Journal of Memory and Language 65 (2011) 264–285

Contents lists available at ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

On-line acoustic and semantic interpretation of talker information

Sarah C. Creel *, Melanie A. Tumlin

Department of Cognitive Science, University of California, San Diego, United States

ARTICLE INFO

Article history:
Received 20 May 2010
revision received 13 June 2011
Available online 8 July 2011

Keywords:
Spoken language comprehension
Eye tracking
Word learning
Representational specificity
Talker variability
Talker characteristic

ABSTRACT

Recent work demonstrates that listeners utilize talker-specific information in the speech signal to inform real-time language processing. However, there are multiple representational levels at which this may take place. Listeners might use acoustic cues in the speech signal to access the talker's identity and information about what they tend to talk about, which then immediately constrains processing. Alternatively, or simultaneously, listeners might compare the signal to acoustically-detailed representations of words, without awareness of the talker's identity. In a series of eye-tracked comprehension experiments, we explore the circumstances under which listeners utilize talker-specific information. Experiments 1 and 2 demonstrate talker-specific recognition benefits for newly-learned words both in isolation (Experiment 1) and with preceding context (Experiment 2), but suggest that listeners do not strongly semantically associate talkers with referents. Experiment 3 demonstrates that listeners can recognize talkers rapidly, almost as soon as acoustic information is available, and can associate talkers with multiple arbitrary referents. Experiment 4 demonstrates that if talker identity is highly diagnostic on each trial, listeners readily associate talkers with specific referents, but do not seem to make such associations when diagnostic value is low. Implications for speech processing, talker processing, and learning are discussed.

Introduction

Perceivers are skilled at extracting vast amounts of information from their environments, and recent work suggests that much of this information is employed to comprehend language in real time. A particularly interesting set of environmental information comes from the speech signal itself. In addition to phonemic information (Allopenna, Magnuson, & Tanenhaus, 1998; Gaskell & Marslen-Wilson, 2002), the speech signal contains important non-phonemic information: prosody (Watson, Tanenhaus, & Gunlogson, 2008), vocal emotional cues (Morton & Trehub, 2001), and acoustic correlates of talker identity (Creel, Aslin, & Tanenhaus, 2008; Nygaard, Sommers, & Pisoni, 1994; Palmeri, Goldinger, & Pisoni, 1993). Talker-varying acoustic cues can be specific enough to identify the talker (Fellowes, Remez, & Rubin, 1997; Remez, Fellowes, & Rubin, 1997; Van Lancker, Kreiman, & Emmorey, 1985; Van Lancker, Kreiman, & Wickens, 1985), but even when they do not identify the exact individual, they can provide information about the talker's sex and age (Peterson & Barney, 1952), height and weight (Krauss, Freyberg, & Morsella, 2002); perceived honesty (Apple, Streeter, & Krauss, 1979), perceived femininity (Ko, Judd, & Blair, 2006), social class (Labov, 1972), region of origin (Clopper & Pisoni, 2004a), and sexual orientation (Pierrehumbert, Bent, Munson, Bradlow, & Bailey, 2004). These sorts of information are sometimes referred to as "indexical" cues in the speech signal (Abercrombie, 1967; Ladefoged & Broadbent, 1957; Peirce, 1903/1998). Indexical information is linked to what the talker is likely to say, and how they are likely to say it. Numerous studies suggest that listeners use such acoustically-specific

* Corresponding author. Address: UC San Diego Cognitive Science 0515, 9500 Gilman Drive, La Jolla, CA 92093-0515, United States. Fax: +1 858 534 1128.
E-mail address: creel@cogsci.ucsd.edu (S.C. Creel).

0749-596X/$ - see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2011.06.005

information in on-line language processing (Creel et al., 2008; Goldinger, 1996, 1998; Van Berkum, van den Brink, Tesink, Kos, & Hagoort, 2008), but the exact mechanisms by which this happens are not well-specified.

The focus of the current study is on how listeners utilize indexical variability to constrain on-line processing. When listeners show facilitated processing for a familiar voice, why do they do so? At least two types of encoding of non-phonemic acoustic information have been investigated. On the one hand, listeners might use this acoustic information to identify the person talking, and know that person likes to talk about particular topics. On the other hand, the listener may have (implicitly) encoded a particular acoustic version of a particular word produced by that speaker, without necessarily being aware of the speaker's identity. The first alternative might be construed as a semantic encoding (Geiselman & Bellezza, 1976, 1977; Geiselman & Crawley, 1983; Van Berkum et al., 2008): listeners may use acoustically-specific information in the speech signal to activate representations of people. Those representations of people, then, are linked to knowledge about how (the listener believes) those people think and act. For instance, a listener might hear someone who speaks with a fundamental frequency (*f*0) of 190 Hz and activate knowledge that person is likely to enjoy shoe shopping, relative to someone with an *f*0 of 100 Hz. Similarly, a listener may keep track of how a particular partner refers to a particular entity (Horton & Gerrig, 2002, 2005; see review by Brennan and Hanna (2009)). Note that the listener's inference about the person may not necessarily be accurate, but is based on that listener's beliefs about particular individuals or groups. This will be referred to as *talker-semantic* information—using acoustic cues in the speech stream to encode or access semantic information about a talker.

The second manner of encoding acoustically-specific information is that, when a word is presented in a realization that is acoustically similar to existing representations, those representations may be more readily accessed and thus recognized faster. This role of acoustically-specific information assumes that the listener encodes the acoustic specifics in a word, either integrally as part of the word's representation, or as a conjunction of that word and a prelexical encoding (McQueen, Cutler, & Norris, 2006; Sjerps & McQueen, 2010) of the talker's particular speech style (e.g., Eisner & McQueen, 2005; Jesse, McQueen, & Page, 2007). On an acoustic-encoding account, a recurrence of that word produced by a talker who said it previously (a particular acoustic instance of "shoe" by a particular talker) should match memory better than a production by another talker, or even a production by the same talker in a different context. This will be referred to as *acoustic match* information—facilitated on-line comprehension due to a closer match to a word's previously-experienced sound-form. Note that the acoustic information that facilitates on-line comprehension does not require the listener to identify the talker, it simply needs to be consistent across instances.

One thing that is not well understood is when and how strongly these different types of talker-specific encoding take place. Some research suggests that acoustic match effects are strongest for individuals who are inexperienced in a language: children (e.g., Houston & Jusczyk, 2000, 2003; Schmale & Seidl, 2009) and second-language (L2) learners (e.g., Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1996; Lively, Logan, & Pisoni, 1993). These individuals may have difficulty recognizing a word or speech sound contrast when it is produced by a new speaker, dissimilar to the one they learned from. Native adults are adept at recognizing natively-produced words regardless of who says them, suggesting that acoustic-match effects decline as expertise increases. Nonetheless, even native adults still show acoustic-match facilitation under some circumstances (e.g., Goldinger, 1996; Nygaard et al., 1994).

Some authors have suggested (Belin, Fecteau, & Bédard, 2004; Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; González & McLennan, 2007) that adults maintain neurologically and cognitively distinct representations for word-forms and for talkers, even though both sets of representations are derived from the same signal. One might also imagine that the same representations contain talker and word information, but that listeners allocate attention to different dimensions of the representations depending on the task, as described by Nosofsky (1989). Either way, it may be that adults can attend to who is talking when this information is relevant to the task at hand—comprehension—but may not attend when this information is deemed uninformative.

Supporting the idea that listeners do not always attend to talker identity, Vitevitch (2003) found that 40% of listeners in a word-shadowing task did not detect a change in talker midway through the task. This "change deafness" effect is reminiscent of change blindness in visual situations (e.g., Rensink, O'Regan, & Clark, 1997); in both cases the apparent deafness or blindness is thought to be a result of attention not being allocated toward the changed aspect of the stimulus. For Vitevitch's task, listeners were presumably allocating attention more to the phonemic aspects of the presented words so that they could produce them accurately. If adult comprehenders have learned how to parse out different aspects of the speech signal for different purposes, they should be able to attend flexibly to talker identity or to phonemic information when the need arises. That is, talker information could be gated "on" or "off" depending on what information listeners are trying to glean, and would be apprehended separately from phonemic information. Of course, it is clear from research on selective attention in speech that listeners cannot completely selectively attend to word information while talker varies (Green, Tomiak, & Kuhl, 1997; Mullennix & Pisoni, 1990). This suggests that attending to phonemic information may entail attending to talker-specific acoustic attributes, even if talker *identity* is not attended: talker-semantic encoding is not automatic, but acoustic-match encoding may be automatic.

In practice, when processing is facilitated by talker-specific information, it is difficult to separate facilitation based on talker-semantic encoding from facilitation based on acoustic-match encoding. Suppose a listener hears female Voice X say "shoe" and male Voice Y say "horse." Later, the listener hears Voice X say "shoe" and Voice Z say "horse." The listener is more likely to recognize "shoe" as

a previously-presented word than "horse" (Goldinger, 1998; Palmeri et al., 1993). Is this because "shoe" has been paired semantically with indexical attributes in Voice X, because there is a better acoustic match to the recent presentation of *shoe*, or both? In the next two sections, we describe evidence for each of these uses of acoustically-specific information. Then, we outline our approach to exploring how and when listeners use such information.

Talker semantic activation in on-line processing

One way in which the listener might show talker-specificity effects in language processing is by using talker-specific information to access their knowledge about the individual speaking. That knowledge about the speaker then influences encoding and constrains processing. Geiselman and colleagues (Geiselman & Bellezza, 1976, 1977; Geiselman & Crawley, 1983) formulated a proposal based on studies of listeners' incidental encodings of the talker for various sentences. Their *voice-connotation hypothesis* suggested that inferred properties of the talker, derived from the speech signal, shade the meaning of what is said. That is, the specifics of sound do not need to be stored, merely the connotation of the sentence. For instance, Geiselman and Bellezza (1977) found that listeners in a sentence-memory task tended to confuse the gender of the sentential agent with the gender of the voice: for instance, listeners were biased to recall a female speaker saying "The queen spent the money" but a male speaker for "The king spent the money." This suggests that memory for talker interacted with the semantics of the sentence. Though the dimension they varied was gender, one might imagine that this would extend to other socially-relevant dimensions, such as age or social class. It also seems that at least some talker characteristics can have gradient, rather than dichotomous, effects on processing: Ko, Judd, and Stapel (2009) found that the perception of vocal femininity in a (fictitious) job applicant's voice modulated listeners' ratings of competence.

More recently, Van Berkum et al. (2008) extended this idea of semantic activation from talker acoustics to on-line spoken language processing, with a wider range of between-talker variables. They presented listeners with sentences spoken by talkers who differed in gender, age, or social status, while measuring event-related potentials (ERPs). Each sentence had a target word, such as wine in "Before I go to bed I like to have a glass of wine." The authors manipulated who spoke the sentence (e.g., a child vs. an adult). They found that listeners produced a larger semantic mismatch ERP (an N400) when the sentence was spoken by an incongruent speaker (a child) rather than a congruent speaker (an adult). This research suggests that listeners used acoustic cues to talker identity to constrain predictions about upcoming sentence material.

In sum, there is ample evidence that listeners use indexical cues in the speech signal to make high-level semantic inferences. This use of talker acoustics is related to the use of discourse context or real-world context: given the current situation, here are the things that are likely to be spoken. Acoustics are only a means to an end, in that many other cues might work equally well as long as they identify the talker (such as a picture of a talker, or an introductory "And then Nancy Drew said," in printed text). This is consistent with a larger literature in sentence processing showing that comprehenders are able to make nuanced predictions of sentence continuations by constructing mental representations of events and their likely participants (Altmann & Kamide, 1999; Kamide, Altmann, & Haywood, 2003; see also Bicknell, Elman, Hare, McRae, & Kutas, 2010). It is also consistent with a literature on partner-specific use of referring expressions in dialog (Brennan & Hanna, 2009; Horton & Gerrig, 2002, 2005; Horton & Keysar, 1996). Thus, one might expect talker identity to constrain online processing much in the same way that sentence context can constrain processing. For instance, Dahan and Tanenhaus (2004) found that verbs like "snack on" in a sentence such as "He likes to snack on candy" can short-circuit phonological competition with an implausible competitor like *candle* (which sounds like *candy* at the beginning of the word, but is far less tasty), with looks to the candy exceeding looks to the candle even before the onset of "candy." Activating knowledge about the identity of the talker might work in exactly the same way, with knowledge about the talker and their mental state constraining reference before the phonological form of a word is available.

Acoustic matching in on-line processing

A second way in which acoustic details of speech can influence on-line processing is by providing a closer acoustic match to previous instances of a particular word. In fact, experimental evidence from young children (e.g., Houston & Jusczyk, 2000) and second-language (L2) learners (e.g., Lively et al., 1993)—both new to a particular language—suggests that they must learn what information in the speech signal changes words' meanings, and what does not. Early in the acquisition process, memory representations of spoken words are highly acoustically specific: young infants have trouble generalizing past the acoustic specifics of their input (Houston & Jusczyk, 2000; Schmale & Seidl, 2009; Singh, Morgan, & White, 2004). Infants also seem to have more difficulty attending to unfamiliar talkers than to familiar ones when speech is embedded in noise (Barker & Newman, 2004), suggesting that talker familiarity plays a role in maintaining attention to speech input. Later in development, infants are better able to generalize to new acoustic realizations, not hindered by talker-related acoustic variability (Houston & Jusczyk, 2000; Schmale & Seidl, 2009). These infant data, suggesting an initial sensitivity to talker with decreasing sensitivity across the first year of life, parallel a widely-held perspective of speech sound acquisition as a narrowing-down from language-universal speech sound contrasts to language-specific speech sound contrasts over the first year of life (Werker & Tees, 2002). After discerning what sounds are meaningful in their native language (phonemes but not talkers), learners are able to accomplish word learning and recognition using only language-relevant information.

Further evidence, though, suggests that sensitivity to acoustic specifics persists well past infancy. Infants at 14 months benefit from talker variability when learning

new words (Rost & McMurray, 2009), suggesting that they still need help figuring out what attributes of the speech signal should remain constant across multiple instances of a single word. Recent work by Richtsmeier, Gerken, Goffman, and Hogan (2009) finds that preschoolers also benefit from high-talker-variability learning. Among adults, listeners benefit from multiple talkers when learning new L2 speech contrasts (Bradlow et al., 1996; Lively et al., 1993) or dialect characteristics (Clopper & Pisoni, 2004b). That is, hearing a range of different talkers producing the same word or accent results in a more generalizable representation. This suggests that across a wide age range, listeners can be sensitive to the talker-specific aspects of speech.

Even adult native speakers are influenced by talker-specific acoustic information in processing. Goldinger (1996) presented listeners with a study list of spoken words from various talkers. Then, listeners were given a test list of words, again from various talkers, and asked to identify each word as old (i.e., previously heard) or new. Listeners were more accurate at judging that a word was "old" when the test talker matched the original talker than when it mismatched the original talker. In a similar task, Palmeri et al. (1993) showed that this talker-specificity effect was not explained by semantic encoding of talker gender with each word: even when a new voice saying an old word had the same gender as the old voice, listeners showed worse recognition. Thus, listeners benefited from acoustic-matching to a previous voice. As mentioned earlier, this acoustically-specific encoding may occur because phonemic information itself is not completely separable from talker-specific acoustic factors (Green et al., 1997; Mullennix & Pisoni, 1990). That is, even adult listeners must use talker-varying acoustic characteristics in order to recognize phonemes.

More recently, Creel et al. (2008) showed that talker-specific acoustics influence on-line word recognition. They presented pairs of words with early phonological overlap (such as *sheep* and *sheet*) to listeners repeatedly throughout the experiment. For some pairs, the talker for both words matched (the male talker said both *sheep* and *sheet*), while for other pairs, the talker differed between the two words (the male talker said *sheep*, the female talker said *sheet*). As each word was heard, listeners selected the corresponding picture out of four pictures displayed on a computer screen. Eye movements to each picture were tracked during the word, to assess how strongly the listener was considering each alternative. By the last block of trials, listeners were significantly less likely to visually fixate a different-talker overlapping picture (female *sheet* as they were hearing male "sheep") than to fixate a same-talker overlapping picture (male *sheet* as they were hearing male "sheep"). A vocabulary-learning experiment, conducted over two days to allow time for word learning, found temporally later but similar effects for novel nonsense words learned as labels for nonsense pictures. These results suggest that listeners do incorporate talker-specific acoustic information into their representations of words (Creel et al., 2008; Goldinger, 1996) and that it is not an artifact of simple semantic encoding of talker gender (Palmeri et al., 1993). We note briefly that results showing within-gender talker-specificity effects do not necessarily rule out semantic encoding. For instance, listeners in Palmeri et al. (1993) might have encoded vocal femininity (Ko et al., 2009) semantically, with talker-specific facilitation resulting from a semantic match to the particular level of femininity. Of course, using the acoustics to make femininity judgments implies that there exist acoustic representations of feminine voice quality to provide a basis for the judgments.

Distinguishing talker-semantic and acoustic-match effects

As outlined above, there are at least two ways in which acoustic specifics might be utilized by a listener in comprehending spoken language, and there is evidence favoring both. Problematically, the influences of these two types of encoding are difficult to pull apart experimentally. For instance, in the word-learning paradigm used by Creel et al. (2008), listeners might be learning one of two things (or both) about *sheep* and *sheet*. First, they might be learning a talker-semantic association between each talker and each referent (the female talker likes to talk about sheep; see Horton & Gerrig, 2002, 2005, for evidence that such associations are learned). Second, as Creel et al. argue, talker-specificity effects might result from encoding both acoustic characteristics and phonemic characteristics of the word form (this particular token of "shee-" is highly familiar and linked to *sheep*). Third, and quite reasonably, *both* types of associations might be formed. Importantly, either or both types of learning would result in a talker-specificity benefit in processing, which in Creel et al. was realized as visual fixations to talker-associated pictures.

The current study explored how listeners use talker information in on-line language processing, using an eye-tracked word recognition paradigm. We began with an artificial vocabulary, rather than real words, because newly-learned words tend to show stronger evidence of talker-specificity effects than do real words (Creel et al., 2008). This is consistent with the notion that novel words will have fewer (zero) traces than real words, allowing talker-specific traces to constitute the entirety of their representations (though they likely are also affected by phonologically-similar traces). Experiment 1 replicated Creel et al.'s second experiment with a simpler paradigm. Listeners learned novel words as labels for 16 unfamiliar pictures, to form a baseline of acoustic-match effects in on-line recognition. Experiment 2 then presented the same artificial vocabulary, but trained participants in the sentence-frame "Click on the X", where X was one of 16 picture labels.

Experiment 1

The goal of Experiment 1 was to follow up on Creel et al. (2008) by developing a word-learning paradigm that could be run in a single test session. Our first step was to come up with a streamlined artificial vocabulary learning task in which talker-specificity effects are evident. (The analogous task in Creel et al., 2008, required two days of training, which is not a trivial accomplishment in the world of undergraduate research participants.) The current task

was intended as the basis of comparison for following experiments. Accordingly, we designed a 16-word vocabulary, which is described in more detail below. The purpose of using an artificial vocabulary, rather than known words, was to increase the visibility of the talker-encoding effect, which was larger for novel words in Creel et al. (Experiment 2) than for known words (Experiment 1). Goldinger (1998)'s results in a shadowing task also suggest that less-frequent words—which, on an exemplar view, would have fewer episodic memory traces—are more susceptible to talker-specificity effects.

The vocabulary consisted of early-overlapping novel-word pairs, such as *boog* and *booj*. Normally, these two words cannot be distinguished until they diverge from each other phonemically (at the /g/ or the /ʤ/, though coarticulatory information might distinguish them slightly sooner). However, as in Creel et al. (2008), an extra talker "feature" distinguished some of the pairs (see Table 1). This meant that for some listeners, *boog/booj* was a *different-talker pair*, in which the female talker always said *boog* and the male talker always said *booj*. For other listeners, *boog/booj* was a *same-talker pair*, in which the female talker always said *boog* and the same female talker always said *booj*. Henceforth, we will refer to trials where pictures from a different-talker pair are displayed as *different-talker trials*, and where pictures from a same-talker pair are displayed, *same-talker trials.* Our hypothesis was that, as in Creel et al., listeners would visually fixate the correct picture sooner on different-talker trials than on same-talker trials. That is, listeners would be able to use talker information to tell apart the two words sooner than using phonemic information alone.

Table 1
Counterbalanced assignment of talkers to words for different participants.

Word pair		IPA	Participants			
			Ppt. 1	Ppt. 2	Ppt. 3	Ppt. 4
1	boog	/bug/	**Male**	**Female**	Male	Female
	booj	/buʤ/	**Female**	**Male**	Male	Female
2	belm	/bɛlm/	**Female**	**Male**	Female	Male
	beln	/bɛln/	**Male**	**Female**	Female	Male
3	darg	/dɑɹg/	Male	Female	**Female**	**Male**
	darge	/dɑɹʤ/	Male	Female	**Male**	**Female**
4	dalm	/dɑlm/	Female	Male	**Male**	**Female**
	daln	/dɑln/	Female	Male	**Female**	**Male**
...						

Note: IPA = International Phonetic Alphabet; Ppt. = Participant. The same counterbalancing was repeated for the other four pairs (/vig/ /viʤ/, /vɔrm/ /vɔrn/, /zɛlm/ /zɛln/, /zɝg/ /zɝʤ/). Pairs in **bold** are different-talker pairs for that participant.

Method

Participants

Thirty-two members of the UCSD undergraduate community took part in the experiment for course credit. All were native speakers of English. One additional participant did not achieve a criterion of 90% correct word identification on training trials within the 2-h time frame of the experiment, and was replaced with another participant.

Materials

Participants learned 16 novel words as labels for unfamiliar black-and-white shapes (see examples in Fig. 1). These shapes have been used in several previous experiments (Creel, Aslin, & Tanenhaus, 2006; Creel, Tanenhaus, & Aslin, 2006; Creel et al., 2008). There were four assignments of shapes to words, with the only constraint on assignment being that any two shapes appearing together did not have the same base shape (e.g., two shapes that both contained a large, solid circle did not appear together).

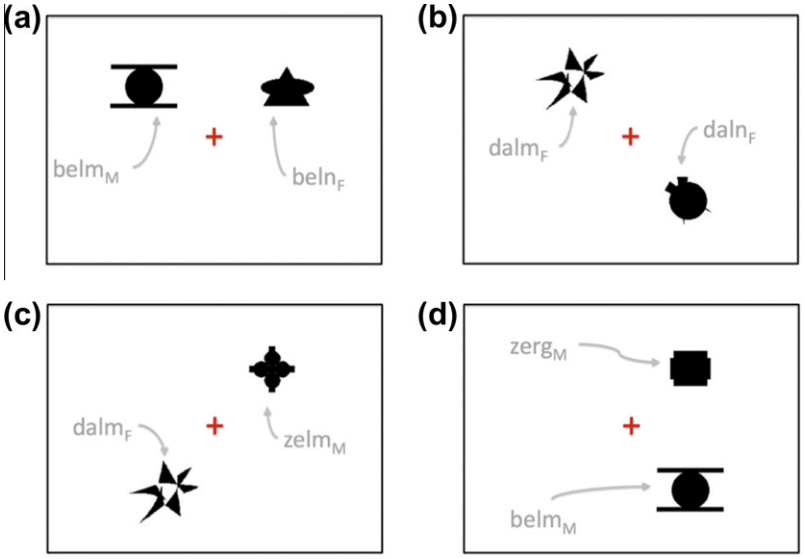

Fig. 1. Four trial types. Subscripts on words indicate talker gender. (a) Onset-match trial with different talkers; (b) onset-match trial with same talker; (c), (d) control trials. Elements in gray were not displayed to participants and serve only to identify objects to the reader. Each picture appeared with only two other pictures throughout the experiment.

The 16 words in the artificial vocabulary consisted of eight pairs of words that overlapped in their onset consonant and nucleus (see Table 1). Each pair shared all but the final consonant. Words were drawn from a set of low-frequency diphones as determined by use of an online phonotactic frequency calculator (Vitevitch & Luce, 2004; details in Appendix). Previous research indicates that lower-frequency words are more susceptible to talker-specificity effects (Goldinger, 1998), presumably because listeners have had fewer exposures to such words, and thus more mutable representations. We presumed that the same would hold true for low-frequency diphones, making talker-specificity effects more evident.

Novel words were additionally constrained in three ways. First, all words had voiced initial consonants so as to provide *f*0 information—a strong correlate of talker identity—as early as possible. Second, all words had voiced final consonants to lengthen the syllable nuclei, as voiced codas in English are preceded by longer nuclei than are unvoiced codas. This lengthened the period of ambiguity between the two words in an onset-overlapping pair. Third, to avoid gross changes to the vowels within a pair, coda consonants in a pair were matched for nasalization and excluded liquids (l, r), though for some pairs, the same liquid was used in the nucleus (e.g., *dalm*, *daln*) to create low-frequency diphones. These constraints should provide listeners with talker information and time to process it prior to disambiguating phonemic information.

Words were recorded from one female talker and one male talker in a sound-attenuated chamber. We chose to use one male and one female in order to have readily evident acoustic differences between talkers. After recording at least two tokens of each word, we selected one token to be used for each word for a given talker based on recording quality and match to the other words in terms of pitch contour. Following selection, all sound files were adjusted to 70 dB SPL in Praat (Boersma & Weenink, 2007), making them identical in overall amplitude. Talkers' averaged productions of each vowel are summarized in Table 2. The two talkers differed in all characteristics ($p < .05$) except for first formant frequency ($p = .22$).

Procedure

There were two phases in this experiment: a training phase, and a test phase. The training phase provided trial-by-trial feedback on correctness, and continued in 128-trial blocks until the participant had scored 90% correct or better in a block. The two-block test phase was identical to the training phase, except that no feedback was provided.

Table 2
Fundamental frequency (*f*0) and first three formant frequencies (F1–F3) of each talker for vowels used in Experiment 1.

Vowel	Male talker				Female talker			
	*f*0	F1	F2	F3	*f*0	F1	F2	F3
/i/	116	282	2587	3242	164	298	2855	3234
/ɛ/	117	636	1639	2832	170	846	1720	2943
/a/	115	795	1150	2533	168	707	1533	2764
/ɔ/	113	416	694	2779	174	478	926	2810
/u/	122	308	1229	2461	184	314	1517	2753
/ɝ/	115	360	1394	1622	169	543	1508	1715

Note: F = formant.

On each training trial, pictures appeared in two of the four screen locations (Fig. 1). The four 200×200 pixel locations were at the points of a square surrounding the center, at 25% of screen width and 20% of height, 75% and 20%, 25% and 80%, and 75% and 80%. The screen was 12×15 in (30.5×38 cm), with resolution set to 1152×864 pixels, and the screen was roughly 24 in. (60 cm) from the participant's eye. Each picture appeared equally often in each screen position, both as the correct choice and as the incorrect choice. Each picture was the target on eight trials per block. At 500 ms after picture onset, the name of one picture was spoken, and the participant guessed which one was the picture named by selecting it with a computer mouse.

Each training trial provided accuracy feedback: once the participant selected one picture, the incorrect picture disappeared and only the correct one stayed on the screen, revealing the correct answer. (See also Magnuson, Tanenhaus, Aslin, & Dahan, 2003, for use of this procedure.) After reaching criterion performance in training (90% correct across a block of training), the participant completed the test phase. Trials in the test phase were identical to the training trials except that they provided no feedback. Eye movements were tracked throughout the procedure.

For a given participant, each shape was the target on two types of trials, onset-matched trials and control trials. On onset-matched trials, the two shapes that appeared were the two members of a pair, such as *boog* and *booj*. On control trials, the same target shape appeared with a particular shape with a dissimilar name, such as *boog* and *zerg*. The *boog* occurred only with these two other shapes. The reason for constraining the number of shapes that appeared together, rather than presenting all possible pairs, was so that we could obtain a sufficiently high number of paired (*boog*/*booj*) trials for analysis. Control (*boog*/*zerg*) trials were matched to experimental trials in the sense that the incorrect shape occurred with that target with equal frequency, so that the only difference between experimental and control trials was the similarity of the label.

Each listener received four different-talker pairs and four same-talker pairs. Across all participants, each word occurred equally in all conditions. That is, for 25% of participants, *boog* was spoken by a male, and *booj* by a female; for another 25%, *boog* was female and *booj* was male; for a third 25%, both *boog* and *booj* were male; and for the last 25%, *boog* and *booj* were both female. Thus, half the time, a particular pair was a different-talker pair, and the rest of the time, it was a same-talker pair. This resulted in four assignments of words to talkers (see Table 1). Importantly, the structure of the artificial vocabulary was such that participants were not *required* to use talker information at all in order to succeed. That is, if talker differences were removed, listeners could still perform with perfect accuracy by using the phonemic information in the words. There were four quasirandom assignments of pictures to words to reduce the effects of particularly easy or difficult word–picture mappings. Combining the four word–talker assignments with the four word–picture mappings generated 16 unique conditions, each of which was run twice.

It is important to note that, *for a given participant, every instance of a particular word was always spoken by the same talker*: word and talker were perfectly correlated.

Equipment and software

Participants were tested individually in a sound-treated room on an Eyelink Remote eye tracker (SR Research, Mississauga, Ontario; www.sr-research.com), which samples gaze position every 4 ms in remote mode. The eye tracker was operated by a PC tower running Eyelink software in DOS mode. A second computer (a Mac) presented experimental stimuli in Matlab, using custom software that relied on the PsychToolBox 3 (Brainard, 1997; Pelli, 1997) and the embedded Eyelink Toolbox (Cornelissen, Peters, & Palmer, 2002). At important time points in the experiment (trial start, sound onset, and answer selection), the Mac sent messages to the PC, which interpolated these messages with time stamps into the eye tracking data stream. The eye tracking ended when the participant selected a response. Each participant wore a small target-like sticker on his or her forehead, allowing the software to pick up the participant's head position and distance, and obviating the need for a chin rest. Pictures displayed on the screen subtended roughly 3° visual angle at a typical viewing distance.

After data collection, files were processed off-line to condense data by variables of interest. Looks within a square region extending 100 pixels above, below, left and right of each location were counted as looks to that location. Looks to physical locations were then recoded as looks to target picture, looks to non-target pictures, looks to other screen areas, looks off-screen, and a "no data" category indicating that the eye was not visible on that sample (e.g., due to blinking). Looks were then averaged across a trial type for each participant. For computational convenience, looks were binned into 50-ms chunks. This means that all gaze-fixation plots display looks centered on a time point (for instance, the time point at 25 ms displays looks from 0 to 50 ms).

The dependent measure was *target advantage* (Fig. 2): the proportion of looks to the correct picture minus the proportion of looks to the incorrect picture. When zero, the two pictures are being fixated equally often, corresponding to equivalent looks to each picture (Fig. 2, top, left side). When positive (Fig. 2, bottom), it indicates that listeners are looking more at the correct picture than the incorrect picture. This number begins to exceed zero about 200 ms after the word being spoken is distinguishable from its alternative(s) (Allopenna et al., 1998), with the delay presumably coming from latency to initiate a saccade (Hallett, 1986). This point will be earlier or later depending on when the listener can distinguish one word from the other.

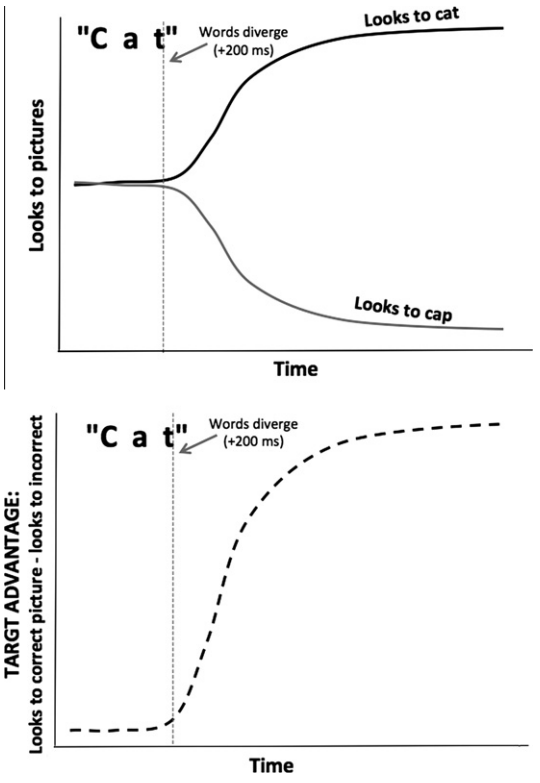

Fig. 2. Illustration of the correspondence between looks to correct and incorrect pictures (top), and target advantage (bottom).

Results

Participants achieved 90% or better performance in about three blocks of training trials ($M = 3.34$, $SD = 1.15$). On the test, different-talker and same-talker trials did not differ in accuracy (95% vs. 93%, $t_1(31) = 1.20$, $p = .24$; $t_2(15) = 1.37$, $p = .19$).

During the test phase, participants fixated the correct picture sooner when they had learned its phonologically-similar competitor from the other talker than when they had learned the competitor from the same talker (Fig. 3). That is, listeners experienced less phonological competition on different-talker trials—when two similar words (*boog, booj*) were learned from different talkers—than on same-talker trials, when two similar words were learned from the same talker.

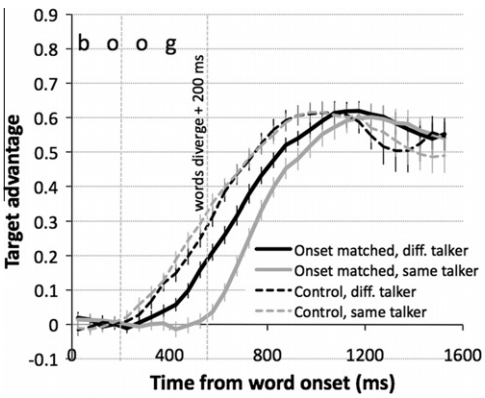

Fig. 3. Experiment 1, preferential target looks (target advantage) over time in onset-matched (solid) and control (dashed) trials. Error bars are standard errors throughout.

As our design was counterbalanced across participants and items, we make statistical decisions based on participants analyses (Raaijmakers, 2003; Raaijmakers, Schrijnemakers, & Gremmen, 1999), reporting by-items analyses (and, where applicable, *minF'* values) by convention. Effect size is reported as generalized eta-squared (Olejnik & Algina, 2003; computational formulas from Bakeman, 2005), which takes into account effect-size inflation in within-participants designs. The data in figures were calculated from by-participants means.

A *t*-test on target advantage for the two levels of Talker (different talker had named each picture, same talker had named each picture) during the time window 200–550 ms tested whether the target looking preference on different-talker trials exceeded looks on same-talker trials.[1] That time window was chosen because (1) the onset of the final consonant of each word occurred at 362 ms on average, and (2) it is standardly assumed that it takes about 200 ms to program and execute an eye movement based on an external signal (Hallett, 1986). This was rounded to the nearest 50-ms interval because data had been downsampled to 50-ms bins; the included bins spanned from just 200 ms to 550 ms (not including 200, but including 550). In all figures, each data point represents the midpoint of a time bin—e.g., a data point at 25 ms represents visual fixations from just after 0 ms up to 50 ms. Thus, the approximate time period where eye movements can be expected based on the ambiguous part of a pair of early-overlap competitors was 0–362 + 200, or (roughly) 200–550 ms. For clarity and conciseness, we do not analyze the control trials, as the theoretical effects of interest are in the onset-matched trials. There was an effect of Talker ($t_1(31) = 2.75$, $p < .01$; $t_2(15) = 1.69$, $p = .11$; mean difference = .059), with higher target advantage on different-talker trials.

Discussion

Participants learned a set of novel words as labels for unfamiliar pictures. After learning, eye tracking measurements indicated that they were able to tell apart two otherwise similar-sounding words learned from *different* talkers (female *boog*, male *booj*) earlier than they could tell apart similar-sounding words learned from the *same* talker (female *boog*, female *booj*). This experiment replicates Creel et al.'s (2008) artificial vocabulary study with a new set of stimuli, and shortens the learning period to a single day.

[1] One note of analytical interest should be made here. In previous uses of this paradigm, trials on which participants made errors were discarded, and we follow this convention here. However, this is slightly problematic in that, presumably, the same decision processes generate both responses and eye movement probability distributions (see, e.g., Ratcliff & Rouder, 1998). When there is uncertainty in this process, noise variability can interfere in both. In cases where participants make errors, this noise influences both the response and the looks at the erroneous alternative prior to a response, so that looks to the wrong shape and wrong responses covary. This means that removing error trials sometimes leads to a pattern where participants are fixating the *correct* alternative at greater-than-chance levels. In most analyses here, if this pattern occurs it should be equivalent across the two conditions to be compared. Note that patterns of significance were nearly identical when error trials were included.

As noted in the Introduction, though, this talker-specificity effect could be coming from one or more uses of talker-related acoustic variability. One possibility is that listeners are encoding talker-related acoustic variability as a sound property, either in the representation of the word form itself or as talker-specific acoustic information (Cutler, Eisner, McQueen, & Norris, 2010; Eisner & McQueen, 2005). Another possibility is that they are processing talker identity as a separate semantic piece of information and using that semantic information to predict pictures, much in the same way as listeners use verb information to constrain phonological competition (Dahan & Tanenhaus, 2004). In fact, it seems logical that they would be doing both: different people certainly think about and talk about different things in the real world, so it would make sense to encode this information.

Of course, this experiment is not the real world. It replicates results of talker-specific information facilitating recognition of isolated words. However, in natural speech, words are nearly always embedded in contexts—acoustic contexts and semantic contexts, as well as syntactic and discourse contexts. Accordingly, in Experiment 2, words were presented in sentence frames ("Click on the X") during training. This means that talker-identifying information is available about half a second earlier than word information (at the latest, at the onset of vocal fold vibration in *click*; it is possible that the burst of sentence-initial /k/ provides talker-identifying information—see Sundara, 2005, on speaker differences in coronal stop bursts).

Context should have different effects depending on what sort of talker information the listener encodes. If listeners' encodings of talker information in Experiment 1 had something to do with talker *identity* (a "semantic" encoding), then talker-specificity effects should be apparent on both words and on the full sentences. If listeners encode the acoustics of test words, then talker-specificity effects might be apparent only on the words, not on the full sentences. It is also possible that listeners associate the entire sentence with the shape (e.g., associating "Clickontheboog" in the female voice with a particular shape). This would predict a drop in talker-specificity on the test trials when the full sentence is no longer available.

Experiment 2

Method

Participants

Thirty-two new participants from the same pool as Experiment 1 took part. An additional five participants were not included in data analysis due to failure to complete the learning phase in the allotted time (3), and poor eye track (2). We tested eight further control participants from the same pool to verify that carrier phrases indeed yielded sufficient information to identify the talkers.

Materials

The same two talkers as in Experiment 1 made new recordings of the same vocabulary items, this time embedded in the sentence "Click on the X." Sound files

were edited as before. To create the sound files that were played during the test trials, which were words in isolation, we extracted the words themselves from the sound files by excising the "Click on the" portion.

Procedure

The procedure was largely similar to Experiment 1, except that on training trials participants heard the sentence "Click on the X" instead of just the word "X." After reaching the 90% correct criterion, participants continued on to two 128-trial blocks of test trials, which were not reinforced. On test trials, only the word was presented, not the sentence. This was done to make the test phase maximally comparable to Experiment 1, so that we could assess whether having talker information available in the lead-in sentence during *learning* had changed the representations formed by learners.

A control talker-identification condition provided an assessment of how rapidly listeners could extract talker-semantic information from the speech signal. In this control condition, a male stick figure and a female stick figure were presented on each trial, as a sound file was played. Instead of associating pictures with words (though the words were present), listeners were asked to click on the figure representing the person who was speaking.

Results

As in Experiment 1, listeners reached 90% correct performance in roughly three blocks of training trials ($M = 2.94$, $SD = 1.05$). Accuracy on different-talker and same-talker trials did not differ during the test (both 94%). Participants were then tested on the words in isolation. Talker-specificity effects on target advantage during the test (Fig. 4) were roughly equivalent to Experiment 1. This was consistent with acoustic-match encoding in context, and also with semantic encoding (relating indexical information to shapes).

The final consonant of each word began at 390 ms on average, so we extended the time window to 200–600 ms (roughly, 0–390 ms plus a 200 ms delay). A *t*-test on target advantage in the time window 200–600 ms with Talker (same-talker, different-talker) as the within-participants factor reached significance, ($t_1(31) = 2.40$, $p = .02$, $t_2(15) = 2.22$, $p = .04$, mean difference = .048), indicating higher target advantage on different-talker trials than same-talker trials.

To assess these results in comparison to Experiment 1, we computed ANOVAs on target advantage with Talker (within participants) and Experiment (between participants) as factors. The time window in the current experiment was truncated to 550 ms, to match Experiment 1. Talker ($F_1(1, 62) = 10.84$, $p = .002$; $F_2(1, 15) = 5.64$, $p = .03$, $minF'(1, 33) = 3.71$, $p = .06$; $\eta_G^2 = .058$) was significant, with a target advantage of .045 for different-talker trials, and −.005 for same-talker trials. Experiment and the Talker × Experiment interaction were not significant (all F's < 1). This does not support a *decrease* in use of talker-specific acoustic information due to a change from sentences to words.

To assess how much talker-semantic encoding might be contributing to the talker-specificity effect, we also checked how much listeners in the current experiment were using talker information in the lead-in sentences to constrain picture selection. Importantly, if listeners were using talker identity, looks on different-talker trials should increase near the beginning of the sentence—as soon as talker identity becomes apparent. That is, target advantage at the beginning of the *sentence* in Experiment 2 should look like target advantage at the beginning of the *word* in Experiment 1. To assess this, we analyzed looks both before and after word onset in each participant's final block of training trials (Fig. 5). During this block, listeners achieved 90% or better accuracy, and were approaching ceiling performance. As a caution, we note that there were fewer trials per participant in this training block (128 trials) than were present in the test trials analyzed (256 trials).

To assess how learning might have differed from Experiment 1, we compared the two experiments' final training blocks in an ANOVA on target advantage, with Experiment (1, 2) as a between-participants factor and Talker as a

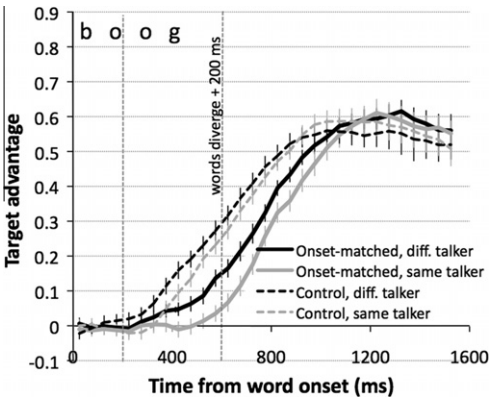

Fig. 4. Experiment 2, target advantage over time for onset-matched (solid) and control (dashed) trials.

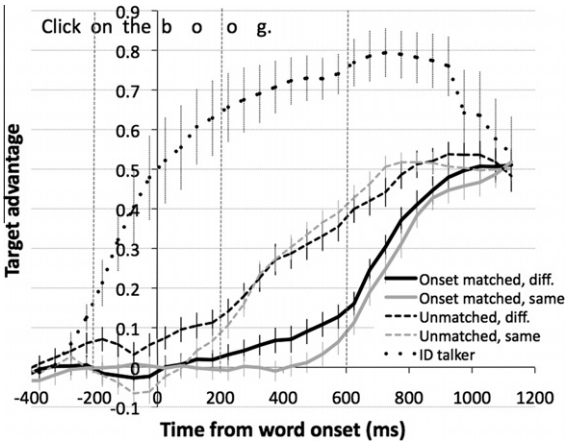

Fig. 5. Experiment 2, target advantage on final block of training trials (thicker lines), compared to control talker-identification condition (dotted black line).

within-participants factor in the post-word-onset time window—200–550 ms—for the final training block in the current experiment and in Experiment 1. If listeners in Experiment 2 were already looking at the correct picture prior to word onset by linking talker identity to a particular shape, then they should show a larger target advantage during the word itself relative to Experiment 1. There was an effect of Talker ($(1, 62) = 4.48$, $p = .04$; $F_2(1, 15) = 5.98$, $p = .03$; $minF'(1, 57) = 2.56$, $p = .12$; $\eta^2_G = .037$), with a mean of .071 for different-talker trials and .014 for same-talker trials. However, there was no effect of Experiment or Talker × Experiment interaction (F's < 1), suggesting that looks in Experiment 2 were not heightened, due to preceding sentence context, relative to Experiment 1.

If talker information semantically constrains processing in the same way that sentence context does, then talker-specific effects should show up prior to word onset. This should happen because indexical information is available well before word onset, but word information is only available at or after word onset. For instance, in Dahan and Tanenhaus' (2004) study on constraining verbs ("snack on"), looks to the more likely alternative (the candy) exceeded looks to its cohort competitor (the candle) prior to word onset. However, listeners in Experiment 2 did not seem to be using talker information prior to word onset. We computed a t-test of the time period beginning 200 ms before word onset to 200 ms after (at which point word information was available as well as talker information; see first time window in Fig. 5). Different-talker trials did not show more looks to the target shape than same-talker trials ($t_1(31) = -0.21$, $p = .84$; $t_2(15) = -.46$, $p = .65$, mean difference = −0.006). Interestingly, a test of the visible difference between same-talker and different-talker control trials *was* significant ($t_1(31) = 2.47$, $p = .02$; $t_2(22) = 2.10$, $p < .05$, mean difference = .09; note that there was a counterbalancing error in the control trials so that not all items occurred in all possible conditions, so the last t-test is run as an independent-samples test in order to include all items). Assuming this result is robust, listeners may have used talker identity information to constrain looking patterns prior to word onset, but only for phonologically dissimilar items—though why this would happen is unclear. One might think that listeners did not show an analogous effect in the experimental trials because they anticipated the upcoming similarity between words—however, this would be inconsistent with what we know about on-line sentence processing, where listeners incorporate the contextual information they have as soon as it becomes available (Dahan & Tanenhaus, 2004; Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995).

A possible, but unlikely, explanation for the weak talker-specificity effects in the lead-in sentences is that lead-ins contained insufficient information to distinguish the two talkers. To rule out this explanation, we ran eight new participants in an eye-tracked talker identification task of 128 trials in length. They heard all *Click on the* sentences used in Experiment 2, but instead of learning names for pictures, they were instructed to click on a picture of a male stick figure when they heard the male talker, and a female stick figure when they heard the female talker. This experiment (dashed black line in Fig. 5) verified that the lead-in material contained sufficient information to distinguish the two talkers: looks during the pre-word time span (200 ms before to 200 ms after word onset) exceeded chance ($t_1(7) = 5.13$, $p = .001$; mean = .47) and also exceeded looks on the different-talker control trials in Experiment 2 ($t_1(8.23) = 4.13$, $p = .006$; mean difference = .39). That is, acoustic cues are available to use talker identity to constrain processing, but listeners seem to be underutilizing them.

Discussion

In this experiment, we wanted to know how presenting talker-identifying information in context (by embedding words in lead-in sentences) would affect listeners' encoding of talker-specific acoustic information. On the one hand, if listeners were making semantic associations between talkers and particular pictures, then listeners should show similar talker-specific effects to Experiment 1, and should also show talker-specific looking patterns during the sentences. On the other hand, if listeners were encoding the acoustic specifics of each word, they should only show talker-specificity effects on the words themselves. Results mostly pointed to the latter alternative: listeners seemed to have encoded talker information along with the words. However, listeners did not seem to use talker *identity* (semantic encoding) to constrain processing during the lead-in sentences, even though a control experiment verified that cues to distinguish the two talkers were available early in the sentence and could drive listeners' eye movements to visual targets.

The results of Experiment 2 suggest a potential divergence between encoding of talker acoustics and talker semantics. On the one hand, listeners are clearly processing acoustic information within the words themselves to identify pictures. On the other hand, listeners appear to glean limited talker-semantic information from introductory sentences, despite evidence from control participants that cues to talker identity in the speech signal are robust. If listeners *can* use talker-identity information in the word-learning and recognition task, why don't they?

When do listeners take full advantage of talker-semantic information?

One potential explanation is that listeners automatically encode acoustic-match information, but that talker-semantic effects require more attention or more processing capacity. Particularly, listeners in a word-learning task might have difficulty encoding so much new information in memory, which prevents them from attending to talker–picture contingencies. Listeners could be encoding at least four types of information: new word forms; new shapes; linkages between word forms and shapes; and linkages between shapes and talkers (or between words and talkers). Any of the first three could be diverting attention from talker-semantic information that listeners would encode in a more naturalistic setting. Alternatively, listeners may simply fail to encode talker identity, doing so only when it is clearly relevant to the situation. It should be kept in mind that another possibility that would look like

Table 3
New information in each experiment.

Experiment	Perceptual information		Associations		Diagnostic value (%)	Talker-specificity effect[a]
	Words	Shapes	Word-shape	Talker-shape		
1	√	√	√	√ (16)	50	.076 ± .168[b]
2	√	√	√	√ (16)	50	.035 ± .093[c,d]
3		√		√ (16)	100	.160 ± .168[e]
4a				√ (16)	50	.029 ± .065[d]
4b				√ (4)	50	.029 ± .203[d]
4c				√ (4)	100	.173 ± .163[d]

Note: Underlined elements had to be encoded to complete a task successfully.
[a] Target advantage on different-talker trials, measured after 16 exposures to each word–talker pairing, starting at beginning of *utterance*.
[b] Related-word trials, single words, final block of training, 200–550 ms after word onset.
[c] Related-word and unrelated-word trials, sentences, final block of training.
[d] From 200 ms before word onset to 200 ms after word onset.
[e] From 200 to 550 ms after sentence onset.

failure to encode is that listeners encode everything, but fail to *use* talker-semantic information because it has low *a priori* probability in the real world. That is, for most words and concepts in a listener's experience, there is little relation to any particular talker. This might change if the listener is aware that a particular talker has a preference for some object or entity, a situation that does occur in the real world. The next two experiments systematically reduce task difficulty to explore the conditions where talker-semantic encoding may be utilized. Both Experiments 3 and 4 eliminate the need to encode new word forms; Experiment 4 additionally eliminates the need to encode novel pictures. Table 3 outlines the different task components for each experiment.

Experiment 3 explores whether listeners have difficulty encoding novel shapes and talker-shape linkages, eliminating the requirement to encode new word forms and word form-shape linkages. Previous work by Horton and Gerrig (2002, 2005) suggests that it is harder for conversational participants to map incoherent (arbitrary) referent categories to particular partners (partners A and B each have knowledge about some frogs and some fish) than it is to map coherent categories to partners (A just knows about fish, B just knows about frogs). Our talker-specific novel shapes are incoherent categories, and thus, it may be relatively difficult to encode talker-identity information about them. Specifically, listeners may not be able to encode talker-shape relationships in the same amount of time it takes them to encode word-shape relationships. If listeners have difficulty mapping an arbitrary set of shapes to talkers, then they should take a longer time to learn talker-shape mappings than listeners in the first two experiments took to learn word-shape mappings, and looks to correct pictures should be fairly slow to emerge. If listeners readily learn mappings between an arbitrary set of shapes and talkers, then they should learn rapidly and show early visual fixations to the target shape.

Experiment 3

In the current experiment, listeners were trained to associate voices and shapes, but did not have to associate *words* with shapes. The shapes were the same ones as used in Experiments 1–2. The voices were also the same, but rather than naming one of the shapes, they claimed to like of one of the shapes. Listeners were asked to learn which one the talker liked, with each talker liking 8 of the 16 shapes.

Method

Participants

Sixteen participants from the same pool as Experiments 1 and 2 took part. Two had participated in earlier experiments a year or more prior to the current experiment. One was in Experiment 2, while the other had participated in a related vocabulary-learning study not reported here. Overall results, as well as statistical significance patterns, were unchanged whether these participants were included or excluded. Analyses presented include both participants.

Materials

Visual stimuli were the same as those used in Experiments 1–2. However, sound stimuli differed from those in Experiments 1–2 in that they did not contain labels for the objects. Sound stimuli were recordings of each talker speaking two phrases, equalized in amplitude. The phrase used during training with feedback and the first block of testing was "Which one do you think I like?" To assess whether listeners had merely overlearned this single test phrase, we switched the phrase during a second block of generalization test trials (leaving talker-shape mappings constant) to "Can you help me find things to play with?" These particular phrases were originally recorded for a child language experiment and were borrowed for the current experiment so that no new recordings had to be made. Visual stimuli were the 16 shapes used in Experiments 1 and 2, which were randomly assigned to appear with each other and to be preferred by one talker or the other.

Procedure

The procedure was quite similar to that for the first two experiments. To match the first two experiments, listeners were trained, in blocks of 128 trials, until they reached 90%

correct performance. On each trial, two pictures appeared. As in Experiments 1 and 2, each picture could appear with one of two other pictures. Because there needed to be a single correct response on each trial, all trials were different-talker trials (one shape belonged to the male talker, one to the female). Listeners received feedback on their correctness on each training trial. Upon reaching criterion performance, they continued to two blocks of unreinforced test trials. The first block of 128 test trials was identical to the training blocks, except that no feedback was provided. The second block of 128 test trials was much the same as the first, except that each talker's sentence was changed from "Which one do you think I like?" to "Can you help me find things to play with?" Visual fixations were tracked throughout.

Results

Participants learned talker-shape pairings faster than word-shape pairings. They reached the 90% correct criterion in an average of 2.0 blocks ($SD = .52$), which was faster than the 3.34 blocks in Experiment 1 (Welch's $t(45.67) = 5.57$, $p < .0001$) or the 2.94 blocks in Experiment 2 (Welch's $t(46.00) = 4.16$, $p = .0001$). This speed suggests that learning associations between talkers and arbitrary shapes is not difficult for participants. Accuracy during the test was 98.8% ($SD = .013$).

We also considered the time course of participants' uptake of talker information on-line (Fig. 6). Specifically, we computed a *t*-test on target advantage from 200 to 550 ms after sentence onset, matching the time window used to compare Experiments 1–2. These analyses only compared different-talker trials, as the current experiment did not contain any same-talker trials. In the first test block, which was matched in number of trials and amount of exposure to Experiment 2's last training block, the target advantage exceeded zero ($t_1(15) = 6.08$, $p < .0001$; $t_2(15) = 9.29$, $p < .0001$; $m = 0.16$). Target advantage in the current experiment exceeded the corresponding target advantage on different-talker trials in test phases of Experiments 1 and 2, in which listeners had *more* exposure to talker-word pairings (three blocks) than in the current experiment (vs. Experiment 1: $t_1(46) = 3.45$, $p = .001$; $t_2(30) = 4.22$, $p = .0002$, $m = .16$ vs. $m = .05$; vs. Experiment 2: $t_1(46) = 4.08$, $p = .0002$; $t_2(30) = 4.61$, $p < .0001$, $m = .16$ vs. $m = .04$). In the second test block, where listeners heard a new phrase from each talker, target advantage also exceeded zero in this early time window ($t_1(15) = 9.45$, $p < .0001$; $t_2(15) = 9.71$, $p < .0001$, $m = .22$).

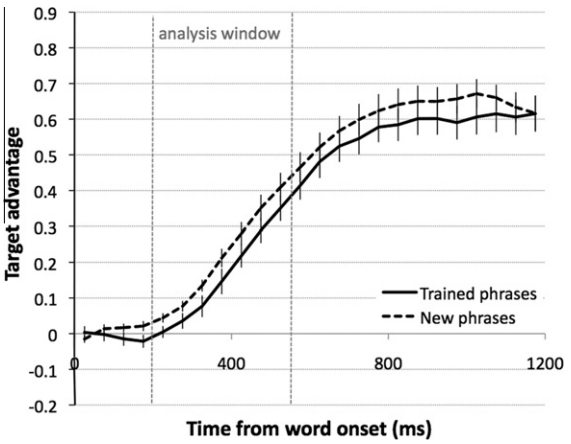

Fig. 6. Experiment 3, target advantage to correct picture (the one that the talker on that trial preferred), when the sentences used during training (solid black) or new sentences (dashed black) were heard. Significance values refer to sentences used in training.

Thus, listeners in this task readily learn arbitrary talker-shape relationships. Further, they are able to recognize which of the two talkers is speaking quite rapidly, even under the burden of learning novel visual shapes and shape-talker relationships. Note that the faster pace of learning in this experiment resulted in an average of two blocks of learning trials. This means that the amount of talker-shape exposure that listeners had experienced when they reached the test trials is roughly the same as in Experiment 2's final block of training. This means that listeners in the current experiment showed stronger early talker-specificity effects in their visual fixations than did listeners in the final training blocks of Experiments 1 and 2, with an equivalent amount of exposure to word–talker pairings.

Discussion

This experiment demonstrated that listeners do not find it prohibitively difficult to associate shapes with one of two talkers, and that they can use talker-related acoustic information very early in the spoken utterance to constrain looks to one picture. This learning is robust to changes in the exact sound that is used during learning. This suggests that talker-semantic information can readily be encoded, even for extremely arbitrary referents, at least when learners are directed toward talker-referent contingencies. Relating this to Experiment 2, where talker-semantic effects were small, the current experiment suggests that participants in Experiment 2 *should* have had plenty of time to encode robust talker-referent pairings by the third block of training trials, yet they did not. It seems clear that the listeners in Experiment 2 were not doing the same thing as listeners in Experiment 3—that they were not strongly encoding—or at least were not utilizing—these talker-shape relationships. Encoding in the current experiment may have been aided by the information that the two talkers had affinities for particular shapes, a talker-semantic relationship that listeners may have previously experienced in everyday life. Experiments 1–2 did not suggest such a relationship, meaning that listeners had no reason to think that word–talker probabilities would differ.

Thus, the difficulty of mapping novel shapes to talkers cannot explain weak talker-semantic encoding in Experiment 2. What differed between the current experiment and Experiment 2? Three factors stand out: encoding novel word forms; attention to talker identity; and trial-by-trial diagnostic value. Participants in Experiment 2 may not have encoded talker–picture contingencies because they were occupied with encoding novel word forms, which diverted them from attending to other contingencies. This would be consistent with Horton and Gerrig's (2005) work

suggesting that encoding talker-referent relationships is dependent on basic memory constraints. Another reason was that Experiment 2's listeners had no reason to think that talker identity was relevant to word learning. By contrast, in the current experiment, listeners *had* to register the talker's identity to select a response, giving them every reason to think talker identity was relevant. Finally, for listeners in Experiment 2, at a trial-by-trial level, talker identity was only diagnostic half the time, because half of the trials were same-talker trials. That is, only 50% of trials could be answered using only talker information. In the current experiment, talker identity was diagnostic all of the time (100%).

To explore the factors that led to encoding of talker-shape pairings, we conducted a final experiment. Experiment 4 duplicated the pre-test portion of Experiment 2, except that real words and real shapes were used. This eliminated the memory load of learning novel phoneme sequences, learning novel shapes, and learning linkages between word forms and shapes, leaving only talker-shape linkages as the new information. It differed from Experiment 3 in that attention to talker was not required to complete the task.

There are two straightforward predictions for Experiment 4. If encoding talker-shape linkages occurs readily when no other new information is present (novel words), then listeners throughout Experiment 4 should readily encode talker-semantic information. No attention to talker was required to complete the task, so if explicitly drawing attention to talker identity is the crucial factor in eliciting strong talker-semantic effects, then we should only see weak talker-semantic effects throughout Experiment 4.

Two more subtle hypotheses are possible. First, if encoding talker-shape mappings is the "bottleneck" in using talker-semantic information, then listeners in Experiment 4 should show talker-semantic effects only when this memory load is lessened. To test this account, we contrasted a high-load association condition (4a: two talkers were linked to eight pictures each) with a low-load association condition (4b and 4c: two talkers were linked to *two* pictures each). If talker-semantic information is more readily used when it is easier to encode talker-shape linkages in memory, then listeners in 4b and 4c should show more evidence of talker-semantic encoding than listeners in 4a. Finally, we also contrasted 50% diagnosticity (4a and 4b) with 100% diagnosticity (4c). If diagnostic value made listeners more likely to utilize picture–talker mappings, then listeners in 4c should show stronger talker-semantic activation (i.e., looks to pictures before word onset) than listeners in 4a and 4b.

The three sub-experiments (4a, 4b, 4c) were run sequentially as we explored these factors. Note that sequential execution creates a confound of time during a single academic quarter, and recent work suggests that time-of-quarter impacts participant conscientiousness (Witt, Donnellan, & Orlando, 2011). It is not transparent to know exactly how a decrease in participant conscientiousness (or attentiveness) would affect the talker-specificity effect, but a reasonable guess is that the talker-specificity effect would be less and less likely as the quarter proceeded. This is the opposite of what was found—talker specificity was evident only in the *last* condition, which was run at the end of the academic quarter. This suggests that, if anything, the confound of time makes our tests more conservative than if conditions had been run simultaneously.

Experiment 4

Method

Participants

In each of three sub-experiments, 16 participants (total $N = 48$) from the same pool as in previous experiments took part. Two had participated in Experiment 2 a year or more prior to the current experiment (one in 4a, one in 4b). Neither participant patterned differently from the other participants in their sub-experiment, and analyses were unchanged with these participants removed. Analyses reported include these participants.

Materials

Visual stimuli were 16 Microsoft Office shapes (circle, square, triangle, star, donut, cloud, lightning bolt, stop sign, smiley face, teardrop, sun, moon, hourglass, pacman, diamond, heart) judged to be familiar to undergraduates. Each shape had black fill rather than color, to maintain a similar level of perceptual distinctiveness relative to the black-and-white shapes in the first three experiments. Shapes were edited to fill a 200×200 pixel square. Because the original talkers were unavailable, two new talkers—one male, one female—provided recordings. Recordings took place in a sound-attenuated chamber, and consisted of each talker saying "Click on the X," where X was one of the 16 shapes specified above. Sound files were recorded to a computer, and were then edited to select artifact-free tokens and to set the average intensity to 70 dB SPL for each sound file.

Procedure

This experiment was designed to be comparable to Experiment 2's final block of training trials. In that trial block of that experiment, we found no talker-specificity effect during the sentence carrier. Since Experiment 2's listeners took 2.94 blocks to reach criterion accuracy, the final block of training trials in Experiment 2 occurred after (on average) 1.94 blocks of exposure—that is, it was roughly the third block of trials. There were 8 instances of each shape-talker pairing per training block (two shapes appearing and one of them being named). This means that listeners during that final training block had heard each word about 16 times ($8 \text{ instances} * 1.94 \text{ blocks} = 15.52$ exposures).

The current experiment followed this pattern, but with real words. Like Experiment 2, words were presented in the frame "Click on the X." Also like Experiment 2, listeners were provided 16 exposures to each shape-talker pairing prior to the trials where eye movements were measured. Experiment 4a, with 16 pictures, matched Experiment 2 in number of trials. Experiments 4b–4c, which had only

four pictures each, were matched on the number of per-picture exposures but not on the total number of trials.

Each target was presented eight times per block, as in Experiment 2. This meant that, for Experiment 4a, with 16 possible target pictures, each block contained 128 trials, while for Experiments 4b–4c (four possible target pictures), each block contained 32 trials. Each participant completed three blocks of trials, but only the third block of trials was analyzed (matching Experiment 2's final training block). This was followed by a questionnaire that assessed memory for word–talker pairings and awareness of the experimental manipulation.

Each target occurred equally often with two other possible pictures. For Experiments 4a and 4b, half of these picture pairings were different-talker picture pairings. Which pictures appeared together and whether a picture pair was a same- or different-talker pair were counterbalanced across participants. For Experiment 4c, all of the picture pairings were different-talker picture pairings to make talker identity diagnostic of the correct answer on 100% of trials. Note also that, rather than being early-overlapping pairs, the shapes used here were all phonologically distinct (early overlap was no more than the first phoneme). In each case, we looked for talker-specific fixation patterns prior to word onset, as an indication of talker-semantic encoding (linking the talker identity to the shape).

Results

Experiment 4a

Accuracy was high ($M = 99.6\%$, $SD = .3\%$). Listeners showed subtle evidence of fixating talker-specific pictures prior to target word onset (200 ms before to 200 ms after) in the third block of trials (Fig. 7). There was a marginal effect of talker on target advantage ($t_1(15) = 1.81$, $p = .09$; $t_2(15) = 1.01$, $p = .33$; mean difference = .026). Thus, removing the requirement to learn novel shapes and words does not create a large talker-semantic effect in the absence of explicit instruction. Yet, perhaps 16 talker-word pairings is too many to encode easily, especially given the arbitrary nature of the picture sets. The next sub-experiment presented only four mappings.

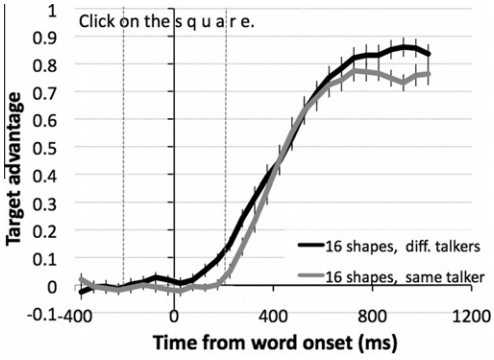

Fig. 7. Experiment 4a, target advantage on same-talker and different-talker trials.

Experiment 4b

Accuracy was high ($M = 99.2\%$, $SD = 1.2\%$). The number of pictures was reduced by a factor of 4 so that the number of picture–talker pairings was 4 (2 per talker). Still, listeners did not show large effects of talker-specificity in the third block of trials (Fig. 8); the visible difference between different-talker and same-talker trials was not significant ($t_1(15) = .41$, $p = .69$; $t_2(15) = .51$, $p = .62$; mean difference = .026), though the effect was numerically identical to Experiment 4a. A reduced number of trials (and thus reduced power) relative to Experiment 4a may have led to failure to detect a subtle effect. This result does not suggest that lightening the memory load in terms of talker–picture mappings generates a strong, early talker-specificity effect.

Experiment 4c

A remaining possibility was that talker information in 4a and 4b did not have enough immediate diagnostic value for participants to exploit it. To address this, Experiment 4c made every single trial a different-talker trial—talker was a diagnostic cue on 100% of trials. Accuracy was high ($M = 99.5\%$, $SD = .9\%$). As shown in Fig. 9, listeners in the third block of trials made above-chance looks to the pictures in the preword time window ($t_1(15) = 4.24$, $p < .0001$; $t_2(15) = 7.53$, $p < .0001$; $M = .173$). In an ANOVA

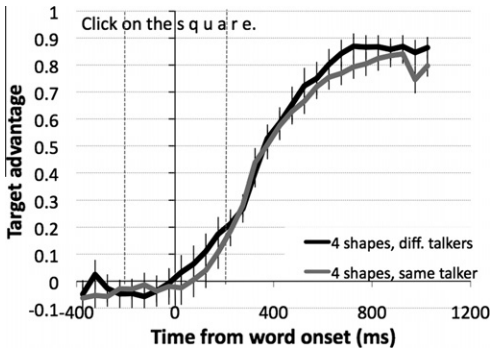

Fig. 8. Experiment 4b, target advantage on same-talker and different-talker trials.

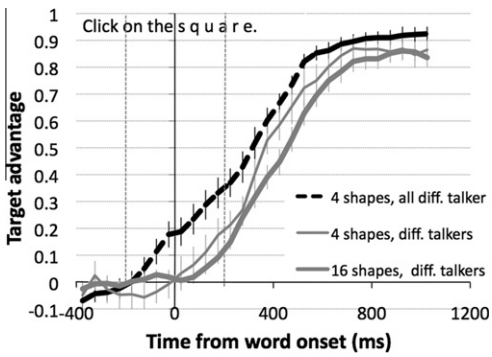

Fig. 9. Experiment 4c, target advantage. All trials were different-talker trials. Different-talker trials from Experiments 4a and 4b provided for comparison (gray). Significance values reflect that black line is greater than each of the gray lines to at least the noted significance level.

on different-talker target advantage with Sub-Experiment (4a, 4b, 4c) as a between-participants factor, Sub-Experiment was significant ($F_1(2, 45) = 4.64$, $p < .02$; $F_2(2, 30) = 12.03$, $p = .0001$, $minF'(2, 71) = 3.35$, $p < .05$, $\eta_G^2 = .17$). The 4c different-talker effect exceeded that for both 4a ($t_1(30) = 3.29$, $p = .003$; $t_2(15) = 4.21$, $p < .001$; .173 vs. .029) and 4b ($t_1(30) = 2.21$, $p = .04$; $t_2(15) = 3.77$, $p = .002$; .173 vs. .029). Examination of the number of participants who showed positive target advantage (Fig. 10) suggests that only Experiment 4c showed a preponderance of participants with positive target advantage.

Questionnaire

Questionnaire results (summarized in Table 4) suggested that listeners had relatively limited explicit awareness of talker-shape pairings in all three sub-experiments. The number of shape-talker pairs recalled was far from complete; listeners sometimes responded "yes" when asked if some of the words had changed voices during the experiment; and few guessed the purpose of the experiment. Eliminating participants who showed awareness of the experimental purpose (3 in 4a, 5 in 4b, 4 in 4c) did not change patterns of significance. Further, correct talker-shape recall did not correlate strongly or significantly with target advantage scores on different-talker trials (Table 4, last column), mirroring Horton's (2007) finding that explicit memory for partner-picture mappings is not strongly related to partner-specific processing benefits in picture naming.

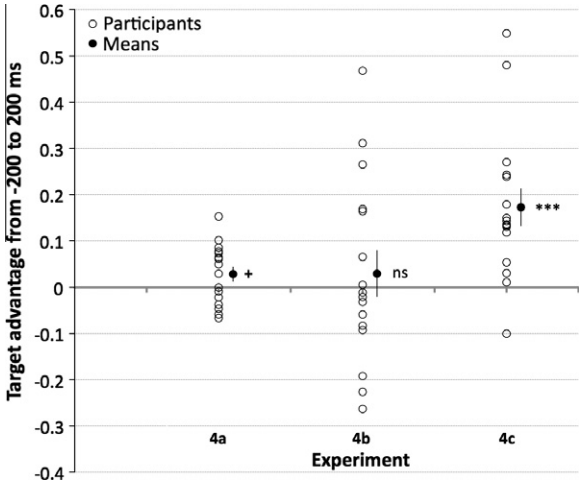

Fig. 10. Experiment 4, target advantage in different-talker condition of each sub-experiment, in the time window −200 to +200 ms from word onset. $^{+}p < .10$; $^{***}p < .001$.

Discussion

The data in Experiment 4 suggest that listeners need high cue diagnosticity to spontaneously utilize talker-specific information in neutral sentence frames, but that awareness of talker–picture mappings need not play a large role. Specifically, listeners only showed strong talker-specific looking patterns during the lead-in sentence ("Click on the") when talker information diagnosed the target on 100% of trials. When talker was only diagnostic on 50% of trials (and uninformative on the rest), listeners showed relatively weak evidence of talker-specific looking patterns, even if a given participant saw only four shapes, which should have provided a relatively low memory load. When listeners did utilize talker information to restrict reference prior to word onset (Experiment 4c), most seemed unaware that they had done so. This supports the notion that talker identity information is something that listeners do not explicitly encode (as demonstrated by Vitevitch, 2003), unless directed to attend to it (as in the current Experiment 3). This talker-semantic effect without awareness echoes Horton's (2007) result that listeners' memory for partner-picture associations is largely unrelated to awareness, and Vitevitch's finding that listeners did not register a mid-experiment change in talker identity.

How do the present results relate to the previous experiments? First, these data provide a potential explanation for Experiment 2's weak talker-semantic effects. Talker-semantic encoding was weak at least in part because talker–picture pairings were not highly diagnostic: listeners *can* implicitly encode talker–picture relationships to predict upcoming material, but talker must be a highly diagnostic cue. Second, the results of Experiment 4c are similar to those in Experiment 3, but without requiring listeners to focus on talker identity to perform the task. Given a situation where talker identity was highly diagnostic, listeners utilized talker–picture mappings automatically. This suggests that directing listeners' attention to talker identity may be sufficient to invoke talker-semantic encoding, but is not necessary.

General discussion

In a series of four experiments, we explored how listeners encode and utilize talker-specific acoustic information

Table 4
Questionnaire data from Experiment 4.

Experiment	Recall word's talker[a]	False recall[a]	Thought word changed voice[b]	Guessed purpose[b]	Shape-looks correlations[c]
4a	25% (26%)	2% (5%)	8/16	3/16	−.04
4b	55% (39%)	5% (14%)	6/16	5/16	−.32
4c	63% (34%)	3% (13%)	6/16	4/16	.18

[a] Means (SDs).
[b] Number of participants.
[c] Target advantage in −200 ms to +200 ms time window correlated with proportion of shapes recalled (2nd column from left). None approached significance.

in on-line spoken language processing. We were interested in both listeners' use of *acoustic-match* information, where word forms are stored in acute acoustic detail, and their use of *talker-semantic* information—linkages between talkers and the things they talked about. We found evidence consistent with use of both types of information. Experiment 1 replicated a study by Creel et al. (2008) which demonstrated that listeners can tell apart newly learned early-phonological-overlap word pairs if the two words are repeatedly spoken by two different talkers. Specifically, listeners distinguished newly-learned words such as *boog* and *booj* faster when the two had been learned from two different talkers (female *boog*, male *booj*) than when both were learned from the same talker (female *boog*, female *booj*). In Experiment 2, we taught the listeners the same novel vocabulary embedded in a lead-in sentence, such as *Click on the boog*. This sentence could give listeners talker-identity information well in advance of the word, yet listeners showed little evidence of using the lead-in sentences, with talker-specificity effects evident only on the words themselves.

Experiments 3 and 4 explored this weak use of talker-semantic information (talker-shape associations) by eliminating various memory-intensive aspects of the task. For these two experiments, listeners' exposure to talker-shape pairings was matched to the exposure level prior to the last block of training in Experiment 2, in which talker-specificity effects were relatively weak. Experiment 3 asked whether participants have difficulty linking particular talkers to arbitrary sets of unfamiliar shapes, without the additional memory load of encoding novel phoneme sequences. Participants learned talker-shape pairings faster than previous participants had learned label-shape pairings. They were able to use talker-identifying information in the sentence almost immediately in order to visually fixate the talker's preferred shape on a trial. This suggested that encoding novel shapes and making arbitrary talker-shape associations is not particularly difficult, but hinted that attention or task demands may play a role. Experiment 4 explored whether listeners used early talker-identifying information in lead-in sentences when the task did not require them to do so. When listeners had to select one of two shapes on a trial, with half talker-diagnostic trials and half non-diagnostic trials, they only weakly used talker identity as a cue, regardless of the number of shape pairings they had to encode. However, when talker was diagnostic on every single trial, and the participant only saw four pictures in the entire experiment, listeners used talker identity cues in the lead-in sentence to fixate a given picture prior to word onset. This suggests that participants in Experiment 2 may not have deciphered the utility of talker-shape pairings as cues because talker information only diagnosed the correct response on a portion of trials. Thus, those participants did not use talker-shape associations to predict responses. In summary, our data suggest that listeners may store acoustic-match information with little effort, but store talker-semantic information when it is attended or is highly diagnostic of a response, and perhaps when memory demands are at a minimum.

The current study is consistent with earlier reports of talker-specific effects in language processing. The first two experiments (as well as Creel et al., 2008, Experiment 2; Goldinger, 1996, 1998) may largely reflect acoustic matching to highly-detailed word-form representations. On the other hand, Experiments 3 and 4c verify that listeners can utilize indexical information in the speech signal quite dramatically, to semantically activate talker-related shapes in the earliest moments of the speech signal. These latter results accord with work by Horton and colleagues (Horton & Gerrig, 2002, 2005) on partner-specific memory for collaborative terms. The four experiments reported here, taken together with existing work (Goldinger, 1996, 1998; Palmeri et al., 1993), suggest that listeners are highly capable of encoding and utilizing talker information both semantically and acoustically, but the circumstances in which listeners use this information are potentially somewhat different. That is, acoustics may be encoded automatically (either with the word form or as an acoustic talker-specific property), while talker-semantic encoding may be sensitive to task utility. The current work does not demonstrate that acoustic encoding is *un*affected by diagnostic value (though see Palmeri et al. (1993), for evidence of automatic encoding of talker-specific word representations). However, it does demonstrate that listeners encode (or utilize) what seems to be talker-semantic information very differently depending on the task they are performing. This may be as simple as changing listeners' *a priori* expectations about the relevance of talker identity, as in Experiment 3. There also seems to be a role for the diagnostic value of indexical information, as exemplified by Experiment 4—high diagnostic value results in greater encoding (or utilization) of talker-shape associations, with little evidence of awareness. Thus, when the listener expects talker identity to be useful, either because it is necessary to succeed in the task (Experiment 3) or because it is an extremely-reliable response cue (Experiment 4), the listener will use talker-semantic information in on-line processing.

It should be kept in mind that the "task utility" of talker-semantic information in the real world may change from situation to situation, and may be greater than that demonstrated in our study. For instance, if a friend says "Where are my shoes?" at the counter of a bowling alley, that person's identity is likely to be highly useful in finding a visual target (shoes of a particular size and color) because the visual array (large numbers of pairs of shoes) will be vastly larger than two pictures.

Multiple roles of acoustic detail in speech processing

At a broader level, the current study suggests that the acoustic details of spoken language have effects at multiple levels in comprehension, which vary from situation to situation. This provides valuable input to our understanding of how listeners use highly-detailed information in the speech signal by demonstrating that talker-specific encoding may occur at multiple loci, but potentially under different circumstances.

First, talker information may be encoded as part of a word's representation. This seems to be the case for young children (Houston & Jusczyk, 2000; Schmale & Seidl, 2009; Singh et al., 2004) and for second-language learners

(Bradlow et al., 1996; Lively et al., 1993). For native adults, talker-specific word encoding might persist, or it might reflect an updating of models of particular talkers' speech habits, or both (e.g., Jesse et al., 2007).

One might reasonably ask why this acoustically-specific storage would be advantageous to the listener, and why semantic encoding would not be *more* advantageous. One possibility is that talker-related acoustic properties are necessary for understanding some speech sounds (e.g., Hillenbrand, Getty, Clark, & Wheeler, 1995; Peterson & Barney, 1952). That is, phoneme identification is somewhat dependent on talker identification. For instance, vowels across different talkers, even within a dialect, are not completely determinable from formant frequencies alone, and equating vowels across talkers is a non-trivial matter (Clopper, 2009). Thus, incorporating talker-specific attributes may be necessary for accurate vowel perception. In support of a role for talker identity in vowel identification, vowels (as well as whole words) are identified much more slowly when talker characteristics change from trial to trial than when they do not (Magnuson & Nusbaum, 2007; Mullennix & Pisoni, 1990; Nusbaum & Morin, 1992). This suggests that every time the talker changes, the listener must change contextual expectations to fit the new talker, meaning that talker-related acoustic characteristics are an important factor in identifying speech sounds. Some consonants also vary demonstrably from talker to talker (Allen, Miller, & DeSteno, 2003; Newman, Clouse, & Burnham, 2001), or do not show transfer of adaptation when talker is changed, suggesting that talker-dependent speech sound perception is not limited to vowels (see Eisner & McQueen, 2005; Kraljic & Samuel, 2005).

With this in mind, we examined individual items in the word-learning experiments, but this did not suggest any consistency in terms of particular vowels, vowel lengths, or differences in vowel lengths, that might conceivably underlie the acoustic talker-specificity effects seen here. It is worth noting that pilot experimentation in our lab suggested that novel words with unvoiced coda consonants (and thus brief vowels) showed weaker talker-specificity effects than the voiced-coda novel words used here. This would be consistent with vowel-based information carrying the talker-specificity effects. Of course, sine-wave speech experiments (Fellowes et al., 1997; Remez et al., 1997) as well as work on talker identification by Van Lancker, Kreiman, and Emmorey (1985), van Lancker, Kreiman, and Wickens (1985) suggest that a large number of acoustic properties differ between, and serve to identify, talkers. The current study was not designed to explore what specific acoustic cues were at work. However, more targeted future work will examine which talker-specific properties may be encoded, and how such properties may generalize to other productions by the same or similar talkers.

Complementing this acoustically-detailed storage is an ability to use acoustic material to access *semantic* (indexical) information about a talker. This information can link directly to referents, rather than being mediated by word form. We found that a semantic-like talker effect was maintained across a change in wording ("Which one do you think I like" to "Can you help me find things to play with") in Experiment 3, and emerged without directing attention to talker in Experiment 4c. This use of talker-identifying information is not even tied to acoustics in particular—acoustic cues are simply one source of information about who is talking. Knowing who is talking, or at least something about them, can then constrain the comprehender's inferences of what the talker may talk about. One reason that talker-semantic information may not be used in on-line comprehension is that listeners may start out with low *a priori* expectations about word–talker associations: such associations are unlikely most of the time, but are likely in certain semantic situations, such as when a talker has a particular interest in a particular concept. This seems to occur in Van Berkum et al. (2008), where indexical cues constrain how the listener thinks a sentence will continue. It also seems to occur in the present Experiment 3. Interestingly, recent work by Staum Casasanto (2008) suggests that visual cues to a talker's ethnicity change the listener's expectations of how the talker will phonetically realize particular words. Thus, activated knowledge about the person talking can adjust listeners' expectations of form as well as meaning.

Further, semantic use of talker-identity cues relates to a literature in referential communication (Brennan & Hanna, 2009; Horton & Gerrig, 2002, 2005; Horton & Keysar, 1996). In particular, Horton and Gerrig (2005) find that individuals in a referential communication task remember associations between conversational partners and shapes, and are better at remembering these associations when they are easily encoded (e.g., one partner knows about fish, while another knows about frogs) than when they are less-easily encoded (each partner knows some fish and some frogs). This difference in memorability may partially explain our participants' apparent failure to encode talker-semantic information in Experiment 2—the set of novel referent shapes was not easily partitioned into two disjoint sets. For that matter, neither were the shapes in Experiment 4. If the two sets of shapes had differed in an easily-identifiable way, such as color differences or membership in two non-overlapping categories, listeners may have more readily encoded talker-related information about the shapes (Galati & Brennan, 2010).

Of course, listeners in Experiment 3 learned our arbitrary talker-shape mappings fairly quickly. This may have been easier in the absence of encoding word forms. Further, it is not particularly surprising that adult humans can encode something that they are asked to pay attention to, and it seems likely that if listeners in Experiment 2 had known that they *should* encode talker information, they may have done so. What is more interesting is the outcome of Experiment 4c, which suggests that listeners readily utilize talker-semantic information *without* explicit instruction to do so. It remains possible that listeners would still be unlikely to encode talker-semantic information under a stronger memory load, even with higher diagnostic value as in Experiment 4c.

Alternative interpretations

We contend that our data reflect two different levels at which talker information is encoded, though we have not definitively proved that this is the case. A different

interpretation of our data is that the differences in the talker-specificity effects we found in the first two experiments vs. the last two were not differences in levels of representation, but a split between implicit and explicit memory. That is, effects of talker-specific representations showed up almost instantaneously during sound presentation in Experiments 3 and 4c because listeners were explicitly aware of the usefulness of talker information as a cue, but showed up more gradually during sound presentation in Experiments 1 and 2 because listeners were only implicitly sensitive to talker information as a cue. This explanation is somewhat compatible with our data. We do suspect that the word-form-representation effects are largely implicit. However, this explanation falters in that listeners in Experiment 4c showed strong talker-specificity effects, but had no more explicit knowledge of the experiment's purpose than listeners in 4a and 4b, who showed only weak talker-specificity effects. As far as our data suggest, listeners are not strongly aware of talker as a semantic predictor except in cases where they are obliged to use it (Experiment 3). This corroborates Horton's (2007) recent work suggesting that listeners encode talker-referent linkages regardless of explicit awareness. Thus, the implicit/explicit distinction may be orthogonal to our levels-of-representation account.

Another interesting account, and one that is supported by recent work (Magnuson & Nusbaum, 2007; Rensink, O'Regan, & Clark, 1997; Vitevitch, 2003), is that the relevant distinction is attentive/inattentive. Magnuson and Nusbaum, for instance, found that the *expectation* of talker changes is enough to elicit interference in a phoneme monitoring task—listeners alerted to the possibility that they were hearing words from two (synthesized) talkers showed interference, while those who were not so alerted did not show interference. That is, listeners may not attend to talker identity unless they are directed to do so. We have argued that listeners engaged in a word learning task in which talkers and words covary do not attend to talker identity. Listeners trying to identify talkers (control condition, Experiment 2) or learn things about that talker (preferred shapes; Experiment 3) had to attend to talker acoustics. Identifying a talker and using her as a predictor may happen predominantly under attentive circumstances. However, this does not explain Experiment 4c, where listeners also showed early patterns of talker-specific looking despite limited awareness of talker–picture relationships. It may be that listeners encode talker-semantic information (talker–picture relationships) incidentally when talker information is highly diagnostic, but attention is required for encoding (or using) talker information in more complex situations. It is also possible that, with a very small set of items *and* high diagnosticity, listeners in Experiment 4c *acoustically* encoded the full sentences, rather than associating talkers with particular shapes.

Implications for language representation

Our data are somewhat consistent with a two-systems account for storage of talker information (e.g., Belin et al., 2000, 2004; González & McLennan, 2007). Listeners may store talker-semantic information separately from acoustic-match information. However, proponents for two systems for speech vs. talker information usually mean two different *acoustic* time scales rather than acoustic vs. semantic, perhaps one at a fine temporal grain and left-lateralized, and the other at a coarse temporal grain and right-lateralized (Poeppel, 2003; Zatorre, Belin, & Penhune, 2002). It is convenient to dichotomize variability in the speech signal into rapidly-varying phonemic characteristics (voice onset time, formant transitions) vs. slow-varying talker-related characteristics (*f*0, formant frequency range), which would neatly lateralize speech leftward and talker rightward. Extending this to our proposed acoustic/semantic distinction, identification of talkers might proceed from coarse-grained analyses, and identification of words might proceed from fine-grained analyses.

Complicating this tidy picture, talker variability crosscuts both coarse and fine acoustic time scales, and listeners can use both coarse and fine-grained information to identify talkers (Allen et al., 2003; Fellowes et al., 1997; Remez et al., 1997). This suggests that talker representations and word representations, even if they function separately, must share some information. This might mean that talker representations, while including some fine-grained information, are *biased* toward coarse-grained information more strongly than fine-grained information, with the reverse being true for speech processing. This bias might then lead to apparent rightward lateralization for voice identification. The neuropsychological literature is somewhat ambivalent on this count. Brain imaging (Belin et al., 2000) and lesion studies (see Belin et al., 2004, and references within) suggest that voice identification is right-lateralized. However, Perrachione, Pierrehumbert, and Wong (2009) find evidence more consistent with leftward lateralization for voice identification. Further, Saygin, Dick, Wilson, Dronkers, and Bates (2003) find evidence that even nonspeech environmental sound perception is impaired in aphasic patients (with left-hemisphere damage), suggesting that sound identification generally is left-lateralized. Without controlling for the temporal "grain" of the information used in identification, though, it is difficult to know whether lateralization patterns result from differences in stimulus class (speech sounds vs. voices), acoustic properties (fine-grained or coarse-grained), or both.

Our data do not speak directly to the idea of differential cue weighting in talker recognition ("coarse" bias) vs. word recognition ("fine" bias), in that it is unknown what acoustic parameters facilitated talker recognition in the current study. In fact, different cues may have facilitated talker recognition in some of our experiments, and talker-specific word recognition in others.[2] However, we can make some tentative assertions about the use of coarse-grained information in talker identification. Poeppel (2003) suggests that the temporal integration window for "coarse" (right-hemisphere) processing is roughly 200 ms. Taking this as a basis, optimal recognition of talkers from coarse-grained

[2] Thanks to J. McQueen for noting this possibility.

properties should occur at a delay of 200 ms from sound onset, plus the time to plan an eye movement (another 200 ms). That is, reliable recognition of talkers should emerge at around 400 ms.[3] However, in talker recognition conditions in the current study (Experiment 3, and the control condition of Experiment 2), listeners distinguished the two talkers (i.e., showed a positive target advantage) by 250–300 ms (Experiment 2: $t_1(7) = 2.68$, $p = .03$; $t_2(15) = 3.85$, $p = .002$; Experiment 3: $t_1(15) = 3.62$, $p = .003$; $t_2(15) = 2.23$, $p = .04$). By comparison, phoneme differences without talker differences—here, the same-talker trials in Experiment 4a–4b—showed a positive target advantage only 50 ms earlier, from 200–250 ms ($t_1(31) = 3.53$, $p = .001$; 250–300 ms by items, $t_2(15) = 4.17$, $p = .0008$). Thus, the speed of our talker recognition effects seem somewhat inconsistent with listeners using *exclusively* coarse-grained information to identify talkers. However, since our study was not designed with this particular question in mind, and only tests a two-talker distinction, further investigation is needed.

What do these potential dichotomies in sound processing mean for accessing information about talkers during on-line language comprehension? One possibility is that talker representations consist of acoustic information encoded as some type of reference frame or predictive model, which describes how a talker routinely produces speech sounds (see Cutler et al., 2010, for a suggestion along these lines). This could then be used alone to identify talkers, or combined with abstracted word representations to produce talker-specific expectations for words. Further, it seems from the current study (particularly Experiment 4) that such talker representations may be accessed without explicit awareness. What, then, allows listeners to distinguish similar-sounding words in the first two experiments? One possibility is that listeners are encoding words and implicitly encoding talker identity, and these two representations are being linked due to their coactivation. On the other hand, since there is overlap in the acoustic properties that identify both talkers and speech sounds (because talkers realize speech sounds differently), it may be impossible for word encoding and word recognition to be unaffected by talker variability, even if the listener is not encoding talker identity.

Implications for language learners

The current study, taken together with previous work (Creel et al., 2008; Goldinger, 1996, 1998; Palmeri et al., 1993), has implications for language learning in that it suggests that native adult listeners represent information about talkers at both acoustic and semantic levels of language representation. That is, acoustically-specific representations of word forms are encoded even when the listener is inattentive to talker identity. They may not be able to disattend to talker-related acoustic features, because at least some of the those acoustic features are crucial to perceiving speech. On the other hand, adults do apparently fail to attend to talker *identity* in speech processing (Vitevitch, 2003), much like they may fail to attend to the visual specifics of an interlocutor (Simons & Levin, 1998). Vitevitch, following Rensink, O'Regan, and Clark (1997), suggests that listeners may only encode talker identity under limited circumstances, and that this may be modulated by attention. We modify this account by suggesting that extremely high diagnostic value, as well as awareness, may increase encoding of talker-semantic mapping.

This raises the question of what information young language learners are likely to take in. For children, acoustic matching might be a stronger influence than for adults, given that children have more recently learned—or are still learning—to focus attention on phonemic aspects of the speech signal. Children do seem to learn to hear through talker variability (Houston & Jusczyk, 2000) during the first year of life, on the same developmental time scale as the loss of nonnative contrasts (Werker & Tees, 2002). Nonetheless, previous work demonstrates that, like adults, preschool-aged children are hindered by talker variability in a word recognition task (Jerger, Martin, & Pirozzolo, 1988; Ryalls & Pisoni, 1997). Unlike adults, children match acoustic characteristics of the talkers in reproducing recognized words (Ryalls & Pisoni, though see Goldinger, 1998, for such effects in adults). This might mean that children are somewhat more sensitive than adults to acoustic-match effects in word learning, with a gradual lessening—but not disappearance—of acoustic-match effects stretching across the course of development, rather than just the first year of life.

Conversely, young children might encode talker-semantic information differently than adults, in one of two ways. First, they might encode talker-semantic information less strongly than adults. This might happen because children have less information about associations between types of voices and personal characteristics than adults do. Children do respond more positively to familiar-sounding than to unfamiliar-sounding voices (Kinzler, Dupoux, & Spelke, 2007). Current work in our lab (Creel, 2010) further suggests that children can use existing and new associations with particular voice types to predict what colored shapes the talker might request. This fits with an account in which listeners easily encode simple mappings between talkers and referents (Galati & Brennan, 2010). Presumably, adults would be able to keep track of more complex linkages than children, though this has not yet been explored. Second, they might encode talker-semantic information more strongly than adults. This might happen because children, who know fewer words (and fewer talkers) than adults, have higher expectations of the chance probability of a particular word coming from a particular talker. That is, their estimates of chance likelihood and of talker-specific-word likelihood are much closer together, making it easier for them to pick up on talker-word relationships. This remains to be tested.

[3] We acknowledge that this 200-ms window is a sliding window rather than some sort of minimum-information threshold. However, it stands to reason that if listeners really are using coarse-grained properties to identify talkers and fine-grained properties to identify phonemes, then one should find evidence of differences in detection speed in a time-sensitive measure. Postulating *no* differences in a time-sensitive measure would seem to make the "coarse-talker" account unfalsifiable.

Conclusion

The current study explores the circumstances under which listeners use talker-related information to facilitate spoken language processing. We find that task conditions dramatically modulate talker-specific processing, and that with full attention toward identifying talkers, listeners can distinguish two talkers on a similar time scale as two words. When asked to learn words, listeners may not utilize talker-semantic information (Experiment 2), but when asked explicitly to link talkers to novel referents, they readily do so (Experiment 3). Also, when within-trial diagnostic value is low, listeners do not seem to use talker information to predict referents (Experiments 4a and 4b), but do use talker information when diagnostic value is high. We argue that there are at least two processes at work, one which involves storing highly-detailed representations of speech, and another which involves semantically associating talkers with information in the outside world, though further empirical tests would be needed to bolster this position. Overall, the study emphasizes the complexity of listeners' use of talker information in on-line language processing.

Appendix

Analysis of individual phoneme and diphone probabilities was conducted using the Phonotactic Probability Calculator (Vitevitch & Luce, 2004). All probabilities are raw calculations estimated using the online version of the PPC, available at http://www.bncdnet.ku.edu/cgi-bin/DEEC/post_ppc.vi.

References

Abercrombie, D. (1967). *Elements of general phonetics*. Chicago: Aldine Publishing. Co.

Allen, J. S., Miller, J. L., & DeSteno, D. (2003). Individual talker differences in voice-onset-time. *Journal of the Acoustical Society of America, 113*, 544–552.

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language, 38*, 419–439.

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247–264.

Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal attributions. *Journal of Personality and Social Psychology, 37*, 715–727.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*, 379–384.

Barker, B., & Newman, R. S. (2004). Listen to your mother! The role of talker familiarity in infant streaming. *Cognition, 94*, B45–B53.

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Science, 8*, 129–135.

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

Bicknell, K., Elman, J. L., Hare, M., McRae, K., & Kutas, M. (2010). Effects of event knowledge in processing verbal arguments. *Journal of Memory and Language, 63*, 489–505.

Boersma, P., & Weenink, D. (2007). Praat: Doing phonetics by computer (Version 5.1.20) [Computer program]. <http://www.praat.org/>. Retrieved 31.10.09.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. (1996). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech perception. *Journal of the Acoustical Society of America, 101*, 2299–2310.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*, 433–436.

Brennan, S. E., & Hanna, J. E. (2009). Partner-specific adaptation in dialog. *Topics in Cognitive Science, 1*(2), 274–291.

Clopper, C. G. (2009). Computational methods for normalizing acoustic vowel data for talker differences. *Language and Linguistics Compass, 3*, 1430–1442.

Clopper, C. G., & Pisoni, D. B. (2004a). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics, 32*, 111–140.

Clopper, C. G., & Pisoni, D. B. (2004b). Effects of talker variability on perceptual learning of dialects. *Language & Speech, 47*, 207–239.

Cornelissen, F. W., Peters, E., & Palmer, J. (2002). The Eyelink toolbox: Eye tracking with MATLAB and the psychophysics toolbox. *Behavior Research Methods, Instruments & Computers, 34*, 613–617.

Creel, S. C. (2010). Considering the source: Preschoolers (and adults) use talker acoustics predictively and flexibly in on-line sentence processing. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the*

Word	Phonemes					Diphones			
	1st	2nd	3rd	4th	*Total*	1st	2nd	3rd	*Total*
bɛlm	.0512	.0729	.0737	.0295	*.2273*	.0032	.0087	.0010	*.0129*
bɛln	.0512	.0729	.0737	.0467	*.2445*	.0032	.0087	.0004	*.0123*
bug	.0512	.0221	.0179		*.0913*	.0012	.0002		*.0014*
budʒ	.0512	.0221	.0108		*.0841*	.0012	.0001		*.0013*
dɑlm	.0518	.0605	.0737	.0295	*.2155*	.0023	.0059	.0010	*.0092*
dɑln	.0518	.0605	.0737	.0467	*.2328*	.0023	.0059	.0004	*.0086*
dɑɹg	.0518	.0605	.0784	.0137	*.2044*	.0023	.0161	.0010	*.0194*
dɑɹdʒ	.0518	.0605	.0784	.0112	*.2020*	.0023	.0161	.0010	*.0194*
vɔɹm	.0224	.0165	.0784	.0295	*.1468*	.0002	.0030	.0034	*.0065*
vɔɹn	.0224	.0165	.0784	.0467	*.1641*	.0002	.0030	.0027	*.0058*
vig	.0224	.0318	.0179		*.0722*	.0010	.0006		*.0016*
vidʒ	.0224	.0318	.0108		*.0650*	.0010	.0005		*.0015*
zɛlm	.0026	.0729	.0737	.0295	*.1786*	.0004	.0087	.0010	*.0101*
zɛln	.0026	.0729	.0737	.0467	*.1959*	.0004	.0087	.0004	*.0095*
zɝg	.0026	.0247	.0179		*.0452*	.0001	.0005		*.0006*
zɝdʒ	.0026	.0247	.0108		*.0380*	.0001	.0010		*.0010*

32nd annual conference of the cognitive science society (pp. 1810–1815). Austin, TX: Cognitive Science Society.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2006). Acquiring an artificial lexicon: Segment type and order information in early lexical entries. *Journal of Memory and Language, 54*, 1–19.

Creel, S. C., Aslin, R. N., & Tanenhaus, M. K. (2008). Heading the voice of experience: The role of talker variation in lexical access. *Cognition, 108*, 633–664.

Creel, S. C., Tanenhaus, M. T., & Aslin, R. N. (2006). Consequences of lexical stress on learning an artificial lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*, 15–32.

Cutler, A., Eisner, F., McQueen, J. M., & Norris, D. (2010). How abstract phonemic categories are necessary for coping with speaker-related variation. In C. Fougeron, B. Kühnert, M. D'Imperio, & N. Vallée (Eds.), *Laboratory phonology 10* (pp. 91–111). Berlin: de Gruyter.

Dahan, D., & Tanenhaus, M. K. (2004). Continuous mapping from sound to meaning in spoken-language comprehension: Immediate effects of verb-based thematic constraints. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 498–513.

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*, 224–238.

Fellowes, J. M., Remez, R. E., & Rubin, P. E. (1997). Perceiving the sex and identity of a talker without natural vocal timbre. *Perception & Psychophysics, 59*, 839–849.

Galati, A., & Brennan, S. E. (2010). Attenuating information in spoken communication: For the speaker, or for the addressee? *Journal of Memory and Language, 62*, 35–51.

Gaskell, G., & Marslen-Wilson, W. D. (2002). Representation and competition in the perception of spoken words. *Cognitive Psychology, 45*, 220–566.

Geiselman, R. E., & Bellezza, F. S. (1976). Long-term memory for speaker's voice and source location. *Memory & Cognition, 4*, 483–489.

Geiselman, R. E., & Bellezza, F. S. (1977). Incidental retention of speaker's voice. *Memory & Cognition, 5*, 658–665.

Geiselman, R. E., & Crawley, J. M. (1983). Incidental processing of speaker characteristics: Voice as connotative information. *Journal of Verbal Learning and Verbal Behavior, 22*, 15–23.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*, 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes: An episodic theory of lexical access. *Psychological Review, 105*, 251–279.

González, J., & McLennan, C. T. (2007). Hemispheric differences in indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 410–424.

Green, K. P., Tomiak, G. R., & Kuhl, P. K. (1997). The encoding of rate and talker information during phonetic perception. *Perception & Psychophysics, 59*, 675–692.

Hallett, P. E. (1986). Eye movements. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance*. New York: Wiley.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America, 97*, 3099–3111.

Horton, W. S. (2007). The influence of partner-specific memory associations on language production: Evidence from picture naming. *Language and Cognitive Processes, 22*, 1114–1139.

Horton, W. S., & Gerrig, R. J. (2002). Speakers' experiences and audience design: Knowing when and knowing how to adjust utterances to addressees. *Journal of Memory and Language, 47*, 589–606.

Horton, W. S., & Gerrig, R. J. (2005). The impact of memory demands on audience design during language production. *Cognition, 96*, 127–142.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*, 91–117.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 1570–1582.

Houston, D. M., & Jusczyk, P. W. (2003). Infants' long-term memory for the sound patterns of words and voices. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 1143–1154.

Jerger, S., Martin, R., & Pirozzolo, F. (1988). A developmental study of the auditory Stroop effect. *Brain & Language, 35*, 86–104.

Jesse, A., McQueen, J. M., & Page, M. (2007). The locus of talker-specific effects in spoken word recognition. In J. Trouvain & W. J. Barry (Eds.), *Proceedings of the 16th international congress of phonetic sciences* (pp. 1921–1924). Dudweiler: Pirrot.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and Language, 49*, 133–156.

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences, 104*, 12577–12580.

Ko, S. J., Judd, C. M., & Blair, I. V. (2006). What the voice reveals: Within- and between-category stereotyping on the basis of voice. *Personality and Social Psychology Bulletin, 32*, 806–819.

Ko, S. J., Judd, C. M., & Stapel, D. A. (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.

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology, 51*, 141–178.

Krauss, R. M., Freyberg, R., & Morsella, E. (2002). Inferring speakers' physical attributes from their voices. *Journal of Experimental Social Psychology, 38*, 618–625.

Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: U Penn.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*, 98–104.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (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*, 1242–1255.

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

Magnuson, J. S., Tanenhaus, M. K., Aslin, R. N., & Dahan, D. (2003). The time course of spoken word learning and recognition: Studies with artificial lexicons. *Journal of Experimental Psychology: General, 132*, 202–227.

McQueen, J. M., Cutler, A., & Norris, D. (2006). Phonological abstraction in the mental lexicon. *Cognitive Science, 30*, 1113–1126.

Morton, J. B., & Trehub, S. E. (2001). Children's understanding of emotion in speech. *Child Development, 72*, 834–843.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47*, 379–390.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative production. *Journal of the Acoustical Society of America, 109*, 1181–1196.

Nosofsky, R. M. (1989). Further tests of an exemplar-similarity approach to relating identification and categorization. *Perception & Psychophysics, 45*, 279–290.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception, speech production, and linguistic structure*. Washington: IOS Press.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science, 5*, 42–46.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*, 434–447.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*, 309–328.

Peirce, C. S. (1903/1998). A syllabus of certain topics of logic. In Peirce Edition Project (Ed.), *The essential Peirce. Selected philosophical writings (1893–1913)* (Vol. 2, pp. 258–270). Bloomington: Indiana University Press (Original work published 1903).

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision, 10*, 437–442.

Perrachione, T. K., Pierrehumbert, J. B., & Wong, P. C. M. (2009). Differential neural contributions to native- and foreign-language talker identification. *Journal of Experimental Psychology: Human Perception and Performance, 35*, 1950–1960.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*, 175–184.

Pierrehumbert, J. B., Bent, T., Munson, B., Bradlow, A. R., & Bailey, J. M. (2004). The influence of sexual orientation on vowel production (L). *Journal of the Acoustical Society of America, 116*, 1905–1908.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as ''asymmetric sampling in time''. *Speech Communication, 41*, 245–255.

Raaijmakers, J. G. W. (2003). A further look at the ''Language-as-Fixed-Effect Fallacy''. *Canadian Journal of Experimental Psychology, 57*, 141–151.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., & Gremmen, F. (1999). How to deal with "The Language-as-Fixed-Effect Fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language, 41*, 416–426.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science, 9*, 347–356.

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*, 651–666.

Rensink, R. A., O'Regan, J. K., & Clark, J. J. (1997). To see or not to see: The need for attention to perceive changes in scenes. *Psychological Science, 8*, 1–6.

Richtsmeier, P. T., Gerken, L., Goffman, L., & Hogan, T. (2009). Statistical frequency in perception affects children's lexical production. *Cognition, 111*, 372–377.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science, 12*, 339–349.

Ryalls, B. O., & Pisoni, D. B. (1997). The effect of talker variability on word recognition in preschool children. *Developmental Psychology, 33*, 441–452.

Saygin, A. P., Dick, F., Wilson, S. W., Dronkers, N. F., & Bates, E. (2003). Neural resources for processing language and environmental sounds: Evidence from aphasia. *Brain, 126*, 928–945.

Schmale, R., & Seidl, A. (2009). Accommodating variability in voice and foreign accent: Flexibility of early word representations. *Developmental Science, 12*, 583–601.

Simons, D. J., & Levin, D. T. (1998). Failure to detect changes to people in a real-world interaction. *Psychonomic Bulletin & Review, 5*, 644–649.

Singh, L., Morgan, J. L., & White, K. S. (2004). Preference and processing: The role of speech affect in early spoken word recognition. *Journal of Memory and Language, 51*, 173–189.

Sjerps, M. J., & McQueen, J. M. (2010). The bounds on flexibility in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 36*, 195–211.

Staum Casasanto, L. (2008). Does social information influence sentence processing? In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the 30th annual conference of the cognitive science society* (pp. 799–804). Austin, TX: Cognitive Science Society.

Sundara, M. (2005). Acoustic-phonetics of coronal stops: A cross-language study of Canadian English and Canadian French. *The Journal of the Acoustical Society of America, 118*, 1026–1037.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632–1634.

Van Berkum, J. J., van den Brink, D., Tesink, C. M., Kos, M., & Hagoort, P. (2008). The neural integration of speaker and message. *Journal of Cognitive Neuroscience, 20*, 580–591.

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I. Recognition of backward voices. *Journal of Phonetics, 13*, 19–38.

Van Lancker, D., Kreiman, J., & Wickens, T. (1985). Familiar voice recognition: Patterns and parameters. Part II: Recognition of rate-altered voices. *Journal of Phonetics, 13*, 39–52.

Vitevitch, M. S. (2003). Change deafness: The inability to detect changes between two voices. *Journal of Experimental Psychology: Human Perception and Performance, 29*, 333–342.

Vitevitch, M. S., & Luce, P. A. (2004). A web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers, 36*, 481–487.

Watson, D., Tanenhaus, M., & Gunlogson, C. (2008). Interpreting pitch accents in online comprehension: H^* vs. $L + H^*$. *Cognitive Science, 32*, 1232–1244.

Werker, J. F., & Tees, R. (2002). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 25*, 121–133.

Witt, E. A., Donnellan, M. B., & Orlando, M. J. (2011). Timing and selection effects within a psychology subject pool: Personality and sex matter. *Personality and Individual Differences, 50*, 355–359.

Zatorre, R. J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences, 6*, 37–46.