 544–52.

Bartholomeus, Bonnie. 1973. Voice identification by nursery school children. Canadian Journal of Psychology 27. 464–72.

Baumann, Oliver, and P. Belin. 2010. Perceptual scaling of voice identity: common dimensions for different vowels and speakers. Psychological Research 74. 110–20.

Beecher, Michael D., P. K. Stoddard, C. L. Horning, and S. E. Campbell. 1991. Recognition of individual neighbors by song in the song sparrow, a species with song repertoires. Behavioral Ecology and Sociobiology 29. 211–5.

——, S. E. Campbell, and J. Burt. 1994. Song perception in the song sparrow: birds classify by song type but not by singer. Animal Behaviour 47. 1343–51.

Belin, Pascal, S. Fecteau, and C. Bédard. 2004. Thinking the voice: neural correlates of voice perception. Trends in Cognitive Sciences 8. 129–35.

Bergman, Thore J., J. C. Beehner, D. L. Cheney, and R. M. Seyfarth. 2003. Hierarchical classification by rank and kinship in baboons. Science 302. 1234–6.

Bolhuis, Johan J., and M. Gahr. 2006. Neural mechanisms of birdsong memory. Nature Reviews Neuroscience 7. 347–57.

Bradlow, Anne R., and T. Bent. 2008. Perceptual adaptation to non-native speech. Cognition 106. 707–29.

Brenowitz, Eliot A., D. Margoliash, and K. W. Nordeen. 1997. An introduction to birdsong and the avian song system. Journal of Neurobiology 35. 495–500.

Bricker, Peter D., and S. Pruzansky. 1966. Effects of stimulus content and duration on talker identification. Journal of the Acoustical Society of America 40. 1441–9.

Brockmann, Mieke, C. Storck, P. N. Carding, and M. J. Drinnan. 2008. Voice loudness and gender effects on jitter and shimmer in healthy adults. Journal of Speech, Language, And Hearing Research 51. 1152–60.

Cheney, Dorothy, and R. Seyfarth. 1980. Vocal recognition in free-ranging vervet monkeys. Animal Behaviour 28. 362–7.

——, and ——. 1999. Recognition of other individuals' social relationships by female baboons. Animal Behaviour 58. 67–75.

Church, Barbara A., and D. L. Schacter. 1994. Perceptual specificity of auditory priming: implicit memory for voice intonation and fundamental frequency. Journal of Experimental Psychology. Learning, Memory, And Cognition 20. 521–33.

Clopper, Cynthia G., and D. B. Pisoni. 2004. Effects of talker variability on perceptual learning of dialects. Language and Speech 47. 207–39.

——, ——, and K. de Jong. 2005. Acoustic characteristics of the vowel systems of six regional varieties of American English. Journal of the Acoustical Society of America 118. 1661–76.

Creel, Sarah C. 2010. Considering the source: preschoolers and adults use talker acoustics predictively and flexibly in on-line sentence processing. Proceedings of the 32nd Annual Conference of the Cognitive Science Society, ed. by S. Ohlsson and R. Catrambone, 1810–15. Austin, TX: Cognitive Science Society.

——, R. N. Aslin, and M. K. Tanenhaus. 2008. Heeding the voice of experience: the role of talker variation in lexical access. Cognition 106. 633–64.

DeCasper, Anthony J., and W. P. Fifer. 1980. Of human bonding: newborns prefer their mothers' voices. Science 208. 1174–6.

Doupe, Allison J., and P. K. Kuhl. 1999. Birdsong and human speech: common themes and mechanisms. Annual Review of Neuroscience 22. 567–631.

Dubbeldam, J. L. 1998. The neural substrate for 'learned' and 'nonlearned' activities in birds: a discussion of the organization of bulbar reticular premotor systems with side-lights on the mammalian situation. Acta Anatomica 163. 157–72.

Eisner, Frank, and J. M. McQueen. 2005. The specificity of perceptual learning in speech processing. Perception & Psychophysics 67. 224–38.

Ey, E., D. Pfefferle, and J. Fischer. 2007. Do age- and sex-related variations reliably reflect body size in non-human primate vocalizations? A review. Primates: Journal of Primatology 48. 253–67.

Falls, J. B. 1982. Individual recognition by sound in birds. Acoustic Communication in Birds, Vol. 2, ed. by D. E. Kroodsma and E. H. Miller, 237–73. New York: Academic Press.

Fellowes, Jennifer M., R. E. Remez, and P. E. Rubin. 1997. Perceiving the sex and identity of a talker without natural vocal timbre. Perception & Psychophysics 59. 839–49.

Fitch, W. Tecumseh. 1997. Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. Journal of the Acoustical Society of America 102. 1213–22.

——, and J. Giedd. 1999. Morphology and development of the human vocal tract: a study using magnetic resonance imaging. Journal of the Acoustical Society of America 106. 1511–22.

Geiselman, Ralph E., and F. S. Bellezza. 1976. Long-term memory for speaker' s voice and source location. Memory & Cognition 4. 483–9.

——, and ——. 1977. Incidental retention of speaker's voice. Memory & Cognition 5. 658–65.

——, and J. Crawley. 1983. Incidental processing of speaker characteristics: voice as connotative information. Journal of Verbal Learning and Verbal Behavior 22. 15–23.

Gentner, Timothy Q., and S. H. Hulse. 2000. Perceptual classification based on the component structure of song in European starlings. Journal of the Acoustical Society of America 107. 3369–81.

Ghazanfar, Asif A., H. K. Turesson, J. X. Maier, R. van Dinther, R. D. Patterson, and N. K. Logothetis. 2007. Vocal-tract resonances as indexical cues in rhesus monkeys. Current Biology 17. 425–30.

Goldinger, Stephen D. 1996. Words and voices: episodic traces in spoken word identification and recognition memory. Journal of Experimental Psychology: Learning, Memory, and Cognition 22. 1166–83.

——. 1998. Echoes of echoes? An episodic theory of lexical access. Psychological Review 105. 251–79.

——, D. B. Pisoni, and J. S. Logan. 1991. On the nature of talker variability effects on recall of spoken word lists. Journal of Experimental Psychology: Learning, Memory, and Cognition 17. 152–62.

González, Julio, and C. T. McLennan. 2007. Hemispheric differences in indexical specificity effects in spoken word recognition. Journal of Experimental Psychology: Human Perception and Performance 33. 410–24.

Gordon, Matthew, and P. Ladefoged. 2001. Phonation types: a cross-linguistic overview. Journal of Phonetics 29. 383–406.

Green, Kerry P., G. R. Tomiak, and P. K. Kuhl. 1997. The encoding of rate and talker information during phonetic perception. Perception & Psychophysics 59. 675–92.

Hammerschmidt, K., and D. Todt. 1995. Individual differences in vocalisations of young barbary macaques, Macaca sylvanus: a multi-parametric analysis to identify critical cues in acoustic signalling. Behaviour 132. 381–99.

Hillenbrand, James, L. A. Getty, M. J. Clark, and K. Wheeler. 1995. Acoustic characteristics of American English vowels . Journal of the Acoustical Society of America 97. 3099–111.

Hintzman, Douglas L. 1986. ''Schema abstraction'' in a multiple-trace memory model. Psychological Review 93. 411–28.

Hirschfeld, Lawrence A., and S. A. Gelman. 1997. What young children think about the relationship between anguage variation and social difference. Cognitive Development 12. 213–38.

Houston, Derek M., and P. W. Jusczyk. 2000. The role of talker-specific information in word segmentation by infants. Perception 26. 1570–82.

Janik, V. M., and P. J. Slater. 1997. Vocal learning in mammals. Advances in the study of behavior, Vol. 26, ed. by P. J. B. Slater, J. S. Rosenblatt, C. T. Snowdon and M. Milinski, 59–98. San Diego: Academic Press.

Johnson, Keith. 2005. Speaker normalization in speech perception. The handbook of speech perception, ed. by D. B. Pisoni and R. Remez, 363–89. Oxford: Blackwell Publishers.

——. 2006. Resonance in an exemplar-based lexicon: the emergence of social identity and phonology. Journal of Phonetics 34. 485–99.

——, E. A. Strand, and M. D'Imperio. 1999. Auditory–visual integration of talker gender in vowel perception. Journal of Phonetics 27. 359–84.

Joos, Martin. 1948. Acoustic phonetics. Language 24. 5–131.

Kinzler, Katherine D., E. Dupoux, and E. S. Spelke. 2007. The native language of social cognition. Proceedings of the National Academy of Sciences 104. 12577–80.

Kisilevsky, Barbara S., S. M. Hains, K. Lee, X. Xie, H. Huang, and H. H. Ye et al. 2003. Effects of experience on fetal voice recognition. Psychological Science 14. 220–4.

Ko, Sei Jin, C. M. Judd, and D. A. Stapel. 2009. Stereotyping based on voice in the presence of individuating information: vocal femininity affects perceived competence but not warmth. Personality and Social Psychology Bulletin 35. 198–211.

Konishi, Masakazu. 1963. The role of auditory feedback in the vocal behavior of the domestic fowl. Zeitschrift für Tierpsychologie 20. 349–67.

Kraljic, Tanya, and A. G. Samuel. 2005. Perceptual learning for speech: is there a return to normal? Cognitive Psychology 51. 141–78.

——, and ——. 2006. Generalization in perceptual learning for speech. Psychonomic Bulletin & Review 13. 262–8.

Krauss, Robert M., R. Freyberg, and E. Morsella. 2002. Inferring speakers' physical attributes from their voices. Journal of Experimental Social Psychology 38. 618–25.

Kreiman, Jody, B. R. Gerratt, K. Precoda, and G. S. Berke. 1992. Individual differences in voice quality perception. Journal of Speech and Hearing Research 35. 512–20.

Kroodsma, D. E., and E. H. Miller (eds). 1996. Ecology and evolution of acoustic communication in birds. Ithaca, NY: Cornell University Press.

Kutas, Marta, and S. A. Hillyard. 1980. Reading senseless sentences: brain potentials reflect semantic incongruity. Science 207. 203–5.

——, and ——. 1984. Brain potentials during reading reflect word expectancy and semantic association. Nature 307. 161–3.

Ladefoged, Peter. 2006. A course in phonetics. 5th ed. Boston: Thomson/Wadsworth.

——, and D. E. Broadbent. 1957. Information conveyed by vowels. Journal of the Acoustical Society of America 29. 98–104.

Lass, N., and W. Brown. 1978. Correlational study of speakers' heights, weights, body surface areas, and speaking fundamental frequencies. Journal of the Acoustical Society of America 63. 1218–20.

Liberman, Alvin M., and I. G. Mattingly. 1985. The motor theory of speech perception revised. Cognition 21. 1–36.

Lind, H., T. Dabelsteen, and P. K. McGregor. 1996. Female great tits can identify mates by song. Animal Behaviour 52. 667–71.

Lively, Scott E., J. S. Logan, and D. B. Pisoni. 1993. Training Japanese listeners to identify English /r / and /l /. II: the role of phonetic environment and talker variability in learning new perceptual categories. Journal of the Acoustical Society of America 94. 1242–55.

Logan, J. S., S. E. Lively, and D. B. Pisoni. 1991. Training Japanese listeners to identify English /r / and /l /: a first report. Journal of the Acoustical Society of America 89. 874–86.

Magnuson, James S., and H. C. Nusbaum. 2007. Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. Journal of Experimental Psychology: Human Perception and Performance 33. 391–409.

Mann, Virginia A., R. Diamond, and S. Carey. 1979. Development of voice recognition: parallels with face recognition. Journal of Experimental Child Psychology 27. 153–65.

Marler, Peter, and M. Tamura. 1964. Culturally transmitted patterns of vocal behavior in sparrows. Science 146. 1483–6.

Martin, C. S., J. W. Mullennix, D. B. Pisoni, and W. V. Summers. 1989. Effects of talker variability on recall of spoken word lists. Journal of Experimental Psychology. Learning, Memory, and Cognition 15. 676–84.

Miller, D. B. 1979a. Long-term recognition of father's song by female zebra finches. Nature 280. 389–91.

——. 1979b. The acoustic basis of mate recognition by female Zebra finches, Taeniopygia guttata. Animal Behaviour 27. 376–80.

Miller, R. L. 1953. Auditory tests with synthetic vowels. Journal of the Acoustical Society of America 25. 114–21.

Mullennix, J. W., and D. B. Pisoni. 1990. Stimulus variability and processing dependencies in speech perception. Perception & Psychophysics 47. 379–90.

Murry, Thomas, and S. Singh. 1980. Multidimensional analysis of male and female voices. Journal of the Acoustical Society of America 68. 1294–300.

Newman, Rochelle S., S. A. Clouse, and J. L. Burnham. 2001. The perceptual consequences of within-talker variability in fricative production. Journal of the Acoustical Society of America 109. 1181–96.

Niedzielski, Nancy. 1999. The effect of social information on the perception of sociolinguistic variables. Journal of Language and Social Psychology 18. 62–85.

Nordeen, K. W., and E. J. Nordeen. 1992. Auditory feedback is necessary for the maintenance of stereotyped song in adult zebra finches. Behavioral and Neural Biology 57. 58–66.

Nosofsky, Robert M. 1989. Further tests of an exemplar-similarity approach to relating identification and categorization. Perception & Psychophysics 45. 279–90.

Nusbaum, Howard C., and T. M. Morin. 1992. Paying attention to differences among talkers. Speech perception, speech production, and linguistic structure, ed. by Y. Tohkura, E. Vatikiotis-Bateson and Y. Sagisaka, 113–54. Washington: IOS Press.

Nygaard, Lynne C., M. S. Sommers, and D. B. Pisoni. 1994. Speech perception as a talker-contingent process. Psychological Science 5. 42–6.

Owren, Michael J., and G. C. Cardillo. 2006. The relative roles of vowels and consonants in discriminating talker identity versus word meaning. Journal of the Acoustical Society of America 119. 1727–39.

Palmeri, Thomas J., S. D. Goldinger, and D. B. Pisoni. 1993. Episodic encoding of voice attributes and recognition memory for spoken words. Journal of Experimental Psychology: Learning Memory and Cognition 19. 309–28.

Perrachione, Tyler K., J. B. Pierrehumbert, and P. C. M. Wong. 2009. Differential neural contributions to native- and foreign-language talker identification. Journal of Experimental Psychology: Human Perception and Performance 35. 1950–60.
——, J. Y. Chiao, and P. C. Wong. 2010. Asymmetric cultural effects on perceptual expertise underlie an own-race bias for voices. Cognition 114. 42–55.
Perry, Theodore L., R. N. Ohde, and D. H. Ashmead. 2001. The acoustic bases for gender identification from children's voices. Journal of the Acoustical Society of America 109. 2988–98.
Pierrehumbert, Janet B., T. Bent, B. Munson, A. R. Bradlow, and J. M. Bailey. 2004. The influence of sexual orientation on vowel production (L). Journal of the Acoustical Society of America 116. 1905–8.
Remez, Robert E., J. M. Fellowes, and P. E. Rubin. 1997. Talker identification based on phonetic information. Journal of Experimental Psychology. Human Perception and Performance 23. 651–66.
Rendall, Drew, M. J. Owren, and P. S. Rodman. 1998. The role of vocal tract filtering in identity cueing in rhesus monkey, Macaca mulatta vocalizations. Journal of the Acoustical Society of America 103. 602–14.
——, P. S. Rodman, and R. E. Emond. 1996. Vocal recognition of individuals and kin in free-ranging rhesus monkeys. Animal Behaviour 51. 1007–15.
Rost, Gwyneth C., and B. McMurray. 2009. Speaker variability augments phonological processing in early word learning. Developmental Science 12. 339–49.
——, and ——. 2010. Finding the signal by adding noise: the role of noncontrastive phonetic variability in early word learning. Infancy 15. 608–35.
Saygin, Ayse P., F. Dick, S. W. Wilson, N. F. Dronkers, and E. Bates. 2003. Neural resources for processing language and environmental sounds: evidence from aphasia. Brain 126. 928–45.
Schacter, Daniel L., and B. A. Church. 1992. Auditory priming: implicit and explicit memory for words and voices. Journal of Experimental Psychology. Learning, Memory, and Cognition 18. 915–30.
Schmale, Rachel, A. Cristià, A. Seidl, and E. K. Johnson. 2010. Developmental changes in infants' ability to cope with dialect variation in word recognition. Infancy 6. 650–62.
——, and A. Seidl. 2009. Accommodating variability in voice and foreign accent: flexibility of early word representations. Developmental Science 12. 583–601.
Simpson, Adrian P. 2009. Phonetic differences between male and female speech. Language and Linguistics Compass 3. 621–40.
Singh, Leher. 2008. Influences of high and low variability on infant word recognition. Cognition 106. 833–70.
——, K. White, and J. Morgan. 2008. Building a word-form lexicon in the face of variable input: influences of pitch and amplitude on early spoken word recognition. Language Learning and Development 4. 157–78.
Singh, Sadanand, and T. Murry. 1978. Multidimensional classification of normal voice qualities. Journal of the Acoustical Society of America 64. 81–7.
Staum Casasanto, Laura. 2008. Does social information influence sentence processing? Proceedings of the 30th Annual Conference of the Cognitive Science Society, ed. by B. C. Love, K. McRae, and V. M. Sloutsky, 799–804. Austin, TX: Cognitive Science Society.
Stoddard, P. K. 1996. Vocal recognition of neighbors by territorial passerines. Ecology and evolution of acoustic communication in birds, ed. by D. E. Kroodsma and E. H. Miller, 356–74. Ithaca: Comstock Books.
Van Berkum, Jos J., D. van den Brink, C. M. Tesink, M. Kos, and P. Hagoort. 2008. The neural integration of speaker and message. Journal of Cognitive Neuroscience 20. 580–91.
Van Lancker, Diana, J. Kreiman, and K. Emmorey. 1985a. Familiar voice recognition: patterns and parameters Part I: recognition of backward voices. Journal of Phonetics 13. 19–38.
——, ——, and ——. 1985b. Familiar voice recognition: patterns and parameters Part II: recognition of rate-altered voices. Journal of Phonetics 13. 39–52.
Waser, Peter M. 1977. Individual recognition, intragroup cohesion and intergroup spacing: evidence from sound playback to forest monkeys. Behaviour 60. 28–74.
Weary, D. M., and J. Krebs. 1992. Great tits classify songs by individual voice characteristics. Animal Behaviour 43. 283–7.
Werker, Janet, and S. Curtin. 2005. PRIMIR: a developmental framework of infant speech processing. Language Learning and Development 1. 197–234.
Zatorre, Robert J., and P. Belin. 2001. Spectral and temporal processing in human auditory cortex. Cerebral Cortex 11. 946–53.
Zuberbühler, K. 2002. A syntactic rule in forest monkey communication. Animal Behaviour 63. 293–9.