Atten Percept Psychophys (2018) 80:784–797
DOI 10.3758/s13414-017-1395-5

Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing

Ja Young Choi[1,2] · **Elly R. Hu**[1] · **Tyler K. Perrachione**[1]

Published online: 7 February 2018

Abstract The nondeterministic relationship between speech acoustics and abstract phonemic representations imposes a challenge for listeners to maintain perceptual constancy despite the highly variable acoustic realization of speech. Talker normalization facilitates speech processing by reducing the degrees of freedom for mapping between encountered speech and phonemic representations. While this process has been proposed to facilitate the perception of ambiguous speech sounds, it is currently unknown whether talker normalization is affected by the degree of potential ambiguity in acoustic-phonemic mapping. We explored the effects of talker normalization on speech processing in a series of speeded classification paradigms, parametrically manipulating the potential for inconsistent acoustic-phonemic relationships across talkers for both consonants and vowels. Listeners identified words with varying potential acoustic-phonemic ambiguity across talkers (e.g., beet/boat vs. boot/boat) spoken by single or mixed talkers. Auditory categorization of words was always slower when listening to mixed talkers compared to a single talker, even when there was no potential acoustic ambiguity between target sounds. Moreover, the processing cost imposed by mixed talkers was greatest when words had the most potential acoustic-phonemic overlap across talkers. Models of acoustic dissimilarity between target speech sounds did not account for the pattern of results. These results suggest (a) that talker normalization incurs the greatest processing cost when disambiguating highly confusable sounds and (b) that talker normalization appears to be an obligatory component of speech perception, taking place even when the acoustic-phonemic relationships across sounds are unambiguous.

Keywords Speech perception · Categorization

During speech perception, listeners extract stable phonemic percepts from highly variable acoustic signals. In particular, differences among talkers give rise to substantial variation in the acoustic realization of speech, resulting in a nondeterministic relationship between speech acoustics and target phoneme categories for both vowels and consonants (Hillenbrand, Getty, Clark, & Wheeler, 1995; Miller & Baer, 1983; Volaitis & Miller, 1992). A core challenge in understanding speech processing is to determine how the mind and brain disambiguate the many-to-many mapping between acoustics and phonemes. A common account of how listeners maintain phonetic constancy across talkers is talker normalization (Johnson, 2005; Nusbaum & Magnuson, 1997; Pisoni, 1997). In talker normalization, listeners extract information about a talker's vocal tract and articulation from their speech and use this information to establish talker-specific correspondences between idiosyncratic acoustic signals and abstract phonological representations. This talker-specific mapping helps disambiguate a talker's intended phonemes by reducing the degrees of freedom between the acoustic realization of speech and its abstract, categorical representation in the listener's mind.

Previous research has consistently shown that phonetic variability related to talker differences introduces additional processing demands ("interference") in speech perception. Recognizing speech in the presence of indexical variability (i.e., speech from multiple talkers) is slower and less accurate

✉ Tyler K. Perrachione
tkp@bu.edu

[1] Department of Speech, Language, and Hearing Sciences, Boston University, 635 Commonwealth Ave., Boston, MA 02215, USA

[2] Program in Speech and Hearing Bioscience and Technology, Harvard University, Cambridge, MA, USA

compared to speech from a single talker (Assmann, Nearey, & Hogan, 1982; Green, Tomiak, & Kuhl, 1997; Magnuson & Nusbaum, 2007; Morton, Sommers, & Lulich, 2015; Mullennix & Pisoni, 1990; Mullennix, Pisoni, & Martin, 1989; Strange, Verbrugge, Shankweiler, & Edman, 1976). Correspondingly, processing speech produced by variable talkers is more computationally intensive, as reflected in increased neurophysiological and electrophysiological response to talker variability in speech (Chandrasekaran, Chan, & Wong, 2011; Kaganovich, Francis, & Melara, 2006; Perrachione et al., 2016; Wong, Nusbaum, & Small, 2004; Zhang et al., 2013). Variability introduces the possibility of alternative interpretations of the incoming signals, which increases the processing demands on the listener (Magnuson & Nusbaum, 2007).

Two mechanisms for talker normalization have been proposed, each focusing on different types of cues to map speech signals to phonetic categories: intrinsic and extrinsic normalization (e.g., Nearey, 1989). For intrinsic normalization, there may be sufficient ancillary information within speech sounds to self-normalize the relevant phonetic dimensions. For example, in vowels, relationships between fundamental frequency and higher formant frequencies may help disambiguate the phonemically relevant formants (Nearey, 1989; Syrdal & Gopal, 1986). Intrinsic talker normalization allows listeners to understand speech when the source is variable and unpredictable, with the caveat that recognition will be slower and more computationally intensive. On the other hand, extrinsic normalization makes use of phonetic information from preceding speech by the talker to reduce the decision space for identifying target sounds. This mechanism develops talker-specific acoustic-to-phonemic correspondences, facilitating speech perception by building perceptual dependencies between target speech sounds and the preceding speech context (Holt, 2006; Kleinschmidt & Jaeger, 2015; Ladefoged & Broadbent, 1957; Sjerps, McQueen, & Mitterer, 2013; Zhang & Chen, 2016).

Although talker normalization reduces ambiguity in acoustic-to-phonemic mapping (Nusbaum & Magnuson, 1997), it is not known whether the perceptual recalibration involved in intrinsic talker normalization occurs even when acoustic-phonetic features convey a target phonemic contrast unambiguously—contemporary models of perceptual adaptation in speech delineate how developing talker-specific correspondences can improve the efficiency of speech processing, but they are silent as to when these refinements must take place (Kleinschmidt & Jaeger, 2015; Norris, McQueen, & Cutler, 2003). For example, the acoustic features that distinguish sounds such as /o/ and /u/ overlap substantially across the productions of different talkers, whereas those of sounds such as /o/ and /i/ are acoustically nonoverlapping (e.g., Hillenbrand et al., 1995). It follows that phonological contrasts with greater chance for acoustic-phonemic uncertainty across talkers should be more difficult for listeners to disambiguate in the presence of indexical variability, whereas it is unclear whether indexical variability should impose any processing cost on the identification of sounds that are wholly perceptually distinct. Studies of the cognitive cost of indexical variability on speech perception have typically used only a single pair of minimally contrasting sounds as target stimuli (e.g., /ba/ vs. /da/; Green et al., 1997; /b/ vs. /p/; Mullennix & Pisoni, 1990; /s/ vs. /t/; Cutler, Andics, & Fang, 2011), leaving it unclear whether the effect of indexical variability would have different consequences when there is more or less potential for overlap in the target contrasts' acoustic-phonetic features across talkers. It is correspondingly unknown whether talker normalization is an obligatory processing step during speech perception, that is, whether it must occur regardless of the potential ambiguity of encountered speech sound contrasts, or whether it is an operation only brought online when potential phonological ambiguity needs resolving. For example, while the acoustics of one talker's /o/ may be the same as another talker's /u/, in no case would we expect one talker's /o/ to be the same as another talker's /i/. Will there nonetheless be a processing cost associated with talker normalization in this latter case?

In this study, we investigated whether talker normalization processes differentially facilitate speech perception as a function of the level of ambiguity of target sound contrasts. In Experiment 1, we parametrically manipulated the potential for acoustic-phonemic ambiguity across talkers for consonant sounds by varying the number of articulatory (and, thereby, acoustic-phonetic) features they shared. In Experiment 2, we varied the potential formant frequency overlap among vowel categories across talkers. For both experiments, we used a speeded classification task (similar to Garner, 1974) to examine how participants' response times for word identification differed as a function of indexical variability and acoustic similarity between target speech sounds. This speeded classification task allows us to examine the effect of an orthogonal dimension (talker) on processing the target dimension (phoneme). Longer response times in the orthogonal (mixed-talker) condition relative to the control (single-talker) condition indicate that the two dimensions of a speech sound stimulus are processed integrally. Unlike classical Garner paradigms, we do not investigate the reverse effect (i.e., whether variation due to differences in phonemes affect classification of talkers) or congruence effects, as these have been studied extensively elsewhere (e.g., Cutler et al., 2011; Green et al., 1997; Mullennix & Pisoni, 1990) and do not bear on the present research questions. Instead, here we are interested in how and when listeners must incur processing costs to solve the challenge of talker variability during speech perception – a line of inquiry that depends specifically on the difference between conditions with and without orthogonal variability due to differences among talkers.

Using response time data, we tested two primary hypotheses: (a) that talker variability introduces additional cost on speech processing, even when there is no potential acoustic ambiguity in the target phonemic contrast, and (b) that the processing cost of talker variability varies as a function of the overlap between speech acoustics and potential phonemic targets. In two ancillary analyses, we also investigated whether (c) the processing cost associated with talker variability in mixed-talker conditions depends on the target phonetic contrasts' baseline discriminability in the corresponding single-talker condition, and (d) whether the within-category distinctiveness of individual vowel tokens affects their recognition in the presence of indexical variability.

Experiment 1: Consonants

Method

Participants

Native speakers of American English (N = 24; 16 female, eight male; mean age = 20.9 ± 3.27 years) participated in this study. All participants had a self-reported history free from speech, hearing, or language disorder. Participants gave informed, written consent overseen by the Institutional Review Board at Boston University.

Stimuli

Stimuli consisted of four naturally spoken English words that shared the same vowel nucleus (/aɪ/), but which started with different consonants (*buy, sigh, tie, pie*). We chose these words because they allowed us to manipulate the phonetic similarity between target words across three levels (low, medium, and high potential ambiguity across talkers) based on the extent to which they shared phonetic features associated with voicing, manner, and place of articulation. The target contrast in the low-ambiguity condition shared none of these features (/b/ in "buy" vs. /s/ in "sigh"); those in the medium-ambiguity condition shared manner but differed in place and voicing (/b/ in "buy" vs. /t/ in "tie"); and those in the high-ambiguity condition shared manner and place while differing only in voicing (/b/ in "buy" vs. /p/ in "pie") (e.g., Allen, Miller, & DeSteno, 2003). The relative acoustic-phonetic dissimilarity of these contrasts is illustrated in Fig. 1.

The words were recorded by two male and two female native speakers of American English. Recordings were made in a sound-attenuated chamber with a Shure SM58 microphone and Roland Quad Capture sound card sampling at 44.1 kHz and 16 bits. Among numerous tokens of the words recorded by these speakers, the best quality recordings with similar pitch contours and amplitude envelopes were chosen as the final stimulus set. Stimuli were normalized for RMS amplitude to 65 dB SPL in Praat (Boersma, 2001).

Procedure

Participants performed a speeded word identification task in which we parametrically varied the potential ambiguity of the target phonemic contrasts and whether words were spoken by a single talker or mixed talkers (see Fig. 2). Stimuli were presented in six blocks of 40 trials each. Each block consisted of two contrasting words. Participants were instructed to indicate on each trial, as quickly as possible, which of the two words they heard. Each target word was presented 20 times in pseudorandom order, with the restriction that the same word not be presented for more than three sequential trials. In half of the blocks, only one speaker's recordings of the two words were presented (single-talker condition); in the other half, tokens from all four speakers were presented (mixed-talker condition). The talker used in the single-talker condition was counterbalanced across participants. In each block, the vowel nucleus of the two words was kept the same while they differed in their onset consonant (e.g., /baɪ / vs. /saɪ/ in the low-ambiguity condition).

Written instructions assigning a number to the two target words (e.g., "buy = 1; pie = 2" in the high-ambiguity condition) were shown to participants for the duration of each block. Participants listened to the stimuli and identified the spoken word on each trial as quickly and accurately as possible by pressing the corresponding key on a number pad. Trials were presented at a rate of one per 2,000 ms. Stimulus delivery was controlled using PsychoPy Version 1.8.1 (Peirce, 2007). The order of conditions was counterbalanced across participants using Latin square permutations.

Data analysis

Accuracy and response time data were analyzed for each participant in each condition. Accuracy was calculated as the proportion of trials where words were identified correctly out of the total number of trials. Response times were log-transformed to more closely approximate a normal distribution, as expected by the model. Only the response times from correct trials were included in the analysis. Outlier trials deviating from each participant's mean log response time in each condition by more than three standard deviations were also excluded from the analysis (less than 1% of total trials). Data were analyzed in R (Version 3.2.1) using linear mixed-effects models implemented in the package lme4 (Version 1.1.6).

We first assessed whether there was an interference effect of talker variability on the response times for speeded classification of words, and whether this effect varied as a function of the amount of potential intertalker ambiguity in the target consonant contrast. Fixed factors in this analysis included

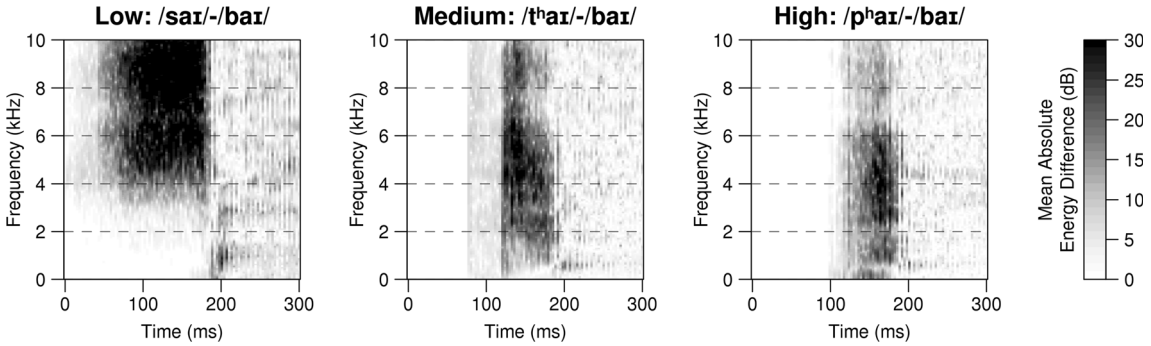

Fig. 1 Potential phonemic ambiguity in the acoustic-phonetic realization of consonant stimuli. Each panel illustrates the mean spectrotemporal difference between stimulus pairs across talkers in the consonant conditions, aligned to the onset of voicing in each stimulus. *Dark shading* shows greater absolute difference between stimuli. In the low-ambiguity condition, /saɪ/ and /baɪ/ differ in terms of manner, place, and voicing, and corresponding spectrotemporal differences can be seen in the high-frequency energy associated with the frication of /s/ and differences in the formant frequencies at the onset of voicing. In the medium-ambiguity condition, /tʰaɪ/ and /baɪ/ differ in terms of place and voicing, and corresponding acoustic-phonetic differences reveal differences in aspiration and onset formant frequencies. In the high-ambiguity condition, /pʰaɪ/ and /baɪ/ differ only in terms of voicing, as evident in the energy differences related to aspiration during voice onset time

indexical variability (single-talker, mixed-talker) and potential phonetic ambiguity (low, medium, high). The model also contained random effects terms of within-participant slopes for indexical variability and phonetic dissimilarity and random intercepts for participants (Barr, Levy, Scheepers, & Tily, 2013). Significance of main effects and interactions was determined by adopting significance criterion of $\alpha = 0.05$, with p values in the mixed-effects linear models based on the Satterthwaite approximation of the degrees of freedom.

Results

Across conditions, participants' word identification accuracy was at ceiling (mean = 99% ± 1%). As such, the primary dependent measure in this study was always response time, consistent with the literature using speeded classification paradigms in speech research (Green et al., 1997; Mullennix & Pisoni, 1990; Nusbaum & Magnuson, 1997; Tomiak, Green, & Kuhl, 1991).

Effects of indexical variability

Compared to the single-talker conditions, response times in the mixed-talker conditions were significantly slower overall (see Fig. 3a and Table 1) (single 787 ms vs. mixed 896 ms; $\beta = 0.047$, $SE = 0.0077$, $t = 6.08$, $p < 9.9 \times 10^{-7}$). Response times in the high-ambiguity condition were significantly slower compared to both the low- ($\beta = 0.026$, $SE = 0.0090$, $t = 2.86$, $p < .009$) and medium-ambiguity conditions ($\beta = 0.041$, $SE = 0.0097$, $t = 4.20$, $p < .0003$) overall. The difference between the average response time in the medium- and low-ambiguity conditions was not significant ($\beta = -0.015$, $SE = 0.0084$, $t = -1.78$, $p = .086$). However, as shown below, these differences were due to the differential increases in processing time required by the respective mixed-talker conditions.

The effect of talker normalization was generally greater for higher ambiguity contrasts than for lower ambiguity ones (see Fig. 3b): There was a significant interaction between indexical variability and potential ambiguity such that the increase in

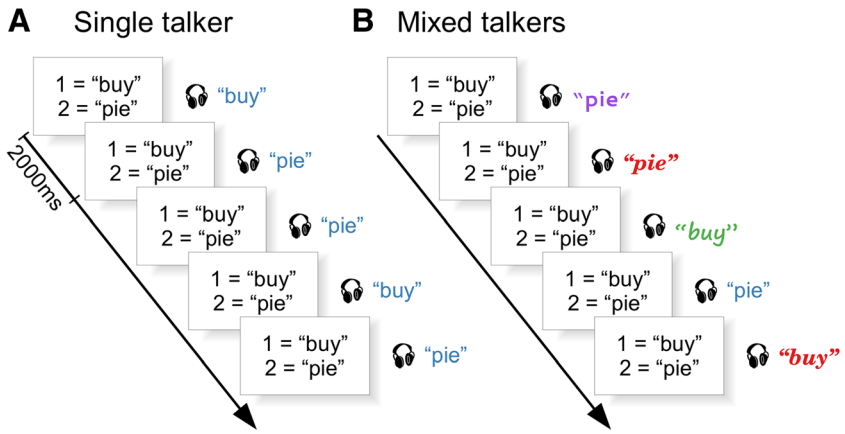

Fig. 2 Task design. Participants performed a speeded word identification task while listening to speech produced by either (**a**) a single talker or (**b**) mixed talkers. The high-potential-ambiguity condition for Experiment 1 is shown

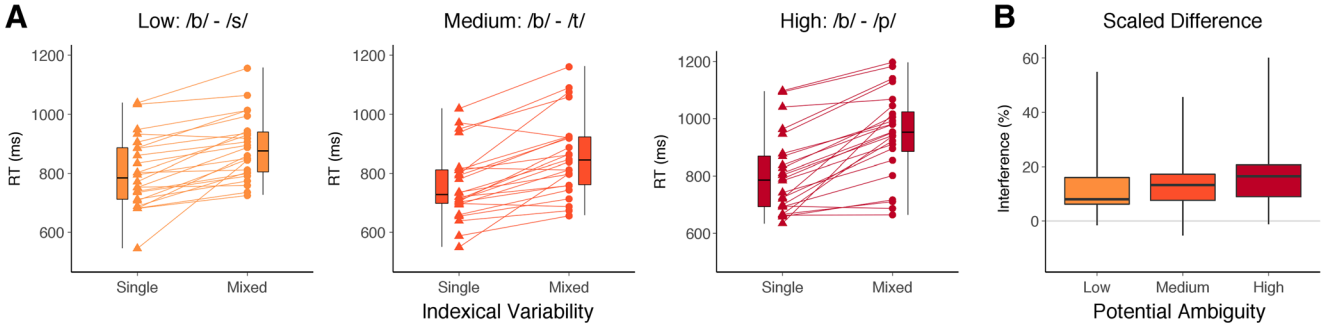

Fig. 3 Effects of indexical variability and potential for acoustic-phonemic ambiguity across talkers on response times for consonant contrasts. **a** Change in response times is shown for individual participants between the single- and mixed-talker conditions across three levels of potential intertalker ambiguity. Box plots in each panel show the distribution (median, interquartile range, extrema) for each variability-by-ambiguity condition. **b** The interference effect of indexical variability is shown for each level of potential ambiguity across talkers. The distribution of differences in response time between the mixed- and single-talker conditions is shown, scaled within-participant to their response time in the single-talker condition: ((mixed – single) / single) × 100. Significant interference was observed for every level of potential intertalker ambiguity; the high-ambiguity condition showed a significantly greater interference effect than either the medium- or low-ambiguity conditions. (Color figure online)

response time between the single- and mixed-talker conditions (i.e., the interference effect) was greater for the high-ambiguity contrast than both the low-ambiguity one (high-ambiguity single/mixed = 805/945 ms vs. low-ambiguity single/mixed = 799/886 ms; $\beta = 0.022$, $SE = 0.0050$, $t = 4.47$, $p < 7.9 \times 10^{-6}$) and the medium-ambiguity one (medium single/mixed = 758/858 ms; $\beta = 0.016$, $SE = 0.0050$, $t = 3.20$, $p < .002$). However, the interaction between the medium- and low-ambiguity conditions was not significant ($\beta = 0.0063$, $SE = 0.0050$, $t = 1.27$, $p = .20$).

Given the presence of both main and interaction effects above, we tested whether the interference effect of indexical variability remained present at every level of potential ambiguity using three separate models—one each for the low-, medium-, and high-potential ambiguity conditions. Response times in the mixed-talker condition were significantly slower than in the single-talker condition for every level of potential ambiguity (low-ambiguity interference: +87 ms, $\beta = 0.047$, $SE = 0.0085$, $t = 5.49$, $p < 1.5 \times 10^{-5}$; medium-ambiguity interference: +100 ms, $\beta = 0.053$, $SE = 0.0087$, $t = 6.13$, $p < 3.1 \times 10^{-6}$; high-ambiguity interference: +141 ms, $\beta = 0.069$, $SE = 0.010$, $t = 6.88$, $p < 5.2 \times 10^{-7}$) (see Table 1). Compared to listening to a single talker, the mixed-talker condition lengthened reaction times by 12% ± 12% in the low-, 14% ± 11% in the medium-, and 18% ± 11% in the high-ambiguity condition (mean ± *SD*) (see Fig. 3B). That is, significant effects of talker normalization were observed for all levels of potential intertalker acoustic-phonemic ambiguity between consonants.

Table 1 Response times (mean ± *SD*, in ms) and interference effects for each level of consonant contrast in Experiment 1

	Potential acoustic-phonemic ambiguity		
	Low	Medium	High
Single talker	799 ± 199	758 ± 176	805 ± 195
Mixed talkers	886 ± 198	858 ± 203	945 ± 229
Difference	87 ± 115	100 ± 111	141 ± 142

Effects of baseline processing speed

Prior studies using Garner speeded classification tasks have shown differences in the degree to which an orthogonal dimension interferes with a target dimension can depend on the relative discriminability of the two dimensions (e.g., Carrell, Smith, & Pisoni, 1981; cf. Mullennix & Pisoni, 1990 vs. Cutler et al., 2011). When dimensions, such as phoneme and talker, are processed integrally, the amount of interference seems to depend principally on the discriminability of the orthogonal dimension (Carrell et al., 1981; Huettel & Lockhead, 1999; Melara & Mounts, 1994). Although in this study the orthogonal dimension was held constant across the various mixed-talker conditions, it is also conceivable that differences in ease with which listeners made the perceptual judgments in the absence of orthogonal interference may have affected the magnitude of that interference. For instance, faster perceptual decisions associated with easier processing may be less susceptible to interference than slower, more processing-intensive ones.

Therefore, as a control, we also investigated whether the magnitude of the interference effects induced by talker variability could be understood in terms of the baseline discriminability of the phonological contrasts in the single-talker condition. The dependent measure in this model was interference (the difference in participants' response times between the mixed- and single-talker conditions) at each level of potential acoustic-phonemic ambiguity. The fixed factor in this analysis was discriminability (participants' mean response time in each single-talker condition). The model also contained random

effects terms of within-participant slopes for discriminability and random intercepts for participants.

The magnitude of processing interference in the mixed-talker conditions relative to the single-talker conditions was not well-characterized by a model of phonetic category discriminability. There was no significant relationship between the amount of interference induced by the mixed-talker conditions and participants' baseline response time in the single-talker conditions ($\beta = -0.13$, $SE = 0.082$, $t = -1.59$, $p = .13$); moreover, the trend for this model was in the opposite direction of a baseline discriminability-based interpretation of interference (i.e., that perceptual decisions made quickly were somewhat more susceptible to interference, not less).

Discussion

The results from the first experiment with consonants show that the magnitude of additional processing cost involved in talker normalization depends on the potential acoustic-phonemic ambiguity between the target phonological contrasts across talkers. The processing cost of talker variability was greatest when the acoustic-to-phonemic mapping was most ambiguous (/b/–/p/) and least when it was most distinct (/b/–/s/). This observation validates the widely stated, but previously untested, assertion that talker normalization facilitates perception of potentially phonetically ambiguous speech sounds by reducing the decision space based on learning the idiosyncratic phonetic realization of target sounds for a given talker (Theodore & Miller, 2010), thereby making speech perception more efficient (Nusbaum & Magnuson, 1997).

In addition to showing that indexical variability elicits a processing delay (e.g., Mullennix & Pisoni, 1990), the results from the experiment with consonants revealed that the effect of talker normalization was always observed—even when the target phonological contrast was acoustically unambiguous across talkers (/b/–/s/). That is, even though no talker's production of /s/ could ever be confused for another talker's production of /b/, the presence of indexical variability nonetheless imposed a significant processing cost on listeners' ability to distinguish this contrast. This observation is not consistent with a restricted model of talker normalization as a process that is brought online only to resolve potential ambiguity among speech sounds. Instead, the observation that talker normalization operates on speech processing in all conditions strongly suggests that this process is an integral part of speech perception.

The nondeterministic relationship between speech acoustics and phonemic categories means that speech perception cannot merely be a process of matching the incoming acoustic signal to an abstract category. Speech perception appears to involve an active process that determines the source of variability and how to resolve the ambiguity given the variability (Heald, Klos, & Nusbaum, 2016), rather than being a strictly passive process of matching the incoming signal to an abstract representation. Influential models of speech perception (e.g., Fowler, 1986; McClelland & Elman, 1986) did not originally account for how listeners are able to resolve the substantial amount of acoustic-phonemic ambiguity that results from the anatomical, articulatory, and dialectal variability across talkers. More recently, authors have begun to develop models of speech processing with an interest in elaborating how the speech perception system balances demands for short-term flexibility to efficiently accommodate the phonetic idiosyncrasies of a particular talker with the needs of maintaining a stable phonology that is robust to generalization in the long term (e.g., Kleinschmidt & Jaeger, 2015; Pierrehumbert, 2016; Sumner, Kim, King, & McGowan, 2014). For example, perception of speech produced by a single talker is predicted to be easier and faster when the listener expects to encounter the same talker repeatedly because they can keep using the same internal representations of the talker-specific speech production. However, when listeners encounter multiple different talkers randomly, they need to either draw upon a larger range of generative models for acoustic-phonemic mapping, or determine enough talker-specific information to select an existing talker-specific model—both of which will incur additional processing costs compared to an accurately predicted talker (Kleinschmidt & Jaeger, 2015). However, even a larger range of generative models can be pared down to just the most relevant given the context (Kleinschmidt & Jaeger, 2015, p. 177)—a process that we might expect to be able to obviate the interference from multiple talkers given a sufficiently unambiguous context, such as /b/ versus /s/. However, the present observation of an interference effect, even in the phonetically unambiguous condition, supports the view that there must be an active and obligatory talker normalization process ongoing in speech perception. That is, even when talker-specific phonetic detail is immaterial to the perceptual decision, the speech perception system must apparently nonetheless expend resources to process trial-by-trial indexical variability.

If this account of talker normalization is correct, then it is important to determine whether this pattern of results generalizes to the perceptual processing of vowels. Prior experiments with processing variability in speech perception have considered vowels and consonants independently (Green et al., 1997; Mullennix & Pisoni, 1990; Strange et al., 1976), sometimes going so far as to suggest that vowel normalization may occur via a distinct, vocal-tract-based approach (Fant, 1973). Thus, in Experiment 2, we applied the design and procedure of Experiment 1 to vowels as the target phonemic contrast. Based on the findings from Experiment 1, we expected to replicate the following effects: (a) that talker normalization would vary as a function of the potential acoustic-phonemic ambiguity across talkers of a given vowel contrast and (b) that the effect of talker normalization would remain significant even for acoustically unambiguous vowel contrasts.

Experiment 2: Vowels

Method

Participants

The same participants (N = 24) who participated in Experiment 1 also completed Experiment 2 during the same visit. The order of the two experiments was counterbalanced across participants.

Stimuli

Stimuli consisted of five naturally spoken English words. They shared the same onset (/b/) and coda (/t/) consonants but had different vowel nuclei (/i/, /ɛ/, /ʌ/, /o/, /u/): *beet, bet, but, boat, boot.* We chose these words because they allowed us to vary the two words in each condition across three levels of potential intertalker ambiguity. For the purpose of this study, we selected vowel category contrasts based on the Euclidean distance between a pair of canonical vowels in F1 × F2 space. Based on the mean F1 and F2 values of all English vowels reported by Hillenbrand and colleagues (1995), we calculated the Euclidean distance of all possible vowel pairs. Among all the vowel pairs, we chose the three vowel pairs with the maximum, median, and minimum Euclidean F1 × F2 distances. The Euclidean distance between canonical vowels was greatest in the low-ambiguity condition (/i/ in *beet* vs. /o/ in *boat*; 1575 Hz), intermediate in the medium-ambiguity condition (/ʌ/ in *but* vs. /ɛ/ in *bet*; 616 Hz), and least in the high-ambiguity condition (/o/ in *boat* vs. /u/ in *boot*; 133 Hz). The acoustic-phonetic similarity of stimuli in the vowel conditions is shown in Fig. 4.

The words were recorded by the same two male and two female native speakers of American English as in Experiment 1, with the same recording and processing procedures. Among numerous tokens of the words recorded by these speakers, the best quality recordings with similar pitch contours and amplitude envelopes were chosen as the final stimulus set.

Procedure

As in Experiment 1, participants performed a speeded word identification task in which we parametrically varied the potential ambiguity of the target vowel contrasts and whether words were spoken by a single talker or mixed talkers. The parameters of stimulus delivery were identical to those of Experiment 1. We manipulated indexical variability as in Experiment 1, presenting recordings from only one talker in half of the blocks (single-talker condition) while presenting tokens from all four talkers in the other half (mixed-talker condition). The talker used in the single-talker condition was counterbalanced across participants. In all blocks, the onset and the coda consonants of the two words were the same while the vowel was manipulated (e.g., *beet* /bit/ vs. *boat* /bot/ in the low-ambiguity condition).

Written instructions assigning a number to the two target words (e.g., "boot = 1; boat = 2" in the high-ambiguity condition) were shown to participants for the duration of each block. Participants listened to the stimuli and identified the spoken word on each trial as quickly and accurately as possible by pressing the corresponding key on a number pad. Trials were presented at a rate of one per 2,000 ms. Stimulus delivery was controlled using PsychoPy Version 1.8.1 (Peirce, 2007). The order of conditions was counterbalanced across participants via Latin square permutations.

Data analysis

Dependent measures included accuracy and response time. Measurement and analysis of these variables was identical to that in Experiment 1. We assessed whether there was an interference effect of talker variability on the response times for speeded classification of words, and whether this effect varied as a function of the potential acoustic-phonemic ambiguity of the target vowel categories. For this analysis, we used a model with the same structure as that in Experiment 1, with fixed factors including indexical variability (single, mixed) and potential phonetic ambiguity (low, medium, high), and random effects terms of within-participant slopes for indexical variability and phonetic ambiguity and random intercepts for participants (Barr et al., 2013). Statistical significance was determined as in Experiment 1.

In this experiment, we conducted an additional analysis to determine whether the acoustic distinctiveness of any specific vowel token was related to how quickly listeners were able to categorize that token in the presence of the indexical variability associated with the mixed-talker condition. That is, we explored whether a highly distinctive token of /u/ would be categorized more quickly than a more ambiguous /u/ when listening to speech from multiple talkers. We operationalized distinctiveness as the distance between each token of one category and the centroid (mean F1× F2) of the other category. For example, in the low-ambiguity condition for vowels, we found the Euclidean distance between each /i/ token and the mean of all the /o/ tokens; and between each /o/ token and the mean of all the /i/ tokens, such that a more distinct /i/ would be further from the /o/ distribution, and a more distinct /o/ would be further from the /i/ distribution. We used this measure of acoustic distinctiveness as a fixed factor in a linear mixed-effects model of response time. We also included overall potential ambiguity as a fixed factor. With this model, we were able to test whether the distinctiveness of individual tokens had an effect on participants' reaction time in categorizing those tokens, and whether that effect varied as a function of the overall potential ambiguity of the contrast they were

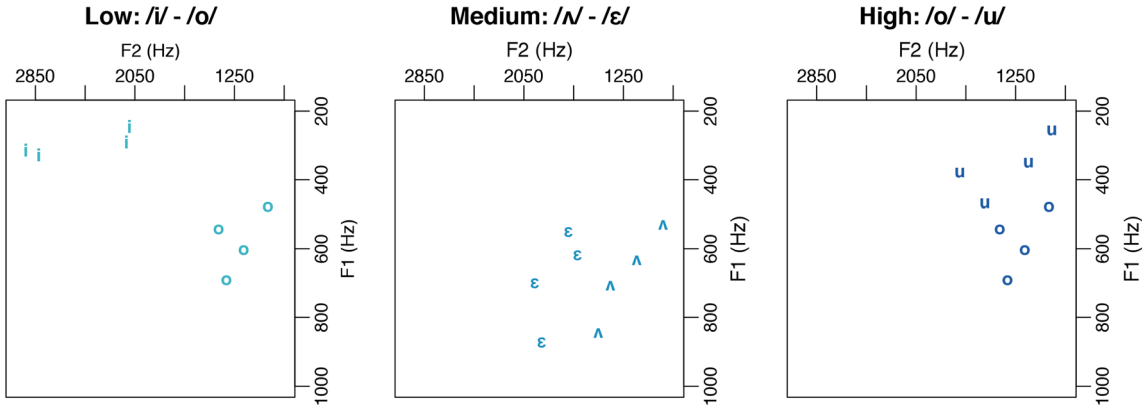

Fig. 4 Potential phonemic ambiguity in the acoustic-phonetic realization of vowel stimuli. Each point on the vowel space chart represents the position of a vowel stimulus spoken by each talker in F1× F2 space. The distance between vowel categories is greatest in the easy phonetic contrast condition and smallest in the hard one. Note that the orientation of the axes is consistent with the articulatory position of the vowels

judging. (This analysis was done for vowels only, because there was no straightforward way to quantitatively operationalize the acoustic distinctiveness between individual consonant tokens in Experiment 1.)

Results

Across conditions, participants' word identification accuracy was at ceiling (mean = 99% ± 2%). As before, the primary dependent measure in this experiment was therefore response time, consistent with the prior literature on interference effects in speech processing.

Effects of indexical variability

Overall response times in the mixed-talker condition were significantly slower than the single-talker condition (see Fig. 5a and Table 2) (single 719 ms vs. mixed 809 ms; $\beta = 0.042$, $SE = 0.0058$, $t = 7.22$, $p < 6.9 \times 10^{-9}$). Compared to the low-ambiguity condition, response times in the high-ambiguity condition were significantly slower ($\beta = 0.046$, $SE = 0.0095$, $t = 4.85$, $p < 4.56 \times 10^{-5}$). Likewise, compared to response times in the medium-ambiguity condition, those in the high-ambiguity condition were significantly slower ($\beta = 0.031$, $SE = 0.0088$, $t = 3.50$, $p < .002$). Response times in the medium-ambiguity condition did not differ from those in the low-ambiguity condition ($\beta = 0.015$, $SE = 0.0077$, $t = 2.00$, $p = .055$). As in Experiment 1, these differences were due to the differential increases in processing time required by the respective mixed-talker conditions.

Like consonants, the effect of talker normalization was generally greater for high-potential ambiguity vowel contrasts than for low-ambiguity ones (see Fig. 5b): There was a significant interaction between indexical variability and acoustic-phonemic ambiguity such that the increase in response time between single- and mixed-talker conditions was greater for the high-ambiguity condition than both the low-ambiguity one (high-ambiguity single/mixed = 737/862 ms vs. low-ambiguity single/mixed = 696/768 ms; $\beta = 0.026$, $SE = 0.0052$, $t = 4.96$, $p < 7.2 \times 10^{-7}$) and medium-ambiguity conditions (medium-ambiguity single/mixed = 724/797 ms; $\beta = 0.025$, $SE = 0.0052$, $t = 4.85$, $p < 1.2 \times 10^{-6}$). The interaction between the low- and medium-ambiguity conditions was not significant ($\beta = 5.07 \times 10^{-4}$, $SE = 0.0052$, $t = 0.10$, $p = .92$).

The presence of both main and interaction effects again required us to test whether the effect of indexical variability was present at every level of potential acoustic-phonemic ambiguity using separate models for each phonetic contrast. The response times in the mixed-talker condition were significantly slower than in the single-talker condition for every level of potential ambiguity (low-ambiguity interference: +72 ms, $\beta = 0.042$, $SE = 0.0078$, $t = 5.39$, $p < 1.8 \times 10^{-5}$; medium-ambiguity interference: +73 ms, $\beta = 0.042$, $SE = 0.0069$, $t = 6.11$, $p < 3.2 \times 10^{-6}$; high-ambiguity interference: +126 ms, $\beta = 0.068$, $SE = 0.0073$, $t = 9.27$, $p < 3.2 \times 10^{-9}$). Compared to listening to a single talker, the mixed-talker condition lengthened reaction times by 11% ± 11% in the low-ambiguity condition, 11% ± 9% in the medium-ambiguity condition, and 17% ± 10% in the high-ambiguity condition (mean ± *SD*) (see Fig. 5b). That is, significant effects of talker normalization were observed for all levels of potential intertalker acoustic-phonemic ambiguity between vowels.

Effects of baseline processing speed

As in Experiment 1, we investigated whether differences in the magnitude of processing interference in the mixed-talker conditions could be understood as a function of stimulus discriminability in the single-talker condition. The dependent measure in this model was the amount of interference at each level of potential acoustic-phonemic ambiguity, and the fixed factor was discriminability (participants' mean response time

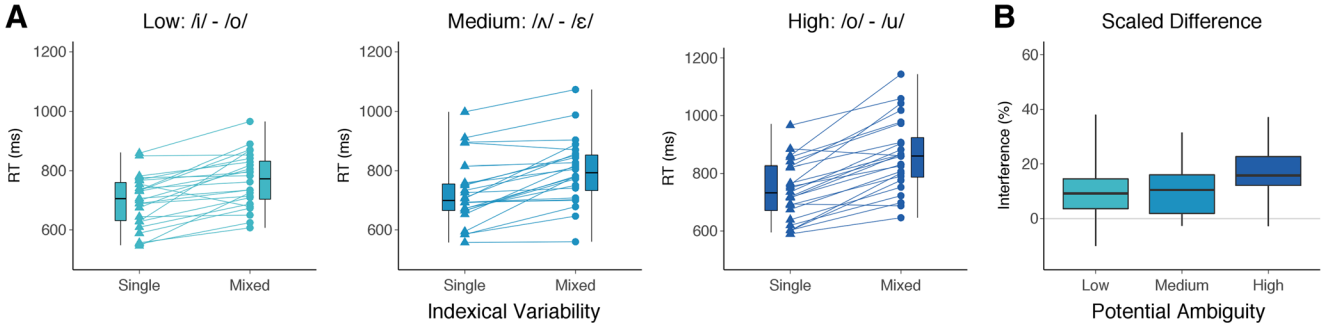

Fig. 5 Effects of indexical variability and potential for acoustic-phonemic ambiguity across talkers on response times for vowel contrasts. **a** Change in response times is shown for individual participants between the single- and mixed-talker conditions across three levels of potential intertalker ambiguity in the vowel conditions. Box plots in each panel show the distribution (median, interquartile range, extrema) for each variability-by-ambiguity condition. **b** The interference effect of indexical variability is shown for each level of potential intertalker ambiguity. The distribution of differences in response time between the mixed- and single-talker conditions is shown, scaled within-participant to their response time in the single-talker condition: ((mixed – single) / single) × 100. Significant interference was observed for every level of potential intertalker ambiguity; the high-ambiguity condition showed a significantly greater interference effect than either the medium- or low-ambiguity conditions

in each single-talker condition). The model also contained the same random effects terms as in Experiment 1.

Again, the magnitude of processing interference in the mixed-talker conditions relative to the single-talker conditions was not well-characterized by a model of baseline phonetic category discriminability. There was no significant relationship between the amount of interference induced by the mixed-talker conditions and participants' baseline response time in the single-talker conditions for vowels ($\beta = -0.13$, $SE = 0.089$, $t = -1.50$, $p = .15$). Like the model for consonants, the trend for the vowel model was in the opposite direction of a baseline discriminability-based interpretation of interference, with perceptual decisions made quickly being slightly (but not significantly) more susceptible to interference than those made more slowly.

Effects of individual token distinctiveness

We also investigated whether the acoustic distinctiveness of individual vowel tokens affected the speed at which listeners categorized them in the mixed-talker condition, and whether this effect varied as a function of the overall potential ambiguity. The acoustic distinctiveness of individual tokens had no effect on the speed with which listeners categorized them, regardless of the potential ambiguity of the two categories (low-ambiguity: $\beta = -5.20 \times 10^{-6}$, $SE = 9.24 \times 10^{-6}$, $t = 0.56$, $p = .58$; medium-ambiguity: $\beta = 9.67 \times 10^{-6}$, $SE = 1.50 \times 10^{-5}$, $t = -0.65$, $p = 0.52$; high-ambiguity: $\beta = -1.49 \times 10^{-5}$, $SE = 3.14 \times 10^{-5}$, $t = -0.48$, $p = .64$) (see Fig. 6). Correspondingly, there were no Condition × Distinctiveness interactions (low vs. medium: $\beta = 1.53 \times 10^{-5}$, $SE = 1.86 \times 10^{-5}$, $t = 0.83$, $p = .41$; medium vs. high: $\beta = 5.50 \times 10^{-6}$, $SE = 3.27 \times 10^{-5}$, $t = 0.19$, $p = .85$; low vs. high: $\beta = 2.14 \times 10^{-5}$, $SE = 3.01 \times 10^{-5}$, $t = 0.71$, $p = .48$). That is, even the most acoustically distinct tokens of a particular vowel category were not identified more quickly than the less distinct tokens during any of the mixed-talker conditions.

Table 2 Response times (mean ± *SD*, in ms) and interference effects for each level of vowel contrast in Experiment 2

	Potential acoustic-phonemic ambiguity		
	Low	Medium	High
Single talker	696 ± 152	724 ± 200	737 ± 180
Mixed talkers	768 ± 172	797 ± 228	862 ± 191
Difference	72 ± 106	73 ± 92	126 ± 97

Discussion

The results of Experiment 2 replicate the findings from Experiment 1, showing that the magnitude of additional processing cost introduced by talker normalization varies as a function of the potential acoustic-phonetic ambiguity of given sound contrast across talkers. Moreover, the effect of talker normalization was significant even when the potential ambiguity in the target sound contrast was essentially nonexistent (/i/–/o/).

Talker normalization in vowel perception has often been described in terms of vocal tract normalization (Fant, 1973). According to this explanation, variability in vowels among talkers is a consequence of anatomical difference in vocal tracts. Thus, the talker normalization process factors out this source of variability so that listeners can reach the same abstract phonetic representation regardless of the talker who produced the speech sound. However, the results from Experiment 2 show that the patterns of talker normalization for vowel perception operate in an analogous way to those of consonant perception observed in Experiment 1, even though

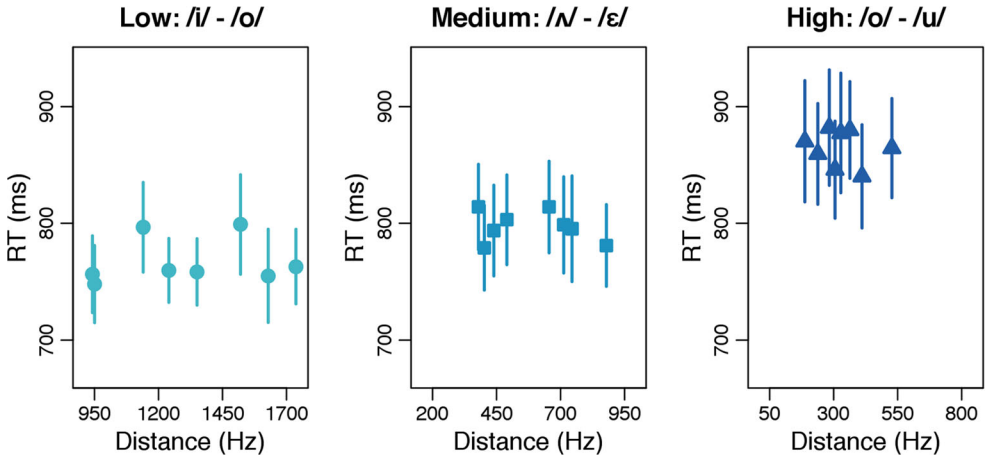

Fig. 6 Acoustic distinctiveness of individual tokens and average response times to them. The distinctiveness of individual tokens was measured as the Euclidean distance within F1× F2 space between each token of one category and the centroid of the contrasting category. Each point is the average response time across participants for each token. *Error bars* represent the standard error of mean. Within each condition, the acoustic distinctiveness of individual tokens did not have a significant effect on response times

normalizing vocal tract anatomy alone does not provide sufficient information about the sources of variability in consonant articulation across talkers—particularly timing (Theodore, Miller, & DeSteno, 2009). The shared pattern of talker normalization for vowel and consonant perception suggests that talker normalization as a cognitive process may operate holistically on both consonants and vowels, and correspondingly is unlikely to be ascribed to a simple normalization of resonance differences due to variability in vocal-tract anatomy. (Indeed, it is a distinct possibility that the mechanisms for phonetic adaptation, such as talker normalization, may reflect a specific instantiation of a more general process for adapting perception to local stimulus statistics; e.g., Laing, Liu, Lotto, & Holt, 2012; Perrachione et al., 2016.)

For all levels of vowel ambiguity in the mixed-talker condition, the acoustic distinctiveness of individual vowel tokens did not make a significant difference in the response times. If speech perception could be achieved via a direct mapping between the acoustic signal and phonetic representation, then we might have expected that less ambiguous stimuli themselves would be perceived faster than more ambiguous ones. Contrary to this expectation, the response times for vowel classification were not affected by the relative distinctiveness of each vowel token, again suggesting that, in the presence of talker variability, intrinsic talker normalization operates comprehensively on all encountered speech signals. The lack of effects of acoustic distinctiveness of individual tokens on response time further supports our finding that talker normalization is an obligatory component of speech perception.

General discussion

The results from the present study further our understanding of how listeners extract stable phonological information from speech signals with substantial acoustic-phonetic variability across talkers. First, the results suggest that the magnitude of the additional processing cost imposed by talker variability on speech processing depends on the potential acoustic-phonetic ambiguity between target phonological contrasts across talkers. The processing cost of talker variability was greatest when the acoustic-to-phonemic mapping was most ambiguous and least (but still present) when the mappings were wholly distinct. Second, in addition to showing that indexical variability elicits a processing delay for talker normalization (e.g., Mullennix & Pisoni, 1990), these results further reveal that the effect of talker normalization is observed even when the target speech sounds are acoustically unambiguous across talkers (e.g., /b/–/s/ and /i/–/o/). That is, even when indexical variability does not obscure the target phonemic contrast, speech perception processes nonetheless appear obligated to normalize the incoming signal.

Crucially, this result cannot be ascribed to the mere presence of acoustic variability, because not all sources of variability impact speech processing. For instance, while acoustic variability due to differences among talkers or speech rate incurs a processing cost (Tomiak et al., 1991), acoustic variability due to differences in amplitude consistently does not impact speech processing (Bradlow, Nygaard, & Pisoni, 1999; Sommers, Nygaard, & Pisoni, 1994).

The acoustics of talkers' speech may be similar for different target sounds, or different for the same target sound. Correspondingly, it is often asserted that the purpose of talker normalization is to reduce the perceptual challenges of talker-related variability in the acoustic realization of speech (Nusbaum & Magnuson, 1997). However, whether talker normalization differentially facilitates the identification of speech sounds that are more or less ambiguous had not previously been demonstrated. Here, we found that the additional processing cost involved in talker normalization during speech

perception is larger for more potentially ambiguous speech sounds, consistent with the view that talker normalization is a resource-dependent mechanism for accommodating indexical variability. For both consonants and vowels, identification of sounds that were most similar in terms of their acoustics showed the greatest effect of talker normalization.

For both consonants and vowels, the effect of talker normalization was significantly greater for the high-ambiguity condition than both the medium- and low-ambiguity conditions, but did not differ significantly between the latter two. One possible explanation for this pattern of results is that the processing cost of indexical variability may be the same as long as there is some unambiguous dimension upon which listeners can base their categorization. For example, in the low-ambiguity condition for vowels, all tokens of /i/ and /o/ are categorically distinct along both F1 and F2 dimensions, regardless of who says them (see Fig. 4). In the medium-ambiguity condition, although /ɛ/ and /ʌ/ have overlapping F1, in the present sample these two phonemes are categorically distinct with regards to F2 (and, indeed, vowel categories tend to exhibit a great deal of front–back distinctiveness, even when varying internally in height; Hillenbrand et al., 1995). In the high-ambiguity condition, however, /u/ and /o/ cannot be wholly distinguished based on either F1 or F2, and correspondingly it is here that we find the greatest cost incurred by talker variability. A similar case may be made for the consonants: The acoustic dimensions distinguishing voiced from voiceless stop consonants /b/ and /p/ are considerably fewer, and substantially more overlapping, across talkers (e.g., Allen et al., 2003; Lisker & Abramson, 1964; Stuart-Smith, Sonderegger, Ratchke, & Macdonald, 2015) than those distinguishing consonants that differ in place or manner. Although we drew on differences in voicing, place, and manner to operationalize different levels of acoustic-phonemic ambiguity across talkers, it is important to acknowledge that these dimensions may not, in fact, be those used in underlying phonological representations, as prior work in speech learning has shown that the relevant acoustic dimensions that listeners can learn to emphasize during speech perception can have considerable situational specificity (e.g., Idemaru & Holt, 2013; Reinisch, Wozny, Mitterer, & Holt, 2014).

Furthermore, our investigation of the effects of acoustic distinctiveness of individual tokens on response time in the mixed-talker condition provides additional evidence that perceptual adjustment to talker is an obligatory process. Within each vowel contrast, the acoustic distinctiveness of any particular token did not affect listeners' response time in the mixed-talker condition, as we might have expected if there were a direct route from acoustics and phonemes. More perceptually distinct within-category tokens were not recognized faster than less distinct ones, suggesting that acoustics-only strategies for processing speech do not operate independently of intrinsic talker normalization. The limited number of tokens in this study requires a conservative interpretation of this result, and further study with a larger number of stimuli with greater within-category variability is warranted to confirm this observation. However, the results of the acoustic dissimilarity analysis for vowels are paralleled by the results of the perceptual dissimilarity analyses for both consonants and vowels: In neither case is the amount of interference explained by participants' baseline processing of these contrasts. Together, these supplementary analyses further suggest that the processing costs incurred by talker normalization reflect an obligatory effort to resolve potential ambiguity in acoustic-phonemic mappings across talkers.

Another type of model that has frequently been used to explain effects of talker variability in speech processing is episodic (or exemplar-based) representations (Goldinger, 1998; Johnson, 1997). Although episodic and talker normalization models are frequently described as incompatible alternatives, no study has yet presented evidence in favor of one model while simultaneously presenting evidence that falsifies the predictions of the alternative process. Rather, talker normalization can be understood as an active cognitive process operating on speech in real time and running in parallel with the episodic memory processes that store traces of encountered speech. Such "hybrid" models of speech processing have also been advocated by others (Luce & McLennan, 2005; Pierrehumbert, 2016; Zhang & Chen, 2016). Although we observed an interference effect due to talker variability in every condition, experiments based in the episodic framework sometimes fail to demonstrate variability-related effects (e.g., Goldinger, Pisoni, & Logan, 1991; McLennan & Luce, 2005). The time-course hypothesis (Luce, McLennan, & Charles-Luce, 2003) has been put forward as one possible explanation for when and why talker variability impacts speech processing. According to this hypothesis, the effect of talker variability emerges only when processing time is slowed down by the nature of the task or stimuli. Fast and easy tasks require access only to abstract representations, but, as processing slows, episodic memories play a greater role. Thus, the effect of talker variability is alleged to slow down speech processing only when the task is harder. Consistent with this hypothesis, the talker-variability effect in the present experiments was greatest in the most ambiguous conditions. However, contrary to the predictions of the time-course hypothesis, our results also showed that there was a significant effect of talker variability, even in the least ambiguous conditions, implying that listeners were affected by talker-specific information even when performing fast and easy tasks. This is paralleled by the discriminability analyses, which hint at the possibility that fast decisions may actually be more susceptible to interference than slower ones. This observation highlights important methodological differences between tasks purporting to demonstrate normalization effects, which principally employ phonological decisions (Mullennix & Pisoni, 1990), versus those

investigating episodic effects, which principally employ memory paradigms or decisions based on semantic and word-level stimulus features (Goldinger, 1996; Palmeri, Goldinger, & Pisoni, 1993; Theodore, Blumstein, & Luthra, 2015).

An alternative interpretation of the talker-specificity effect in memory for speech has been proposed by Theodore and colleagues (2015), implicating attention during encoding, rather than processing time, as the source of the "time-course" effect. In their study, participants exhibited a talker-specificity effect only when their attention was directed to the talker's identity. Voice processing is indeed an attention-demanding process rather than an automatic process (Mullennix & Howe, 1999). In contrast, the results from the present study—and, indeed, all prior speeded classification studies incorporating talker variability—revealed that talker-variability effects in online speech perception emerge even when participants make decisions quickly and are not directed to pay attention to talker identity. Although explicit instruction to attend to voices may motivate participants to allocate additional attentional resources to voice processing—and to demonstrate talker normalization effects where they otherwise might not (Magnuson & Nusbaum, 2007)—our results suggest that, for natural stimuli, simultaneous processing of talker-specific information is an integrated and indeed mandatory part of speech perception. The robust effect of talker variability even in the absence of explicit instructions to focus on talker information further suggests that talker normalization is an active process during speech perception, rather than taking effect solely during the encoding of episodic memory (cf. Goldinger, 1998).

Previous studies of perceptual accommodation of between-talker variability in speech processing have also put forward the supposition that perceptual adjustments to voice may be mandatory. For example, Mullennix and Pisoni (1990) found that the indexical dimension cannot be selectively ignored when phonetic classifications are required—a result that has been taken to mean that allocation of attention to talker-specific information is mandatory. However, the scope of this assertion had not previously been investigated, particularly when there is no acoustic reason for between-talker variability to confound the perception of target speech sounds, such as for phonological contrasts that are acoustically unambiguous across talkers. In the present study, we specifically addressed whether talker-specific processing is obligatory even when there is no potential ambiguity in the acoustic-to-phonemic correspondence for the given stimuli. Although the processing cost of talker normalization is smaller for less ambiguous sounds, the effect nonetheless remains significant even when there is no potential ambiguity across talkers between two sounds. In both the vowel and consonant conditions, the target sounds in the easy phonetic contrasts were wholly acoustically distinct. This unambiguous acoustic information alone could presumably be sufficient for listeners to perceive the sounds accurately. Nevertheless, we observed a significant additional processing cost for listening to mixed talkers relative to a single talker, demonstrating that listeners are engaging in intrinsic talker normalization on a trial-to-trial basis for unambiguous sounds in the mixed-talker condition, just as they do for more ambiguous speech sounds. This result suggests that talker normalization is indeed an obligatory part of speech processing rather than an ancillary cognitive process that is brought online only to facilitate the perception of potentially ambiguous sounds.

Conclusion

The results from this study show (a) that the extent to which intrinsic talker normalization affects speech perception depends on the potential ambiguity between target speech sounds, and (b) that talker normalization is indeed an obligatory component of speech perception, even when there is no potential ambiguity to resolve.

Acknowledgements We thank Sara Dougherty and Terri Scott for their assistance. Research reported in this article was supported by the NIDCD of the National Institutes of Health under award number R03DC014045. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

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.

Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *Journal of the Acoustical Society of America, 71,* 975–989.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5,* 341–345.

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perceptual Psychophysics, 61,* 206–219.

Carrell, T. D., Smith, L. B., & Pisoni, D. B. (1981). Some perceptual dependencies in speeded classification of vowel color and pitch. *Perception & Psychophysics, 29,* 1–10.

Chandrasekaran, B., Chan, A. H. D., & Wong, P. C. M. (2011). Neural processing of what and who information during spoken language processing. *Journal of Cognitive Neuroscience, 23,* 2690–2700.

Cutler, A., Andics, A., & Fang, Z. (2011). *Inter-dependent categorization of voices and segments*. 17th meeting of the International Congress of Phonetic Sciences, Hong Kong.

Fant, G. (1973). *Speech sounds and features*. Cambridge, MA: MIT Press.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14,* 3–28.

Garner, W. R. (1974). *The processing of information and structure*. Potomac, MD: Erlbaum.

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.

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology–Learning, Memory, & Cognition, 17,* 152–162.

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.

Heald, S., Klos, S., & Nusbaum, H. C. (2016). Understanding speech in the context of variability. In G. Hickok & S. Small (Eds.), *Neurobiology of language* (pp. 195–208). San Diego, CA: Academic Press.

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.

Holt, L. L. (2006). Speech categorization in context: Joint effects of nonspeech and speech precursors. *Journal of the Acoustical Society of America, 119,* 4016–4026.

Huettel, S. A., & Lockhead, G. R. (1999). Range effects of an irrelevant dimension on classification. *Perception & Psychophysics, 61,* 1624–1645.

Idemaru, K., & Holt, L. L. (2013). Specificity of dimension-based statistical learning in word recognition. *Journal of Experimental Psychology–Human Perception and Performance, 40,* 1009–1021.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–155). San Diego, CA: Academic Press.

Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 363–389). Malden, MA: Blackwell.

Kaganovich, N., Francis, A. L., & Melara, R. D. (2006). Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. *Brain Research, 1114,* 161–172.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148–203.

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

Laing, E. J. C., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a tune: Talker normalization via general auditory processes. *Frontiers in Psychology, 3,* 203.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20,* 384–422.

Luce, P. A., & McLennan, C. T. (2005). Spoken word recognition: The challenge of variation. In D. B. Pisoni & R. E. Remez (Eds.), *The handbook of speech perception* (pp. 591–609). Malden, MA: Blackwell.

Luce, P. A., McLennan, C. T., & Charles-Luce, J. (2003). Abstractness and specificity in spoken word recognition: Indexical and allophonic variability in long-term repetition priming. In J. Bowers & C. Marsolek (Eds.), *Rethinking implicit memory* (pp. 197–214). Oxford, UK: Oxford University Press.

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

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

McLennan, C. T., & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology–Learning, Memory, & Cognition, 31,* 306–321.

Melara, R. D., & Mounts, J. R. W. (1994). Contextual influences on interactive processing: Effects of discriminability, quantity, and uncertainty. *Perception & Psychophysics, 56,* 73–90.

Miller, J. L., & Baer, T. (1983). Some effects of speaking rate on the production of [b] and [w]. *Journal of the Acoustical Society of America, 73,* 1751–1755.

Morton, J. R., Sommers, M. S., & Lulich, S. M. (2015). The effect of exposure to a single vowel on talker normalization for vowels. *Journal of the Acoustical Society of America, 137,* 1443–1451.

Mullennix, J. W., & Howe, J. N. (1999). Selective attention in perceptual adjustments to voice. *Perceptual and Motor Skills, 89,* 447–457.

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

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America, 85,* 365–378.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America, 85,* 2088–2113.

Norris, D., McQueen, J., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47,* 204–238.

Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as s cognitive process. In K. Johnson & J. W. Mullenix (Eds.), *Talker variability in speech processing* (pp. 109–132). San Diego, CA: Academic Press.

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, J. W. (2007). PsychoPy: Psychophysics software in Python. *Journal of Neuroscience Methods, 162,* 8–13.

Perrachione, T. K., Del Tufo, S. N., Winter, R., Murtagh, J., Cyr, A., Chang, P…Gabrieli, J. D. E. (2016). Dysfunction of rapid neural adaptation in dyslexia. *Neuron*, *92,* 1383–1397.

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics, 2,* 33–52.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson & J. W. Mullenix (Eds.), *Talker variability in speech processing* (pp. 9–32). San Diego, CA: Academic Press.

Reinisch, E., Wozny, D. R., Mitterer, H., & Holt, L. L. (2014). Phonetic category recalibration: What are the categories? *Journal of Phonetics, 45,* 91–105.

Sjerps, M. J., McQueen, J. M., & Mitterer, H. (2013). Evidence for precategorical extrinsic vowel normalization. *Attention, Perception, & Psychophysics, 75,* 576–587.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition: I. Effects of variability in speaking rate and overall amplitude. *Journal of the Acoustical Society of America, 96,* 1314–1324.

Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America, 60,* 213–224.

Stuart-Smith, J., Sonderegger, M., Ratchke, T., & Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology, 6,* 505–549.

Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology, 4,* 1015. https://doi.org/10.3389/fpsyg.2013.01015

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America, 79,* 1086–1100.

Theodore, R. M., Blumstein, S. E., & Luthra, S. (2015). Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis. *Attention, Perception, & Psychophysics, 77,* 1674–1684.

Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific phonetic detail. *Journal of the Acoustical Society of America, 128,* 2090–2099.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: Contextual influences. *Journal of the Acoustical Society of America, 125,* 3974–3982.

Tomiak, G. R., Green, K. P., & Kuhl, P. K. (1991). Phonetic coding and its relationship to talker and rate normalization. *Journal of the Acoustical Society of America, 90,* 2363.

Volaitis, L. E., & Miller, J. L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America, 92,* 723–735.

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience, 16,* 1173–1184.

Zhang, C., & Chen, S. (2016). Towards an integrative model of talker normalization. *Journal of Experimental Psychology–Human Perception and Performance, 42,* 1252–1268.

Zhang, C., Pugh, K. R., Mencl, W. E., Molfese, P. J., Frost, S. J., Magnuson, J. S.,…Wang, W. S.-Y. (2013). Functionally integrated neural processing of linguistic and talker information: An event-related fMRI and ERP study. *NeuroImage*, *124*, 536–549.