JASA ARTICLE

Multiple sources of acoustic variation affect speech processing efficiency[a)]

Alexandra M. Kapadia, Jessica A. A. Tin, and Tyler K. Perrachione[b)]

Department of Speech, Language, and Hearing Sciences, Boston University, 635 Commonwealth Avenue, Boston, Massachusetts 02215, USA

ABSTRACT:

Phonetic variability across talkers imposes additional processing costs during speech perception, evident in performance decrements when listening to speech from multiple talkers. However, within-talker phonetic variation is a less well-understood source of variability in speech, and it is unknown how processing costs from within-talker variation compare to those from between-talker variation. Here, listeners performed a speeded word identification task in which three dimensions of variability were factorially manipulated: between-talker variability (single vs multiple talkers), within-talker variability (single vs multiple acoustically distinct recordings per word), and word-choice variability (two- vs six-word choices). All three sources of variability led to reduced speech processing efficiency. Between-talker variability affected both word-identification accuracy and response time, but within-talker variability affected only response time. Furthermore, between-talker variability, but not within-talker variability, had a greater impact when the target phonological contrasts were more similar. Together, these results suggest that natural between- and within-talker variability reflect two distinct magnitudes of common acoustic–phonetic variability: Both affect speech processing efficiency, but they appear to have qualitatively and quantitatively unique effects due to differences in their potential to obscure acoustic–phonemic correspondences across utterances.

 https://doi.org/10.1121/10.0016611

(Received 31 March 2022; revised 14 November 2022; accepted 7 December 2022; published online 11 January 2023)

[Editor: Matthew B. Winn] Pages: 209–223

I. INTRODUCTION

A long-standing challenge in the scientific study of speech perception has been to explain how listeners can access stable linguistic percepts in the face of highly variable acoustic signals. From even the earliest acoustic analyses of speech, it was noted that there is considerable variability in speech acoustics, including in the phonetic dimensions that are important for distinguishing linguistically contrastive phonological categories (Potter and Steinberg, 1950). This has been termed the *lack of invariance problem* in speech perception (e.g., Pisoni, 1981). Extensive work has been conducted to characterize the immense degree of variability that the speech perception system must overcome: Speech acoustics vary depending on the specific configuration of a talker's vocal tract, the local phonetic environment in continuous speech, differences in vocal source and vocal tract anatomy and physiology between different talkers, talkers' emotional states, the reverberant characteristics of the environment, sociocultural factors such as speakers' dialects, and so forth. This variation poses a challenge not only for the mature, intact speech perception system, but also for infants who must learn the relevant phonological contrasts in their native language in the face of between- and within-talker variation (Pierrehumbert, 2003; van der Feest *et al.*, 2022). Although rarely noted explicitly, the lack of invariance problem in speech perception is not wholly unlike the challenges that natural environments pose to perceptual systems more generally; for instance, in vision, recognizing an object may require overcoming variations due to scenes that involve different intensities, colors, or orientations of illumination; partial occlusion by other objects; or more or less canonical viewing orientations. In this report, we consider how such different sources of variability in the acoustic signal for speech affect the speed and accuracy of spoken word identification.

The principal source of acoustic variability affecting phonemic contrasts in speech is differences in the vocal tract resonance and articulatory dynamics among different talkers (Kleinschmidt, 2019). Numerous studies have shown that these differences incur costs in terms of speech processing efficiency: Listeners are slower and less accurate to recognize the content of speech when it is spoken by multiple different talkers compared to listening to a single consistent talker (Green *et al.*, 1997; Mullennix *et al.*, 1989; Choi *et al.*, 2018, 2022; Morton *et al.*, 2015; Perrachione *et al.*, 2016; Stilp and Theodore, 2020; Kapadia and Perrachione, 2020; Heald and Nusbaum, 2014). These "talker variability" effects are impressive in their reliability, not only across studies, but also across manipulations that are specifically designed to attenuate the effect of talker variability:

[a)]This paper is part of a special issue on Reconsidering Classic Ideas in Speech Communication.

[b)]Electronic mail: tkp@bu.edu

Listening to speech from mixed talkers incurs additional processing costs even when there is no potential acoustic ambiguity between the target speech contrasts (Choi *et al.*, 2018) and when the talkers are highly personally familiar to listeners (Magnuson *et al.*, 2021).

Much of the early research on processing speech variability considered acoustic variation among talkers as a source of noise, proposing a variety of computational solutions whereby variable speech acoustics could be "normalized" to mitigate the irrelevant variation and extract stable phonological categories (e.g., Nearey, 1989; Sussman, 1986). Indeed, much of the prior psycholinguistic work on the cognitive processes behind accommodating variability in speech acoustics has explicitly termed these operations "talker (or speaker) normalization" or "talker adaptation" (e.g., Johnson, 2005; Pisoni, 1997; Sjerps *et al.*, 2019; Wong *et al.*, 2004; Zhang and Chen, 2016). Recent work has moved away from the idea of explicit normalization, seeing variation among talkers instead as an inherent part of the representational and computational architecture of speech processing (Kleinschmidt and Jaeger, 2015; Scott, 2019; Heald *et al.*, 2016). However, both the idea of normalization and its recent reinterpretation as meaningful systematicity begs the question of whether there is something "privileged" about the kind of variability in speech acoustics that results from differences between talkers, as opposed to any other source of variation, such as differences in speech acoustics within a talker from utterance to utterance. Historically, considerably less attention has been paid to the question of within-talker variability in speech acoustics from the perspective of speech perception, although this question is prominent in the domain of talker identification, where listeners must be able to "tell together" within-talker variability to reliably identify a talker's voice across different utterances (Lavan *et al.*, 2019a; Lavan *et al.*, 2019b; Lee *et al.*, 2019; Perrachione *et al.*, 2019).

One line of evidence suggesting that between-talker variation in speech perception may indeed be a privileged kind of variation comes from studies of talker-specific speech processing (Souza *et al.*, 2013). When listeners are familiar with a talker's voice (and therefore, presumably, their phonetic idiosyncrasies), they are more accurate at perceiving that talker's speech, including in adverse listening situations, such as background noise, and even for novel speech content that they had not heard from that talker before (Nygaard *et al.*, 1994). This phenomenon makes sense in the context of studies showing there is structure in talker variability (Kleinschmidt, 2019), which listeners can exploit to make generalizable predictions about a particular talker's speech in different contexts and even for different phonetic contrasts (Allen *et al.*, 2003; Clayards, 2018; Theodore and Miller, 2010). Indeed, the presumption that talker variability is a privileged kind of variability is also evident in the literature on second language learning, where "high variability training paradigms" are routinely studied as an approach to enhance the speed, accuracy, and generalizability of learning to perceive novel phonological contrasts (Perrachione *et al.*, 2011; Sadakata and McQueen, 2014; Kingston, 2003; Clopper and Pisoni, 2004; Barcroft and Sommers, 2005). In this literature, "high variability" is almost always implemented using stimuli produced by multiple different talkers, as opposed to any other kind of variability, such as multiple tokens from a single talker, in different coarticulatory configurations, in different kinds of reverberant environments, or with different kinds or levels of background noise.

Despite the prominence of talker variability in the prior literature, there is more limited evidence that some (though not all) other sources of variability affect speech perception. For instance, while listeners are less likely to recognize that they had heard a word previously if it is spoken by a different talker (Palmeri *et al.*, 1993), they are also less likely to recognize that they had heard a word before if it is spoken by the same talker but at a different rate (Bradlow *et al.*, 1999). Trial-by-trial variability in speech rate is similarly deleterious for on-line speech identification accuracy (Sommers and Barcroft, 2006; Uchanski *et al.*, 1992) and speed (Newman *et al.*, 2001). However, simple differences in stimulus amplitude (loudness) do not affect either recognition accuracy or memory, suggesting that phonetic, but not simply acoustic, variability impacts speech processing (Sommers *et al.*, 1994; Nygaard *et al.*, 1995; cf. Pufahl and Samuel, 2014). Similarly, although there are differences in intelligibility across talkers (Bradlow *et al.*, 1996), and listeners have intelligibility advantages for a familiar talker (Nygaard *et al.*, 1994; Holmes *et al.*, 2018), not all speech from a single talker is equally intelligible (Smiljanic and Bradlow, 2009). Within-talker variation in speech can lead to differences in not just perception, but also memory for speech (Keerstock and Smiljanic, 2019). Such results challenge the common notion that there is something *special*, as opposed to something merely *salient*, about the impacts of between-talker variability on speech processing.

Similarly, asymmetric perceptual interference due to variability between speech segments and talker identity has been used to argue for talker-to-phoneme directional processing dependencies between these two types of information (Mullennix and Pisoni, 1990). However, subsequent work has shown that the degree of within- vs between-talker variability in segmental and voice contrasts can reverse the apparent direction of this processing dependency (Cutler *et al.*, 2011). This underscores a critical limitation of Garner-like paradigms (Garner, 1974) more generally: that the direction of processing dependency effects depends specifically on the magnitude of variation chosen for each dimension, rather than something inherent about the processing order between dimensions (e.g., Huettel and Lockhead, 1999).

A related argument against the idea that between-talker differences are a privileged source of acoustic variability in speech is that the magnitude of acoustic differences between talkers is also variable (e.g., Perrachione *et al.*, 2019). Presumably, therefore, the cognitive or perceptual consequences of variability between more similar-sounding talkers should be less than that between more different-sounding talkers. However, the predictions of certain models of talker-specific speech processing seem to be explicitly at odds with this idea. For example, the *active control framework* for processing speech variability proposes that a special, computationally intensive mode of speech processing is engaged whenever listeners encounter (or expect) variation in speech due to different talkers (Nusbaum and Magnuson, 1997; Magnuson *et al.*, 2021). No allowance is made in this framework for whether the resources demanded by that mode of perception depend on how much variability there might be in the target speech. Indeed, there is some evidence supporting this view: Talker variability appears to impose a fixed amount of processing cost, regardless of how many different talkers might be encountered (Kapadia and Perrachione, 2020; Mullennix and Pisoni, 1990), and these costs can persist even when listeners have context that cues the target talker (Choi *et al.*, 2022; Morton *et al.*, 2015). Furthermore, a fixed amount of acoustic variability in speech stimuli (due to minor variations in voice pitch) may impose greater or lesser processing costs depending on whether listeners are told that those variations are due to differences between two talkers, as opposed to exemplar variation within a single talker (Magnuson and Nusbaum, 2007).[1] However, a separate line of evidence appears to support the idea that the degree of between-talker variability does incur different amounts of processing cost: Smaller variations in voice acoustics (namely, voice pitch) have been shown to incur smaller processing costs compared to larger variations in voice pitch (Stilp and Theodore, 2020)–an observation that appears to be consistent with an idea that talker-variability effects reflect the contribution of domain-general auditory processing mechanisms (e.g., Sjerps *et al.*, 2013).

A better theoretical understanding of within- vs between-talker variability effects on speech processing is constrained by the quality and diversity of empirical data regarding these effects. A concern about the current understanding of processing variability in speech is that much of the prior literature relies on simple two-alternative forced choice tasks, where listeners decide between two target words (e.g., "boot" vs "boat") or phonological contrasts (e.g., /b/ vs /p/) spoken by one or many talkers (e.g., Mullennix and Pisoni, 1990; Green, Tomiak, and Kuhl, 1997; Choi *et al.*, 2018; Kapadia and Perrachione, 2020; cf. Morton *et al.*, 2015; Magnuson *et al.*, 2021; Perrachione *et al.*, 2016). Such tasks may in principle be accomplished by prioritizing lower-level acoustic analyses and, as such, may not reveal as much about the consequences of talker variability on speech recognition as they do for merely speech perception (cf. Hickok and Poeppel, 2007). For instance, prior work has suggested that speech processing effects observed in small, closed-set tasks are not always found in larger or open-set tasks, and vice versa (Sommers, Kirk, and Pisoni, 1997; Clopper, Pisoni, and Tierney, 2006). While talker variability effects are also routinely seen in open-set speech recognition experiments (e.g., Perrachione *et al.*, 2016; Magnuson *et al.*, 2021; Saltzman *et al.*, 2021; Sommers *et al.*, 1997), much of the recent theoretical work on processing talker variability has focused on conclusions from two-alternative forced choice tasks. To critically reconsider the conclusions of such work (some of it our own), in the present manuscript, we further explore the question of whether manipulation of within- or between-talker variability has an effect on speech processing when listeners must decide between a larger number of phonological contrasts.

In this paper, we reconsider the classic idea–whether stated or assumed–in speech perception research that there is something privileged about the acoustic–phonetic variability in speech that arises due to differences among talkers. We adapted a speeded word identification paradigm that has seen extensive use for characterizing talker-variability effects on speech processing efficiency (e.g., Choi *et al.*, 2018, 2022; Choi and Perrachione 2019; Kapadia and Perrachione, 2020; Stilp and Theodore, 2020) to examine whether other sources of acoustic variability in speech incur similar costs, and how the processing demands of these different sources of variability interact. Specifically, we investigated how speech processing efficiency is affected by not only two levels of *between-talker variability* (single vs multiple talkers), but also two levels of *within-talker variability* (single vs multiple acoustically distinct exemplars of the target words from each talker), as well as two levels of variability that affect the degrees of freedom of the word identification decision that listeners must make on each trial (a two-word choice vs a six-word choice). Finally, we also revisit the question of whether the additional processing costs incurred by stimulus variability depend on the inherent degree of potential confusability of the target contrasts (Choi *et al*, 2018; Stilp and Theodore, 2020; Sommers and Barcroft, 2006), and, if so, how this varies with respect to whether that variability comes from between- or within-talker acoustic variation. By investigating whether these sources of acoustic variability in speech have differential effects on listeners' speech processing efficiency, accuracy, and response times, we will be able to better understand whether there are unique cognitive operations that specifically accommodate between-talker variability in speech.

II. METHODS

A. Participants

Native speakers of American English ($N = 24$; 18 female, 6 male; age 18–24, mean = 20.0 years) completed this study. All participants had a self-reported history free from speech, language, or hearing disorders and no familiarity with the talkers used in the experiment. Participants provided informed written consent, approved and overseen by the Institutional Review Board at Boston University.

B. Stimuli

Stimuli consisted of six minimally contrastive monosyllabic words of the form /bVt/. These words all shared the same onset and coda phonemes and differed only by their medial vowel. The set of vowels included /ɪ/, /ɛ/, /æ/, /ʌ/, /o/,

/u/, corresponding to the English words "bit," "bet," "bat," "but," "boat," and "boot." The use of multiple words allowed us to manipulate both the number of possible target words for listeners to identify (*target-word variability*), as well as the degree of potential distinctiveness of the target words, either within or between talkers (*phonological-contrast similarity*).

Acoustic differences due to *between-talker variability* were introduced into the stimuli by obtaining recordings from two male and two female native speakers of American English. Between-talker variability may introduce processing costs because of potential acoustic–phonemic ambiguity across talkers (e.g., one talker's [o] may be acoustically similar to another talker's [u]) (Hillenbrand *et al.*, 1995; Choi *et al.*, 2018). The degree of between-talker acoustic variability on the principal vowel acoustic dimensions (F1 and F2) can be seen by comparing the top row (panels A–D) to the bottom row (panels E–H) in Fig. 1.

Acoustic differences due to *within-talker variability* were introduced into the stimuli by prompting speakers to produce each word with combinations of (*i*) low, medium, and high pitch (within the speakers' natural pitch range) and (*ii*) shorter and longer durations, as well as with rising or falling intonation. These eight variations (3 pitches $\times$ 2 durations + 2 contours) for each of the six words from each of the four talkers made up the final 192-stimulus corpus. The degree of within-talker variability on the principal vowel acoustic dimensions (F1 and F2) can be seen by comparing the right (panels C, D, G, and H) vs the left (panels A, B, E, and F) of Fig. 1. While there is some variation from recording to recording, individual talkers tended to be largely internally consistent in their vowel acoustics, especially insofar as they did not overlap with adjacent categories. This is consistent with the observation that speakers tend to be highly consistent in the acoustic realization of their vowels over time (Heald and Nusbaum, 2014). This also contrasts with the realization of vowels in the mixed-talker condition, where the acoustics of different talkers' categories were more likely to overlap. Instead, a major source of phonetic variability in the within-talker condition was differences in voice pitch, which affect the realization of vowels by increasing or decreasing the harmonic composition of the formants. By explicitly instructing speakers to produce the target words with different vocal pitch, we likely introduced greater moment-to-moment variation in within-talker voice pitch than is present in ecological speaking/listening conditions (e.g., Van Stan *et al.*, 2015; Lee and Kreiman, 2022).

Natural speech samples were digitally recorded in quiet in a sound-attenuated booth using a Shure MX153 microphone (Niles, IL) and Roland Quad Capture (Los Angeles, CA) sound card sampling at 44.1 kHz and 16 bits. Stimuli were normalized for root mean square (RMS) amplitude to 65 dB SPL using Praat (Boersma, 2001), as amplitude variation has previously been shown not to interfere with speech processing response time in lexical decision tasks (Bradlow, Nygaard, and Pisoni, 1999; Sommers *et al.*, 1994; Nygaard *et al.*, 1995); however, stimuli nonetheless retained the

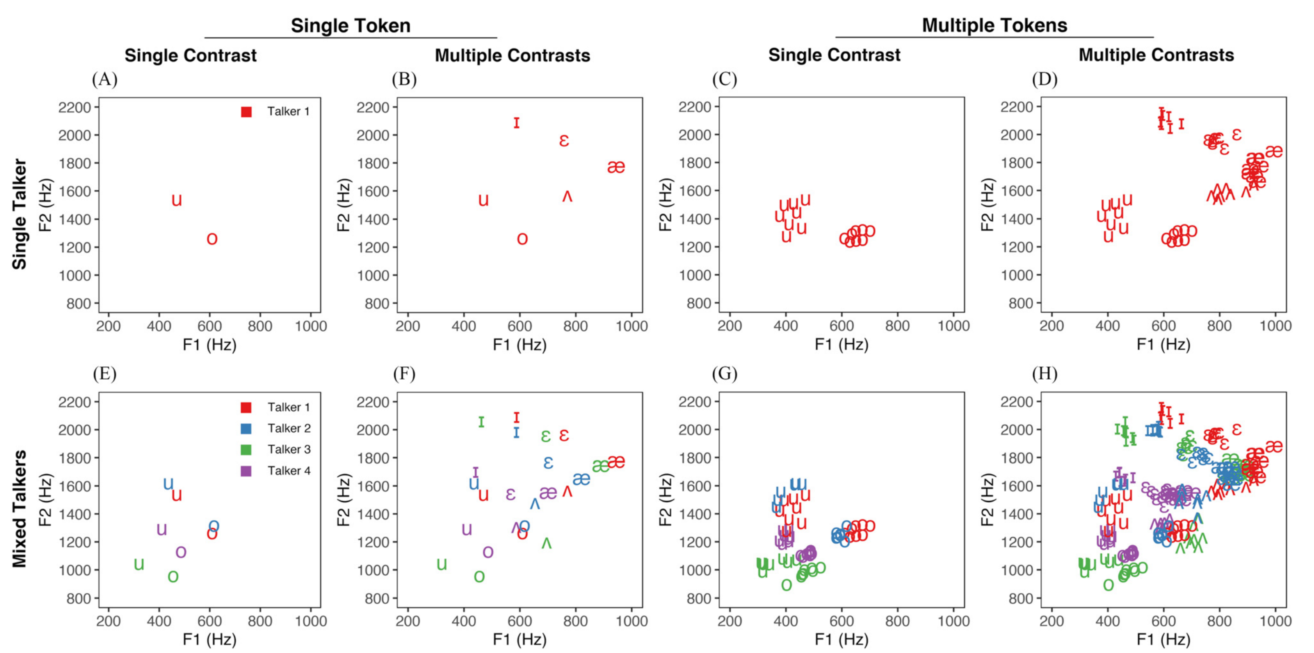

FIG. 1. (Color online) Manipulation of sources of acoustic–phonetic variability by condition. Vowel space plots show the degree of variability in the principal acoustic dimensions (F1, F2) of the vowels in the target words "bat," "bet," "bit," "boat," "boot," and "but." Colors correspond to individual talkers. Phonemic symbols correspond to the vowel category/word. Each panel depicts one of the task conditions: (A) single-talker, one-exemplar, two-word choice; (B) single-talker, one-exemplar, six-word choice; (C) single-talker, multiple-exemplars two-word choice; (D) single-talker, multiple-exemplars, six-word choice; (E) multiple-talker, one-exemplar, two-word choice; (F) multiple-talker, one-exemplar, six-word choice; (G) multiple-talker, multiple-exemplars, two-word choice; (H) multiple-talker, multiple-exemplars, six-word choice. For the two-word choice conditions, "boot" /u/ and "boat" /o/ are shown here, but each participant heard all 15 possible pairwise phonemic contrasts in separate blocks.

natural between- and within-talker variation in their amplitude envelopes. The natural variation in length across talkers and tokens was also preserved in the recordings, as phonetic variability due to speech rate has also been shown to incur processing costs (Green *et al.*, 1997; Bradlow *et al.*, 1999; Sommers and Barcroft, 2006). (To account for potential effects of stimulus duration on response time in the experiment, each *stimulus* was modelled as a random factor in our linear mixed-effects model; see details below.)

C. Procedure

The experiment consisted of a $2 \times 2 \times 2$ factorial design, through which we manipulated *between-talker variability*, *within-talker variability*, and *target-word variability*. *Between-talker variability* was operationalized as the number of talkers whose speech was heard during one condition of the experiment, with two levels: low variability (a single talker) and high variability (all four talkers). *Within-talker variability* was operationalized as the number of distinct recordings of each target word produced by each talker in a condition, with two levels: low variability (one exemplar per word per talker) and high variability (eight exemplars per word per talker). *Target-word variability* was operationalized as the number of phonemic contrasts, i.e., the number of possible target words in each condition, with two levels: two-word choice (one phonological contrast) and six-word choice [multiple (15) phonological contrasts]. Across all levels of all factors, there were eight unique conditions (Table I). The order of these conditions was counterbalanced across participants using Latin square permutations.

To measure how these three factors affected speech processing efficiency, we asked participants to perform a speeded word identification task in each of the eight conditions above. Participants were seated in a sound-attenuated booth. Stimulus delivery was controlled using PsychoPy2 (v1.83.03) (Peirce, 2007) with presentation via Sennheiser HD-380 Pro headphones (Old Lyme, CT). Participants heard words presented one at a time, and indicated which word they had heard by selecting the appropriate target from an on-screen array using a mouse. Participants were instructed to choose the target word as quickly as possible.

TABLE I. Experimental conditions with levels of independent variables. Each factor has a low value and a high value. Two-word choice conditions have 15 blocks to permit within-subjects comparison of words between the two- and six-word choice conditions, and to allow us to investigate effects of phonological similarity on the degree of interference from between- and within-talker variability. Each condition has the same number of trials.

Condition	Talkers	Exemplars	Words	Trials
A	1	1	2	240
B	1	1	6	240
C	1	8	2	240
D	1	8	6	240
E	4	1	2	240
F	4	1	6	240
G	4	8	2	240
H	4	8	6	240

Stimuli were presented in eight blocks of 240 trials each. Response options were displayed on a screen, with each printed word placed in a circle around a central point. The position of each target word on the screen was fixed throughout the experiment to reduce response complexity. The location of target words on the screen was randomized across participants. For the multiple phonemic-contrast conditions, all six target words were displayed; for the one-contrast conditions, only the two relevant options were visible (Fig. 2). Trials were presented at a rate of one per 2000 ms. The cursor position was reset to the center of the screen at the start of every trial, to ensure equal distance to each target. To become familiar with the paradigm and response demands (including position of the target words on the screen), participants first completed 60 practice trials analogous to condition B (single-talker, one-exemplar, six-word choices). The practice stimuli were spoken by a different talker than those in the rest of the experiment.

In the single-talker conditions (A, B, C, D), the talker was consistent across all trials, and the particular talker used in these conditions was counterbalanced across participants. In the multiple-talker conditions (E, F, G, H), recordings from all four talkers were presented with equal frequency within each condition; the presentation order of stimuli was pseudorandomized to ensure that speech from the same talker was never presented on adjacent trials, because even unexpected talker continuity can improve speech processing efficiency (Kapadia and Perrachione, 2020; Carter *et al.*, 2019).

During the two-word choice conditions (A, C, E, and G), participants decided which of two possible words they heard on each trial. Word-pair combinations were blocked within participants so each participant responded to all 15 possible two-word combinations during each condition. The order of these word-pair combination blocks was randomized across participants. During the six-word choice conditions (B, D, F, and H), participants decided which of all six possible words they heard on each trial.

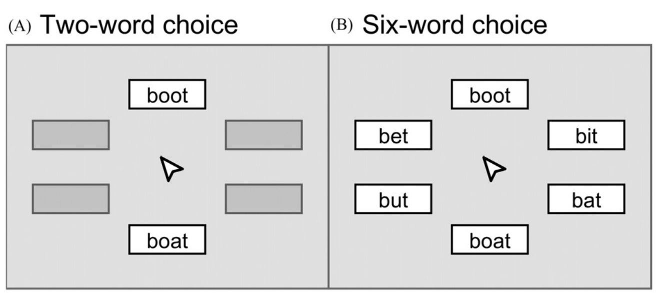

FIG. 2. Task interface. Participants indicated the word they heard on every trial by selecting it with a mouse cursor. (A) On two-word choice trials, only the two possible responses were indicated. A block where the choices were "boot" and "boat" is shown, but participants heard all 15 possible word pairs, blocked by pair, with the possible responses indicated as appropriate. (B) On six-word choice trials all possible responses were available. The cursor was automatically centered at the start of every trial, equidistant from the response targets.

During the low within-talker variability conditions (A, B, E, and F), participants heard only one recording per word per talker, whereas during the high within-talker variability conditions (C, D, G, and H), participants heard all eight exemplar recordings of each word by each talker, which were presented in random order subject to the constraints above. Within each condition, participants heard each target word, talker (in the case of mixed-talker blocks), and within-talker variant (combination of word duration, pitch height, and pitch contour) an equal number of times; however, some individual recordings were heard more or less often in conditions G and H to preserve the balance of other factors and keep the number of trials constant across all conditions.

D. Data analysis

Accuracy and response time were recorded for each trial. Accuracy was calculated as the proportion of correct trials out of total trials in each condition. Response time was measured in milliseconds from the onset of each stimulus. Incorrect trials, as well as trials with response times faster or slower than three standard deviations from the participant's mean in that condition, were excluded from analysis of response time (2.68% of all trials). For statistical analysis, response times were log-transformed to improve normality, as expected by the linear models.

Analyses were conducted in R using (generalized) linear mixed-effects models implemented in the packages *lme4* (v1.1.6) and *lmerTest*. The significance of fixed effects terms was determined by applying the relevant contrast coding scheme, with criterion $\alpha = 0.05$ and *p*-values for model terms based on the Satterthwaite approximation of the degrees of freedom. Where appropriate, *post hoc* pairwise comparisons between levels of the fixed factors were conducted using *difflsmeans*, and significance of multiple comparisons was corrected by controlling the family-wise error rate using the Holm–Bonferroni method.

III. RESULTS

A. Efficiency

Because of classic speed-accuracy tradeoffs (Green and Luce, 1973; Heitz, 2014), the aggregate processing costs associated with stimulus variability can be operationalized as differences in a metric called *efficiency* (Townsend and Ashby, 1978). As in previous work, we calculated efficiency as the quotient of mean accuracy and mean response time per participant per condition (Lim *et al*., 2019; Kapadia and Perrachione, 2020).

Broadly, as the amount of variability increased via any of the independent variables, speech processing efficiency decreased (Fig. 3). Word identification was most efficient when participants chose between two possible words, with only one exemplar of each word, spoken by a single talker; and it was least efficient when participants chose among six words, spoken by multiple talkers, who each produced multiple exemplars of each target word.

We analyzed the effects of the independent variables on participants' word identification efficiency using a linear mixed-effects model. The model's fixed-effects terms included categorical factors for *between-talker variability* (single-talker vs multiple-talkers), *within-talker variability* (one-exemplar vs multiple-exemplars), *target word variability* (two-word choice vs six-word choice), and all two- and three-way interactions. Sum (deviation) coded contrasts were applied to all categorical terms. Because efficiency is calculated as a summary statistic over all trials, resulting in one value per participant per condition, the maximal random-effects structure could include only by-participant intercepts. The form of the model of efficiency (in R notation) was

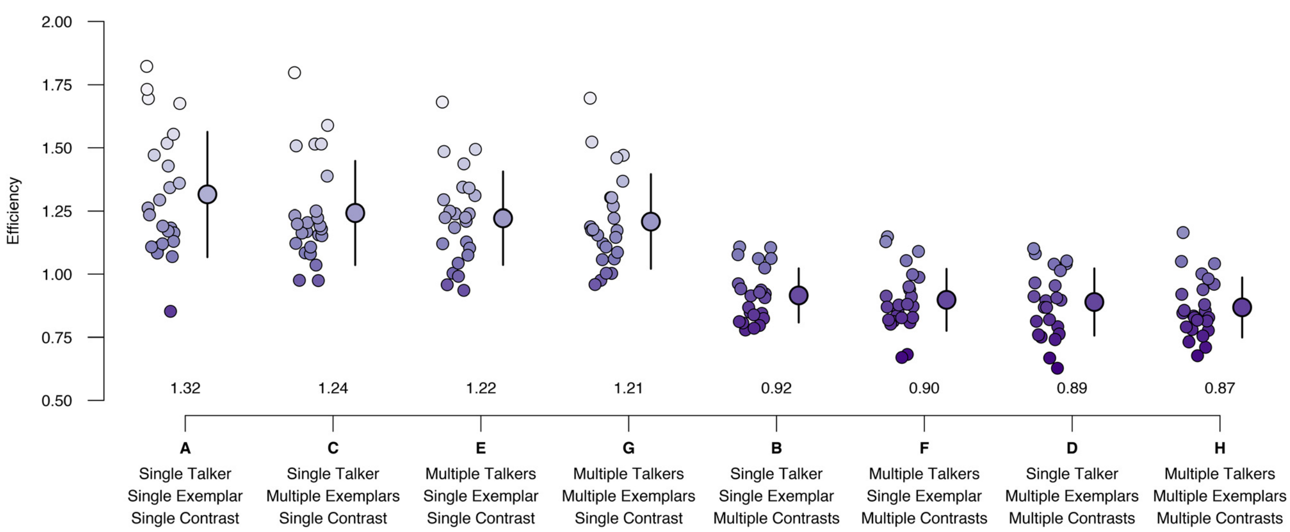

FIG. 3. (Color online) Word identification efficiency in each condition across participants, ordered by group mean. Efficiency was higher in conditions with two-word choices vs six-word choices, with single talkers vs multiple talkers. and with low vs high within-talker variability; but these factors did not interact. Small points show mean efficiency per condition per participant. Large points with error bars show group mean ± standard deviation, with the group mean value per condition reported above the abscissa. Shading denotes efficiency on a linear scale from most efficient (light) to least efficient (dark).

TABLE II. Efficiency model. *Significant after Holm–Bonferroni correction, $\alpha = 0.05$.

Effects	β	s.e.	df	t	p
Between-talker variability (multiple vs single talkers)	0.021	0.006	161	3.253	0.001*
Within-talker variability (multiple vs single exemplars)	0.018	0.006	161	2.783	0.006*
Target word variability (six- vs two-word choices)	0.177	0.006	161	27.711	$\ll$ 0.0001*
Between-talker × Within-talker variability	0.007	0.006	161	1.107	0.270
Between-talker × Target word variability	0.011	0.006	161	1.741	0.084
Within-talker × Target word variability	0.004	0.006	161	0.596	0.552
Between-talker × Within-talker × Target word variability	0.008	0.006	161	1.251	0.213

$$efficiency \sim talkers * exemplars * words + (1|participant).$$

Efficiency was significantly reduced by every source of variability: both between- and within-talker variability, as well as the number of possible target words (Table II). However, there were no significant two- or three-way interactions between these factors, suggesting they had independent and additive effects on speech processing efficiency. A marginal interaction betwixt *between-talker variability* and *target-word variability* suggested that the processing costs of talker variability may be comparatively smaller when the decision space is larger.

The observed differences in efficiency may have arisen due to differences in accuracy, response time, or both. Furthermore, the various sources of stimulus variability may have differential impacts on speech processing speed vs decision outcomes. To disentangle the consequences of the three sources of variability on listeners' decision outcomes (accuracy) vs processing speed (response time), we next consider the effects of these factors on each of the dependent variables separately.

B. Accuracy

Overall, participants' accuracy was very high, approaching ceiling performance (Fig. 4). Word identification was most accurate in conditions where the amount of stimulus variability was minimal, and fell modestly as the amount of variability increased, particularly as the number of possible target words increased.

We analyzed whether the three sources of variability affected word identification accuracy on each trial using a generalized linear mixed-effects model for binomial data (correct = 1, incorrect = 0). The model's fixed-effects terms included categorical factors for *between-talker variability* (single-talker vs multiple-talkers), *within-talker variability* (one-exemplar vs multiple-exemplars), *word-choice variability* (two-word choice vs six-word choice), and all two- and three-way interactions. Sum (deviation) coded contrasts were applied to all categorical terms. The model's random-effects terms included by-participant slopes for all fixed-effects terms, by-participant intercepts, and by-stimulus (item) intercepts. The overall model form (in R notation) was

$$accuracy \sim talkers * exemplars * words + (1 + talkers * exemplars * words|participant) + (1|stimulus).$$

Word identification was significantly less accurate when words were spoken by multiple talkers compared to a single talker (Table III). However, within-talker variability

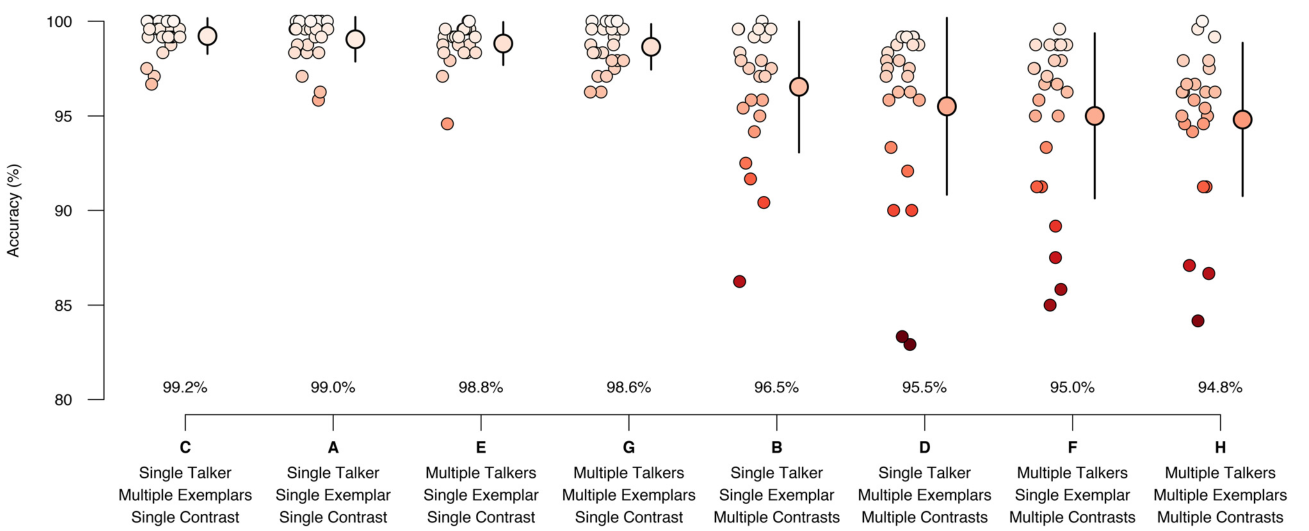

FIG. 4. (Color online) Accuracy in each condition across participants, ordered by group mean. Accuracy was higher in conditions with two-word choices vs six-word choices and in conditions with single talkers compared to conditions with multiple talkers. Small points show mean accuracy per condition per participant. Large points with error bars show group mean ± standard deviation, with the group mean value per condition reported above the abscissa. Shading denotes accuracy on a linear scale from highest (light) to lowest (dark).

TABLE III. Accuracy model. *Significant after Holm–Bonferroni correction, $\alpha = 0.05$.

Effects	β	s.e.	z	p
Between-talker variability (multiple vs single talkers)	0.177	0.088	2.007	0.045*
Within-talker variability (multiple vs single exemplars)	0.021	0.070	0.300	0.764
Target word variability (six- vs two-word choices)	0.789	0.079	9.964	$\ll$ 0.001*
Between-talker × Within-talker variability	−0.015	0.063	−0.241	0.810
Between-talker × Target word variability	0.031	0.049	0.620	0.535
Within-talker × Target word variability	−0.053	0.065	−0.882	0.411
Between-talker × Within-talker × Target word variability	−0.049	0.061	−0.799	0.424

did not affect word identification accuracy. Accuracy was also significantly lower when listeners had to decide between six possible targets compared to just two. None of the interaction terms was significant, suggesting that between-talker and word-choice variability had independent and additive effects on accuracy, which were neither moderated nor compounded by the additional presence of within-talker variability.

C. Response time

Participants' time to identify the target was, generally speaking, more susceptible to the different sources of stimulus variability than their aggregate efficiency or accuracy. As the amount of variability increased between conditions, participants' response times tended to slow (Fig. 5). Response times were fastest when trial-by-trial stimulus variability was minimal (the single-talker, single-exemplar, two-word choice condition) and slowest when trial-by-trial stimulus variability was maximal (the multiple-talker, multiple-exemplar, six-word choice condition).

We analyzed whether the independent variables affected word identification response time on each correct trial using a linear mixed-effects model with the same structure as that for the accuracy data (above). The form of the model of response time (in R notation) was

$$\log_{10}(RT) \sim talkers * exemplars * words + (1 + talkers * exemplars * words|participant) + (1|stimulus).$$

This model revealed significant main effects of all three factors (Table IV). Response time slowed with the introduction of any source of variability, whether between-talker, within-talker, or due to more potential target words. These factors also had a complicated pattern of interaction on participants' response time: There was a significant interaction betwixt between-talker and within-talker variability, as well as between-talker and target-word variability. Although there was no two-way interaction between within-talker variability and target-word variability, there was a significant three-way interaction between all of these sources of variability, suggesting that the presence of multiple forms of variability had either mediating or compounding effects on listeners' processing time during the task.

To unpack these interactions, we performed a series of pairwise comparisons to understand how changing the amount of one kind of stimulus variability (e.g., between-

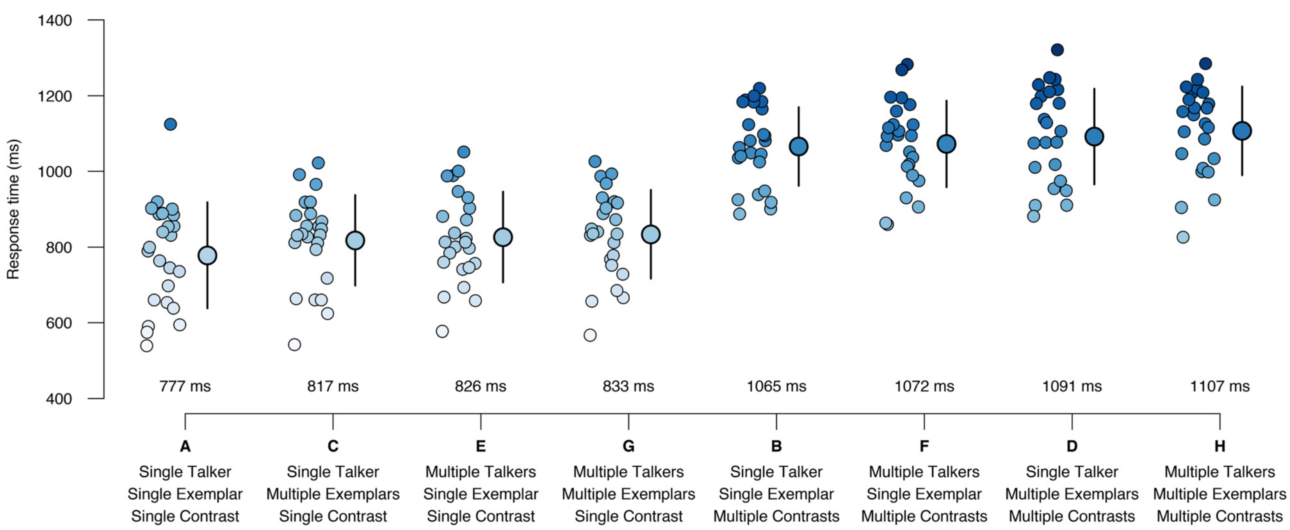

FIG. 5. (Color online) Response time in each condition across participants, ordered by group mean. Response times were fastest in condition A (when participants chose between two words spoken by a single talker, with one recording per word) and became slower as more variability was added, whether from additional within-talker variability, between-talker variability, or more word choices. The condition with the greatest amount of variability (H) was slowest. Small points show mean response time per condition per participant. Large points with error bars show group mean ± standard deviation, with the group mean value per condition reported above the abscissa. Shading denotes response time on a linear scale from fastest (light) to slowest (dark).

https://doi.org/10.1121/10.0016611

TABLE IV. Response time model. *Significant after Holm–Bonferroni correction, $\alpha = 0.05$.

Effects	β	s.e.	df	t	p
Between-talker variability (mixed vs single talkers)	−0.008	0.003	23.616	−3.151	<0.005*
Within-talker variability (multiple vs single exemplars)	−0.006	0.001	32.449	−4.564	≪0.001*
Target word variability (six- vs two-word choices)	−0.061	0.003	23.491	−21.010	≪0.001*
Between-talker × Within-talker variability	−0.004	0.002	24.555	−2.147	0.042*
Between-talker × Target word variability	−0.005	0.002	24.263	−2.866	<0.009*
Within-talker × Target word variability	0.002	0.002	24.590	0.996	0.329
Between-talker × Within-talker × Target word variability	−0.005	0.001	27.243	−4.646	≪0.001*

talker variability) affected response time while holding the other sources of variability (e.g., within-talker and word-choice variability) constant (Table V).

Response times were significantly slower when listening to multiple vs single talkers when there was only one exemplar and one phonological contrast (conditions E vs A). However, when any other source of variability was present (multiple exemplars, or multiple possible contrasts), introducing additional variability from multiple talkers did not further slow response times vs the corresponding single-talker condition (conditions G vs C, F vs B, and H vs D).

Response times were also significantly slower when listening to multiple vs one exemplar per talker in the absence of other sources of variability (conditions C vs A), revealing that within-talker phonetic variability alone has a significantly detrimental effect on speech processing. Introducing within-talker variability did not further slow the time to decide between two words when there was already variability due to multiple talkers (conditions F vs E), but, interestingly, did further slow response times in all conditions with multiple target words (conditions D vs B and H vs F).

Finally, response times were always significantly affected by the number of target words listeners had to consider during the trial. Regardless of other sources of variability, selecting a response during six-word choice conditions was always significantly slower than during two-word choice conditions.

D. Phonological contrast effects

We next considered whether the degree of acoustic–phonetic similarity of the target phonological contrast affected listeners' speech processing efficiency, and whether this was mediated by the presence of between- or within-talker variability. Prior work has suggested that phonological contrasts with greater acoustic similarity are processed more slowly, and that between-talker variability has an even larger effect on response times for proximal contrasts, likely due to the greater possibility of acoustic–phonetic overlap in these categories across talkers. However, it is important to note that between-talker variability has a deleterious effect on response time even for phonological contrasts that are acoustically unambiguous across talkers, such as /i/ vs /o/ (Choi *et al.*, 2018).

Here, we aimed to replicate that result using a wider range of phonological contrasts, as well as to examine whether this effect is similarly susceptible to the presence of within-talker variability. We operationalized phonological contrast dissimilarity as the distance between the centroid of two vowel categories in F1 × F2 space. We hypothesized that greater distance between target categories would result in faster response times. Having seen that within-talker variability also affects response time for single contrasts (conditions C vs A), we also tested whether this effect would be susceptible to the degree of phonological contrast dissimilarity.

TABLE V. Pairwise effects of high vs low variability for each source (between-talkers, within-talkers, or phonetic contrasts). For condition labels (A–H), refer to Table I and Fig. 1. *Significant after Holm–Bonferroni correction, $\alpha = 0.05$.

Variability effects	Conditions	Δ RT (ms)	Interference (%)	β	s.e.	t	df	p
Between-talker variability (multiple vs single talker)								
One exemplar and one contrast	E > A	48.3	6.21	0.043	0.012	3.816	24.9	< 0.001*
Multiple exemplars and one contrast	G > C	16.0	1.95	0.010	0.006	1.519	23.8	0.142
One exemplar and multiple contrasts	F > B	6.8	0.64	0.003	0.004	0.760	22.6	0.455
Multiple exemplars and multiple contrasts	H > D	15.1	1.38	0.008	0.006	1.399	23.1	0.175
Within-talker variability (multiple vs one exemplar)								
Single talker and one contrast	C > A	39.7	5.10	0.027	0.009	3.128	24.4	< 0.005*
Multiple talkers and one contrast	G > E	7.3	0.89	−0.008	0.005	−1.464	28.7	0.154
Single talker and multiple contrasts	D > B	26.2s	2.46	0.014	0.005	2.985	28.7	< 0.006*
Multiple talkers and multiple contrasts	H > F	34.5	3.22	0.019	0.005	4.197	29.2	< 0.001*
Word-choice variability (multiple vs one phonemic contrast)								
Single talker and one exemplar	B > A	287.7	36.98	0.140	0.011	12.31	23.0	≪ 0.001*
Multiple talkers and one exemplar	F > E	246.2	29.79	0.099	0.007	15.09	29.5	≪ 0.001*
Single talker and multiple exemplars	D > C	274.2	33.54	0.130	0.007	18.29	23.0	≪ 0.001*
Multiple talkers and multiple exemplars	H > G	273.3	32.79	0.130	0.006	21.57	24.1	≪ 0.001*

For each target word, we took the position of its vowel in F1 × F2 space from the measurements made by Hillenbrand *et al*. (1995). We then calculated the Euclidean distances between all pairs of vowels (in log Hertz). In operationalizing the acoustic similarity of vowel category pairs, we chose to use the values from data in Hillenbrand *et al*., rather than stimuli from the present experiment, because those represent acoustic averages based on a much larger and more balanced sample of speakers. Therefore, those measurements should be more representative of our listeners' lifetime experience with vowel productions from diverse talkers, which presumably guided their behavior during the present experiment.

First, to determine whether the processing demands of between-talker variability scaled as a function of the similarity of the target phonological contrast, we submitted participants' response times from the two-word choice, low within-talker variability conditions (A, E) to a linear mixed-effects model with a categorical fixed factor of *between-talker variability* (single vs multiple talkers), a continuous fixed factor of *phonological contrast dissimilarity* (as above), and their interaction. The model's random-effects terms included by-participant slopes for all fixed factors, by-participant intercepts, and by-stimulus intercepts.

Contrasts on the fixed factors demonstrated the expected, significant effect of *between-talker variability* ($\beta = -0.036$, $s.e. = 0.012$, $df = 24.078$, $t = -3.094$, $p < 0.005$), such that response times were faster for single than mixed talkers. The effect of *phonological contrast dissimilarity* was also significant ($\beta = -0.021$, $s.e. = 0.003$, $df = 23.214$, $t = -6.319$, $p \ll 0.001$), such that response times were faster for more acoustically dissimilar contrasts (e.g., /æ/ vs /o/) and slower for more acoustically similar ones (e.g., /ɛ/ vs /æ/). Furthermore, there was a significant interaction between these terms ($\beta = 0.008$, $s.e. = 0.004$, $df = 23.223$, $t = 2.084$, $p < 0.05$), such that the between-talker variability effect was larger for acoustically similar contrasts and smaller for acoustically dissimilar contrasts [Fig. 6(A)].

Second, to determine whether the processing demands of within-talker variability also scaled as a function of the similarity of the target phonological contrast, we submitted participants' response times from the two-word choice, low between-talker variability conditions (A, C) to a linear mixed-effects model with a categorical fixed factor of *within-talker variability* (single vs multiple exemplars), a continuous fixed factor of *phonological contrast dissimilarity* (as above), and their interaction. The model's random effects terms were as above.

Contrasts on the fixed factors demonstrated the expected, significant effect of *within-talker variability* ($\beta = -0.033$, $s.e. = 0.013$, $df = 23.394$, $t = -2.486$, $p < 0.03$), such that response times were slower when listeners were hearing multiple exemplars of the target word. The effect of *phonological contrast dissimilarity* was also significant ($\beta = -0.020$, $s.e. = 0.004$, $df = 23.141$, $t = -4.615$, $p < 0.001$), such that response times were faster for more acoustically dissimilar contrasts and slower for more acoustically similar ones. However, the interaction between these terms was not significant

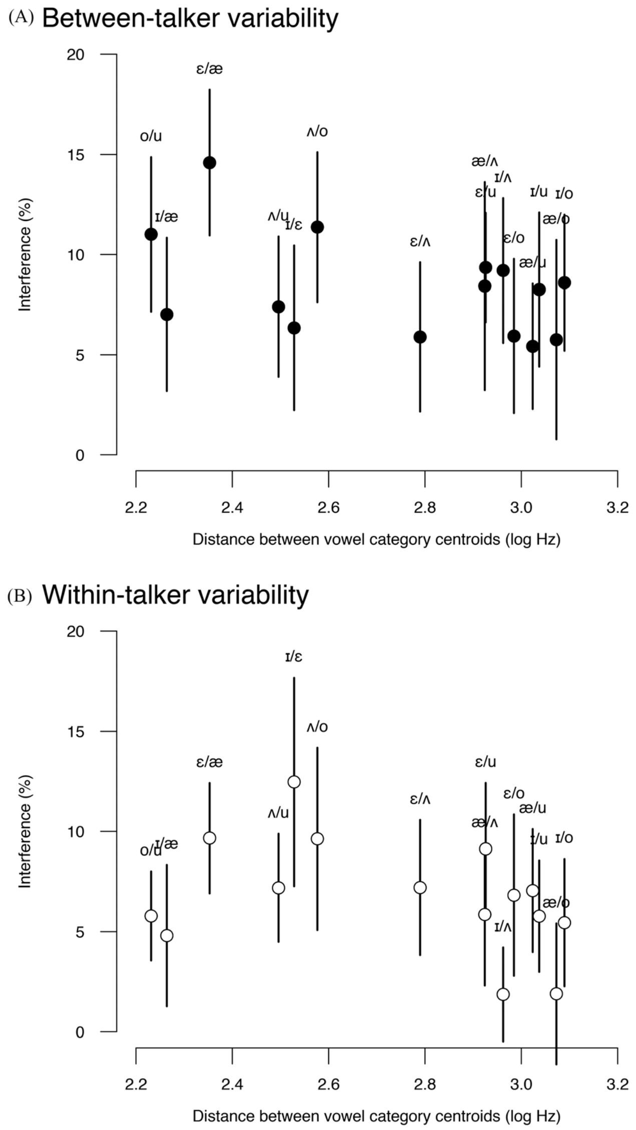

FIG. 6. The interference effect of between- and within-talker variability as a function of phonological contrast dissimilarity. To illustrate the proportional differences between response times for each vowel contrast in the high- vs low-variability levels of each factor, the mean response time data are collapsed into an "interference effect" of variability: (high-variability RT – low-variability RT)/low-variability RT (Choi *et al*., 2018). (A) The interference effect of between-talker variability was significantly greater for more similar phonological contrasts. **(B)** The interference effect of within-talker variability was not significantly related to the acoustic similarity of the target contrast. Points show the mean, and error bars show the standard error of the mean, across participants.

($\beta = 0.007$, $s.e. = 0.004$, $df = 23.044$, $t = 1.646$, $p = 0.113$), such that the within-talker variability effect was not systematically affected by the similarity of the target phonological contrast [Fig. 6(B)].

IV. DISCUSSION

In this paper, we examined the consequences of three sources of variability on speech processing efficiency. We first examined their effects on overall efficiency as an aggregate measure of response accuracy and response speed, since speed-accuracy tradeoffs can obscure effects on one or the other measure (Green and Luce, 1973; Heitz, 2014; Townsend and Ashby, 1978). Overall efficiency was affected by all three sources of variability in the stimuli: both between- and within-talker variability, as well as the number of response options. When variability was introduced for each factor, speech processing efficiency declined. However, the factors did not interact. Looking at the constituent measures of accuracy and response time separately, this lack of interaction may have reflected differential speed-accuracy tradeoffs across the different sources of variability, as the pattern of interference effects on accuracy and response time independently was more nuanced.

With respect to word recognition accuracy, participants' performance was very high, and the deleterious effects of variability were minimal. This is perhaps unsurprising, since the target speech reflected sampling from a common population that should have been highly familiar to our listeners, even if they did not have prior exposure to these particular talkers or recordings. In this way, listeners could draw on their lifetime experience with similar speech to support their performance, consistent with our casual experience of finding speech perception trivially easy and mostly error-free in everyday listening settings. However, accuracy was nonetheless significantly impacted by two sources of variability: between-talker acoustic–phonetic variation and word-choice variability. When listeners had to identify words spoken in contexts with multiple talkers, they were less accurate than when identifying words spoken by a single talker. This replicates extensive prior research showing talker-variability effects on speech perception accuracy (e.g., Morton *et al.*, 2015). Moreover, when listeners had more possible response choices, they also made more word identification errors, although they maintained very high accuracy overall in even the most variable condition (95%). These target-variability effects parallel classical findings in both speech (Sommers *et al.*, 1997) and psychology more generally (e.g., Beck and Kastner, 2009) that competing choices are distracting and impair performance.

Interestingly, within-talker variability appeared to have no effect on listeners' word identification accuracy, regardless of the amount of variability in the other levels. Because participants were able to maintain a high degree of accuracy across the experiment, it suggests that the principal consequences of variability on speech processing efficiency arise due to differences in processing time, rather that decision outcomes (Choi and Perrachione, 2019; Kapadia and Perrachione, 2020). The lack of a within-talker-variability effect on accuracy must be considered within the scope of the stimuli used in this experiment, which reflect acoustic–phonemic distributions that were highly familiar for our listeners. In the case that, for instance, listeners' experiences with acoustic–phonemic mappings and the experimental dimensions of within-talker variability were misaligned–such as when listening to foreign-accented speech (Xie and Jaeger, 2020; Vaughn *et al.*, 2019)–it is straightforward to see how unfamiliar patterns of within-talker variability could undercut listeners' word identification accuracy.

Indeed, response times were the most susceptible to manipulations of the three sources of variability. Conditions where there was more between-talker variability, within-talker variability, and potential target word choices all had significantly slower processing time than the corresponding low-variability condition. Significant interactions between these factors indicated that these sources of variability also had a complex pattern of compounding or attenuating effects on listeners' response time. Notably, the introduction of within-talker variability increased response times in almost every case vs the analogous condition with only one token per talker per target (with the exception of conditions E vs G, where the number of potential target words was few and acoustic–phonetic variability was already present from multiple talkers).

Interestingly, the classic and widely replicated finding that response times are slower for speech from multiple talkers compared to speech from a single talker was only observed in the present study when the amount of variability in the other conditions was minimized (conditions E vs A). When variability from either of the other sources was present, adding multiple talkers no longer had significantly deleterious effects on processing speed. This is surprising considering prior work suggesting that between-talker variability is the largest potential source of variation in speech acoustics (Mullennix and Pisoni, 1990; Kleinschmidt, 2019). Alternatively, the costs of having to make decisions about speech from multiple talkers may instead be reflected in differences in listeners' accuracy. However, the finding that between-talker variability did not impose further processing costs on top of within-talker variability also challenges the idea that these sources of acoustic–phonetic variability are accommodated by dissociable underlying processes.

Notably, regardless of the amount of variability in the other factors, increasing participants' decision space from two words to six words significantly increased their response times. One interpretation of this effect is that adding more phonological contrasts increases the number of possible interpretations of the signal that a listener must consider, leading to more perceptual processing and longer response times. However, we believe it to be unlikely that increasing the number of possible perceptual interpretations of a given speech sample will continue to increase the demands on listeners' perceptual processing (Munroe, 2009). In real life, there are essentially infinitely many possible speech signals that listeners may hear, yet speech content is nonetheless recognized not only in finite time, but also impressively quickly. A more likely explanation for this stark difference in response time between the two- and six-word conditions is listeners' added uncertainty in indicating the correct response from the expanded on-screen array. Extensive work in psychology has shown how increasing the number of possible discrete responses results in increasing delay to indicate a response over and above additional perceptual processing demands (reviewed in Proctor and Schneider, 2018).

On the one hand, the present results suggest that adding between-talker variability does not significantly increase processing costs when there are more than two possible response choices. This potentially challenges theoretical conclusions as to the mechanisms for processing talker variability that have been derived from experiments involving only two-alternative forced choice paradigms (e.g., Choi and Perrachione, 2019; Choi *et al.*, 2022). On the other hand, studies using other paradigms that involve multiple or free responses have also shown effects of talker variability (Perrachione *et al.*, 2016; Sommers *et al.*, 1997; Magnuson *et al.*, 2021; see especially Saltzman *et al.*, 2021). As such, it is possible that the added uncertainty (and thus delay) of indicating the correct response introduced both a ceiling effect and an additional source of behavioral noise that obscured within- or between-talker variability effects in the six-word choice conditions. Ecological speech processing rarely involves deciding between just two (or six) possible responses over and over, raising an important challenge for researchers in this domain to develop novel tasks by which between- and within-talker variability effects can be measured in more ecologically realistic designs.

Finally, we replicated previous observations of phonological contrast dissimilarity on speech processing efficiency (Choi *et al.*, 2018; Sommers and Barcroft, 2006; cf. Stilp and Theodore, 2020). Word identification decisions for more similar (here, more acoustically proximal in $F1 \times F2$ space) vowel contrasts were made more slowly than for vowel contrasts that were more acoustically distinct. Furthermore, this effect interacted with the presence of between-talker variability, such that between-talker variability imposed greater relative processing costs when the phonological contrasts were more similar, and smaller relative processing costs when the contrasts were more distinct. This makes sense when considering how between-talker variability affects the principal phonetic dimensions of a target vowel contrast: Because different talkers have different vocal tract lengths, the absolute frequencies of their F1 and F2 resonances will differ. When the phonological contrasts are closer in acoustic space, there is greater likelihood for acoustic–phonetic mismatch between talkers; for example, the F2 in one talker's /o/ may be more similar to the F2 in another talker's /u/ than their /o/, leading to greater acoustic–phonemic ambiguity across talkers. Indeed, it has been suggested that the reason talker variability imposes processing costs, even for acoustically unambiguous tokens (Choi *et al.*, 2018), is because, ecologically, a situation with multiple talkers increases the likelihood that there *will* be ambiguity, which the speech processing system must be prepared to accommodate (Magnuson and Nusbaum, 2007; Magnuson *et al.*, 2021).

Interestingly, we did not observe a significant interaction between phonological contrast similarity and within-talker variability. That is, response times to the more acoustically similar phonological contrasts were not more affected by within-talker variability than the more acoustically distinct ones. On the one hand, this might suggest that there is something unique about between- vs within-talker variation, such that between-talker variation is more likely to result in acoustic–phonemic mismatches, which, in turn, disproportionately confounds the processing of acoustically similar phonological contrasts and results in an overall decrement in word identification accuracy, as noted above. However, this result must be interpreted with respect to both the *physical dimension* and *magnitude* of acoustic–phonetic variability introduced by the between- vs within-talker variability levels in the present study. Just like the direction of processing dependencies in classic Garner interference paradigms depends on the relative difficulty (or salience) of variation along either physical dimension (Cutler *et al.*, 2011; Huettel and Lockhead, 1999), so too must the effects of within- vs between-talker processing costs be considered with respect to how physically dissimilar stimuli become due to the variation those manipulations introduce. Looking at Fig. 1(H), it is clear that there is considerably more opportunity for acoustic–phonological mismatch *across* talkers than *within* a talker for the stimuli in the present experiment. That is, when encountering a new talker in a multiple-talker condition, there is greater likelihood for confusion between the newly heard acoustics and vowel categories, especially if a listener had anchored to a context based on the preceding token (Choi and Perrachione, 2019; Laing *et al.*, 2012; Stilp and Assgari, 2018; Morton *et al.*, 2015; Johnson, 1990). Given modern advances in speech stimulus resynthesis, it should be possible in future work to create conditions that parametrically vary the degree of acoustic–phonemic variability both within and between talkers. This would, in turn, allow us to better ascertain whether there is something inherently unique about these sources of variability vis-à-vis phonological contrastiveness, or whether these two sources only appear different because, in natural speech, they typically entail different magnitudes of variation along the phonologically relevant acoustic–phonetic dimensions (as in the present study).

The suggestion that the shared vs distinct effects of between- and within-talker variability on speech processing efficiency simply reflect the degree of acoustic variability underlying these distinctions parallels a larger question in the literature on speech variability: Namely, whether the effects of variation are categorical (i.e., all or nothing) or whether they are graded by the magnitude of variation. For instance, we previously showed that talker variability effects do not scale with the number of talkers (Kapadia and Perrachione, 2020), suggesting that the mere presence of variability, not its magnitude, is categorically deleterious. However, what if the amount of acoustic variability among talkers was less? Stilp and Theodore (2020) suggested that smaller between-talker differences in F0 could produce smaller aggregate talker variability effects. Further work should be done to systematically parameterize the degree of between- vs within-talker variability and understand its graded vs stepwise consequences on speech processing efficiency (e.g., Nusbaum and Magnuson, 1997; Magnuson *et al.*, 2021).

Ultimately, these results show that within-talker variability also imposes a cost on speech processing efficiency. What implications does this finding have for understanding ecological speech perception, which tends to feel effortless and only very rarely results in errors? In this experiment, we intentionally manipulated the degree of within-talker variability in ways that are somewhat unnatural. First, it is rare for listeners to encounter a series of discrete, disconnected words with random pitch height changes. In natural speech, within-talker acoustics usually change continuously from word to word, in a way that supports listeners' ability to discern the relevant phonetic contrasts for identifying words (Johnson, 1990; Choi and Perrachione, 2019). However, in natural settings, speakers also intentionally introduce larger magnitude (within-talker) variability to highlight or make salient particular linguistic content. That is, speakers may increase their pitch, intensity, or the duration of some target speech to highlight it in the discourse (Hirshberg and Pierrehumbert, 1986). In this way, changing speech acoustics to refocus listeners' attention and highlight communicatively relevant content reflects a functional purpose for within-talker variability: It does incur a processing cost for listeners when speakers *want* to ensure that their listeners more thoroughly encode particular linguistic information. It is notable that, in our present results, within-talker variability did not have an effect on accuracy, though it did impact processing time, consistent with the idea that phonologically unambiguous variation can highlight information without leading to errors on the part of listeners.

V. CONCLUSIONS

Taken together, these results suggest that multiple sources of acoustic variability impose both shared and unique processing costs on speech perception. Between-talker variability imposed costs on both word identification accuracy and processing time, which interacted with the degree of phonological contrastiveness. This pattern of results likely reflects the greater propensity for mismatch between acoustic encoding and phonological categories across talkers and the additional time required to adapt to or normalize these differences. Within-talker variability primarily imposed costs on processing time, but not accuracy, and did not interact with the degree of phonological contrastiveness, suggesting that within-category acoustic variability can demand additional cognitive effort on the part of listeners, but is not detrimental for comprehension. Finally, the complexity of the task's decision space has the largest effect on speech processing efficiency, which may primarily reflect response-selection rather than perceptual-processing demands, and which highlights the need for future work to examine talker-variability effects in more ecological listening scenarios and tasks.

ACKNOWLEDGEMENTS

We thank Sung-Joo Lim, Melanie Matthies, Yaminah Carter, Terri Scott, Ja Young Choi, Jayden Lee, Chinazo Otiono, Grace Mecha, Amabel Antwi, Kamilah Harruna, Rita Kou, and Michelle Njoroge. This work was supported by the National Institute on Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health under Grant Nos. R03 DC014045 (to T.K.P.), R01 DC004545 (to Gerald Kidd), and T32 DC013017 (to Christopher Moore).

[1]Although a recent attempt at replicating this result was not successful (Luthra *et al.*, 2021), it should be noted that in the replication attempt, the amount of stimulus variability failed to impose added processing costs in either the nominally within- or between-talker condition. This ultimately leaves it uncertain whether the same magnitude of acoustic variation would incur different processing costs if it were interpreted as within- vs between-talker variability.

Allen, J. S., Miller, J. L., and DeSteno, D. (**2003**). "Individual talker differences in voice-onset-time," J. Acoust. Soc. Am. **113**, 544–552.
Barcroft, J., and Sommers, M. S. (**2005**). "Effects of acoustic variability on second language vocabulary learning," Stud. Second Lang. Acquis. **27**, 387–414.
Beck, D. M., and Kastner, S. (**2009**). "Top-down and bottom-up mechanisms in biasing competition in the human brain," Vision Res. **49**, 1154–1165.
Boersma, P. (**2001**). "Praat, a system doing phonetics by computer," Glot Int. **5**, 341–345.
Bradlow, A. R., Nygaard, L. C., and Pisoni, D. B. (**1999**). "Effects of talker, rate, and amplitude variation on recognition memory for spoken words," Percept. Psychophys. **61**, 206–219.
Bradlow, A. R., Torretta, G. M., and Pisoni, D. B. (**1996**). "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," Speech Commun. **20**, 255–272.
Carter, Y. D., Lim, S.-J., and Perrachione, T. K. (**2019**). "Talker continuity facilitates speech processing independent of listeners' expectations," in *19th International Congress of Phonetic Sciences*, August 2019, Melbourne.
Choi, J. Y., Hu, E. R., and Perrachione, T. K. (**2018**). "Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing," Atten. Percept. Psychophys. **80**(3), 784–797.
Choi, J. Y., Kou, R. S. N., and Perrachione, T. K. (**2022**). "Distinct mechanisms for talker adaptation operate in parallel on different timescales," Psychon. Bull. Rev. **29**, 627–637.
Choi, J. Y., and Perrachione, T. K. (**2019**). "Time and information in perceptual adaptation to speech," Cognition **192**, 103982.
Clayards, M. (**2018**). "Individual talker and token covariation in the production of multiple cues to stop voicing," Phonetica **75**(1), 1–23.
Clopper, C. G., and Pisoni, D. B. (**2004**). "Effects of talker variability on perceptual learning of dialects," Lang. Speech **47**, 207–239.
Clopper, C. G., Pisoni, D. B., and Tierney, A. T. (**2006**). "Effects of open-set and closed-set task demands on spoken word recognition," J. Am. Acad. Audiol. **17**(5), 331–349.
Cutler, A., Andics, A., and Fang, Z. (**2011**). "Inter-dependent categorization of voices and segments," in *17th Meeting of the International Congress of Phonetic Sciences*, August 2011, Hong Kong.
Garner, W. R. (**1974**). *The Processing of Information and Structure* (Erlbaum, Potomac, MD).
Green, D. M., and Luce, D. (**1973**). "Speed-accuracy tradeoff in auditory detection," in *Attention and Performance IV*, edited by S. Kornblum (Academic Press, New York), pp. 547–569.
Green, K. P., Tomiak, G. R., and Kuhl, P. K. (**1997**). "The encoding of rate and talker information during phonetic perception," Percept. Psychophys. **59**(5), 675–692.
Heald, S., Klos, S., and Nusbaum, H. C. (**2016**). "Understanding speech in the context of variability," in *Neurobiology of Language*, edited by G. Hickok and S. Small (Academic Press, San Diego), pp. 195–208.
Heald, S. L. M., and Nusbaum, H. C. (**2014**). "Talker variability in audio-visual speech perception," Front. Psychol. **5**, 698.
Heitz, R. P. (**2014**). "The speed-accuracy tradeoff: History, physiology, methodology, and behavior," Front. Neurosci. **8**, 150.

Hickok, G., and Poeppel, D. (**2007**). "The cortical organization of speech processing," Nat. Rev. Neurosci. **8**, 393–402.
Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.
Hirshberg, J., and Pierrehumbert, J. (**1986**). "The intonational structuring of discourse," in *24th Annual Meeting of the Association for Computational Linguistics*, July 1986, New York, NY, pp. 136–144.
Holmes, E., Domingo, Y., and Johnsrude, I. (**2018**). "Familiar voices are more intelligible, even if they are not recognized as familiar," Psychol. Sci. **29**, 1575–1583.
Huettel, S. A., and Lockhead, G. R. (**1999**). "Range effects of an irrelevant dimension on classification," Percept. Psychophys. **61**, 1624–1645.
Johnson, K. (**1990**). "The role of perceived speaker identity in $F0$ normalization of vowels," J. Acoust. Soc. Am. **88**, 642–654.
Johnson, K. (**2005**). "Speaker normalization in speech perception," in *The Handbook of Speech Perception*, edited by D. B. Pisoni and R. E. Remez (Blackwell, Malden, MA), pp. 363–389.
Kapadia, A. M., and Perrachione, T. K. (**2020**). "Selecting among competing models of talker adaptation: Attention, cognition, and memory in speech processing efficiency," Cognition **204**, 104393.
Keerstock, S., and Smiljanic, R. (**2019**). "Clear speech improves listeners' recall," J. Acoust. Soc. Am. **146**, 4604–4610.
Kingston, J. (**2003**). "Learning foreign vowels," Lang. Speech **46**, 295–349.
Kleinschmidt, D. F. (**2019**). "Structure in talker variability: How much is there and how much can it help?," Cognition Neurosci. **34**(1), 43–68.
Kleinschmidt, D. F., and Jaeger, T. F. (**2015**). "Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel," Psychological Rev. **122**(2), 148–203.
Laing, E. J. C., Liu, R., Lotto, A. J., and Holt, L. L. (**2012**). "Tuned with a tune: Talker normalization via general auditory processes," Front. Psychol. **3**, 203.
Lavan, N., Burston, L. F. K., and Garrido, L. (**2019a**). "How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices," Br. J. Psychol. **110**, 576–593.
Lavan, N., Burton, A. M., Scott, S. K., and McGettigan, C. (**2019b**). "Flexible voices: Identity perception from variable vocal signals," Psychon. Bull. Rev. **26**, 90–102.
Lee, Y., Keating, P., and Kreiman, J. (**2019**). "Acoustic voice variation within and between speakers," J. Acoust. Soc. Am. **146**, 1568–1579.
Lee, Y., and Kreiman, J. (**2022**). "Acoustic voice variation in spontaneous speech," J. Acoust. Soc. Am. **151**, 3462–3472.
Lim, S.-J., Shinn-Cunningham, B. G., and Perrachione, T. K. (**2019**). "Effects of talker continuity and speech rate on auditory working memory," Atten. Percept. Psychophys. **81**, 1167–1177.
Luthra, S., Saltzman, D., Myers, E. B., and Magnuson, J. S. (**2021**). "Listener expectations and the perceptual accommodation of talker variability: A pre-registered replication," Atten. Percept. Psychophys. **83**, 2367–2376.
Magnuson, J. S., and Nusbaum, H. C. (**2007**). "Acoustic differences, listener expectations, and the perceptual accommodation of talker variability," J. Exp. Psychol. Hum. Percept. Perform. **33**, 391–409.
Magnuson, J. S., Nusbaum, H. C., Akahane-Yamada, R., and Saltzman, D. (**2021**). "Talker familiarity and the accommodation of talker variability," Atten. Percept. Psychophys. **83**, 1842–1860.
Morton, J. R., Sommers, M. S., and Lulich, S. M. (**2015**). "The effect of exposure to a single vowel on talker normalization for vowels," J. Acoust. Soc. Am. **137**, 1443–1451.
Mullennix, J. W., and Pisoni, D. B. (**1990**). "Stimulus variability and processing dependencies in speech perception," Percept. Psychophys. **47**, 379–390.
Mullennix, J. W., Pisoni, D. B., and Martin, C. S. (**1989**). "Some effects of talker variability on spoken word recognition," J. Acoust. Soc. Am. **85**(1), 365–378.
Munroe, R. (**2009**). "Extrapolating," https://xkcd.com/605/ (Last viewed November 14, 2022).
Nearey, T. M. (**1989**). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. **85**, 2088–2113.
Newman, R. S., Clouse, S. A., and Burnham, J. L. (**2001**). "The perceptual consequences of within-talker variability in fricative production," J. Acoust. Soc. Am. **109**(3), 1181–1196.
Nusbaum, H. C., and Magnuson, J. (**1997**). "Talker normalization: Phonetic constancy as a cognitive process," in *Talker Variability in Speech Processing*, edited by K. A. Johnson and J. W. Mullennix (Academic Press, San Diego), pp. 109–132.
Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1994**). "Speech perception as a talker-contingent process," Psychol. Sci. **5**(1), 42–46.
Nygaard, L. C., Sommers, M. S., and Pisoni, D. B. (**1995**). "Effects of stimulus variability on perception and representation of spoken words in memory," Percept. Psychophys. **57**, 989–1001.
Palmeri, T., Goldinger, S., and Pisoni, D. (**1993**). "Episodic encoding of voice attributes and recognition memory for spoken words," J. Exp. Psychol. **19**(2), 309–328.
Peirce, J. W. (**2007**). "PsychoPy: Psychophysics software in Python," J. Neurosci. Meth. **162**, 8–13.
Perrachione, T. K., Del Tufo, S. N., Winter, R., Murtagh, J., Cyr, A., Chang, P., Halverson, K., Ghosh, S. S., Christodoulou, J. A., and Gabrieli, J. D. E. (**2016**). "Dysfunction of rapid neural adaptation in dyslexia," Neuron **92**(6), 1383–1397.
Perrachione, T. K., Furbeck, K. T., and Thurston, E. J. (**2019**). "Acoustic and linguistic factors affecting perceptual similarity judgments of voices," J. Acoust. Soc. Am. **146**, 3384–3399.
Perrachione, T. K., Lee, J., Ha, L. Y. Y., and Wong, P. C. M. (**2011**). "Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design," J. Acoust. Soc. Am. **130**, 461–472.
Pierrehumbert, J. B. (**2003**). "Phonetic diversity, statistical learning, and acquisition of phonology," Lang. Speech **46**, 115–154.
Pisoni, D. B. (**1981**). "Some current theoretical issues in speech perception," Cognition **10**, 249–259.
Pisoni, D. B. (**1997**). "Some thoughts on 'normalization' in speech perception," in *Talker Variability in Speech Processing*, edited by K. Johnson and J. W. Mullenix (Academic Press, San Diego, CA), pp. 9–32.
Potter, R. K., and Steinberg, J. C. (**1950**). "Toward the specification of speech," J. Acoust. Soc. Am. **22**, 807–820.
Proctor, R. W., and Schneider, D. W. (**2018**). "Hick's law for choice reaction time: A review," Q. J. Exp. Psychol. **7**, 1281–1299.
Pufahl, A., and Samuel, A. G. (**2014**). "How lexical is the lexicon? Evidence for integrated auditory memory representations," Cogn. Psychol. **70**, 1–30.
Sadakata, M., and McQueen, J. M. (**2014**). "Individual aptitude in Mandarin lexical tone perception predicts effectiveness of high-variability training," Front. Psychol. **5**, 1318.
Saltzman, D., Luthra, S., Myers, E. B., and Magnuson, J. S. (**2021**). "Attention, task demands, and multitalker processing costs in speech perception," J. Exp. Psychol.: Hum. Percept. Perform. **47**(12), 1673–1680.
Scott, S. K. (**2019**). "From speech and talkers to the social world: The neural processing of human spoken language," Science **366**, 58–62.
Sjerps, M. J., Fox, N. P., Johnson, K., and Chang, E. F. (**2019**). "Speaker-normalized sound representations in the human auditory cortex," Nat. Commun. **10**, 2465.
Sjerps, M. J., McQueen, J. M., and Mitterer, H. (**2013**). "Evidence for precategorical extrinsic vowel normalization," Atten. Percept. Psychophys. **75**(3), 576–587.
Smiljanic, R., and Bradlow, A. R. (**2009**). "Speaking and hearing clearly: Talker and listener factors in speaking style changes," Lang. Linguist. Compass. **3**, 236–264.
Sommers, M. S., and Barcroft, J. (**2006**). "Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification," J. Acoust. Soc. Am. **119**(4), 2406–2416.
Sommers, M. S., Kirk, K. I., and Pisoni, D. B. (**1997**). "Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format," Ear Hear. **18**(2), 89–99.
Sommers, M. S., Nygaard, L. C., and Pisoni, D. B. (**1994**). "Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude," J. Acoust. Soc. Am. **96**, 1313–1324.
Souza, P., Gehani, N., Wright, R., and McCloy, D. (**2013**). "The advantage of knowing the talker," J. Am. Acad. Audiol. **24**(8), 689–700.

https://doi.org/10.1121/10.0016611

Stilp, C. E., and Assgari, A. A. (**2018**). "Perceptual sensitivity to spectral properties of earlier sounds during speech categorization," Atten. Percept. Psychophys. **80**, 1300–1310.

Stilp, C. E., and Theodore, R. M. (**2020**). "Talker normalization is mediated by structured indexical information," Atten. Percept. Psychophys. **82**, 2237–2243.

Sussman, H. M. (**1986**). "A neuronal model of vowel normalization and representation," Brain Lang. **28**(1), 12–23.

Theodore, R. M., and Miller, J. L. (**2010**). "Characteristics of listener sensitivity to talker-specific phonetic detail," J. Acoust. Soc. Am. **128**, 2090–2099.

Townsend, J. T., and Ashby, F. G. (**1978**). "Methods of modeling capacity in simple processing systems," in *Cognitive Theory*, edited by J. Castellan and F. Restle (Erlbaum, Hillsdale, NJ), Vol. 3, pp. 200–239.

Uchanski, R. M., Miller, K. M., Reed, C. M., and Braida, L. D. (**1992**). "Effects of token variability on vowel identification," in *The Auditory Processing of Speech: From Sounds to Words*, edited by M. E. H. Schouten (Mouton de Gruyter, Berlin), pp. 291–302).

van der Feest, S. V. H., Rose, M. C., and Johnson, E. K. (**2022**). "Showing strength through flexibility: Multi-accent toddlers recognize words quickly and efficiently," Brain Lang. **227**, 105083.

Van Stan, J. H., Mehta, D. D., Zeitels, S. M., Burns, J. A., Barbu, A. M., and Hillman, R. E. (**2015**). "Average ambulatory measures of sound pressure level, fundamental frequency, and vocal dose do not differ between adult females with phonotraumatic lesions and matched control subjects," Ann. Otol. Rhinol. Laryngol. **124**, 864–874.

Vaughn, C., Baese-Berk, M., and Idemaru, K. (**2019**). "Re-examining phonetic variability in native and non-native speech," Phonetica **76**, 327–358.

Wong, P. C. M., Nusbaum, H. C., and Small, S. L. (**2004**). "Neural bases of talker normalization," J. Cogn. Neurosci. **16**(7), 1173–1184.

Xie, X., and Jaeger, T. F. (**2020**). "Comparing non-native and native speech: Are L2 productions more variable?," J. Acoust. Soc. Am. **147**, 3322–3347.

Zhang, C., and Chen, S. (**2016**). "Toward an integrative model of talker normalization," J. Exp. Psychol.: Hum. Percept. Perform. **42**(8), 1252–1268.