Psychonomic Bulletin & Review (2022) 29:627–634
https://doi.org/10.3758/s13423-021-02019-3

BRIEF REPORT

Distinct mechanisms for talker adaptation operate in parallel on different timescales

Ja Young Choi[1] · Rita S. N. Kou[2] · Tyler K. Perrachione[2]

Accepted: 23 September 2021 / Published online: 3 November 2021

Abstract

The mapping between speech acoustics and phonemic representations is highly variable across talkers, and listeners are slower to recognize words when listening to multiple talkers compared with a single talker. Listeners' speech processing efficiency in mixed-talker settings improves when given time to reorient their attention to each new talker. However, it remains unknown how much time is needed to fully reorient attention to a new talker in mixed-talker settings so that speech processing becomes as efficient as when listening to a single talker. In this study, we examined how speech processing efficiency improves in mixed-talker settings as a function of the duration of continuous speech from a talker. In single-talker and mixed-talker conditions, listeners identified target words either in isolation or preceded by a carrier vowel of parametrically varying durations from 300 to 1,500 ms. Listeners' word identification was significantly slower in every mixed-talker condition compared with the corresponding single-talker condition. The costs associated with processing mixed-talker speech declined significantly as the duration of the speech carrier increased from 0 to 600 ms. However, increasing the carrier duration beyond 600 ms did not achieve further reduction in talker variability-related processing costs. These results suggest that two parallel mechanisms support processing talker variability: A stimulus-driven mechanism that operates on short timescales to reorient attention to new auditory sources, and a top-down mechanism that operates over longer timescales to allocate the cognitive resources needed to accommodate uncertainty in acoustic-phonemic correspondences during contexts where speech may come from multiple talkers.

Keywords Auditory attention · Talker normalization · Speech perception · Auditory streaming · Talker variability

Despite considerable variability in the acoustic realization of speech sounds across talkers, listeners successfully extract accurate phonetic information from speech signals (Johnson & Mullennix, 1997; Potter & Steinberg, 1950). However, maintaining robust speech perception when faced with talker variability imposes additional processing demands on listeners, which manifest as lower accuracy and/or slower response time for speech perception tasks involving mixed talkers relative to a single talker (e.g., Assmann et al., 1982; Green et al., 1997; Mullennix & Pisoni, 1990). These processing costs appear to be incurred automatically when listeners encounter talker variability (Lim et al., 2019a; Magnuson & Nusbaum, 2007), even when such variability does not obfuscate the phonetic content of the target speech (Choi et al., 2018), and even when listeners have lifelong familiarity with the acoustic-phonemic mappings of the target talkers (Magnuson et al., 2021). Theoretical accounts of speech perception have attempted to explain how listeners become disencumbered by talker variability in terms of access to episodic memory (Goldinger, 1998; Kleinschmidt & Jaeger, 2015), extrinsic normalization via acoustic context (Johnson, 1990; Laing et al., 2012; Sjerps et al., 2019), intrinsic normalization of via secondary acoustic cues (Nearey, 1989; Sussman, 1986), allocation of additional cognitive resources (Nusbaum & Magnuson, 1997), and, more recently, feedforward reorientation of auditory attention (Bressler et al., 2014; Choi & Perrachione, 2019a; Kapadia & Perrachione, 2020; Lim et al., 2021).

An important contribution to understanding how listeners resolve talker variability comes from the role played by preceding speech context during word identification. For instance, early studies showed that talker-specific variation in the fundamental frequency of a preceding sentence could bias

✉ Tyler K. Perrachione
tkp@bu.edu

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

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

the interpretation of an ambiguous vowel sound (Johnson, 1990), reflecting a phenomenon known as "extrinsic normalization" of speech acoustics. Recent work has expanded on this to show how ongoing speech context improves speech processing efficiency by allowing listeners to form a coherent auditory stream (Bregman, 1990) with the current talker as its source. In mixed-talker contexts, the duration and temporal proximity of preceding speech—but not the richness of its phonetic content—affect word recognition efficiency (Choi & Perrachione, 2019a). Similarly, the cognitive demands of processing words spoken by multiple talkers are reduced when stimuli are blocked by talker (Perrachione et al., 2011; Stilp & Theodore, 2020). The improvements to accuracy and response time imparted by talker continuity appear to be automatic, immediate, and independent of listeners' perceptual expectations (Bressler et al., 2014; Carter et al., 2019; Kapadia & Perrachione, 2020; Lim et al., 2019a; Morton et al., 2015), suggesting that talker continuity improves speech processing efficiency by feedforward capture of selective auditory attention (Shinn-Cunningham, 2008). Correspondingly, talker discontinuity (the abrupt change from one talker to another) appears to incur processing costs by disrupting listeners' attention to one auditory object and requiring them to refocus their attention on a new source (Lim et al., 2021; Lim et al., 2019b; Mehrai et al., 2018; Wong et al., 2004). Thus, under an *auditory streaming framework*, the accuracy and response time differential between mixed- and single-talker speech contexts can be understood as speech processing efficiency gains via successful allocation of feedforward auditory selective attention versus efficiency losses from ongoing attentional disruption and reorientation.

An untested prediction of the auditory streaming framework of talker adaptation is that there should be some duration of preceding speech from a continuous talker that is sufficient for fully capturing a listener's auditory attention, thereby rendering their speech processing maximally efficient. That is, in a context where a listener is hearing multiple different talkers in turn, there should be some duration of continuous speech from one talker that would allow speech processing to be as efficient as though the listener were in a single-talker context. Prior work has shown that target words are recognized more efficiently when they are preceded by a brief carrier phrase from the same talker, but the durations of carrier phrases tested (300 and 600 ms) did not fully ameliorate the additional processing costs from the mixed-talker context (Choi & Perrachione, 2019a). Extrapolating linearly from the trend in this prior report, we hypothesized that a continuous speech context of approximately 1,100 ms should allow a listener to become fully adapted to a talker, even in a mixed-talker context.

In contrast, the *active control model* of processing variability in speech (Magnuson & Nusbaum, 2007) postulates a different mechanism behind talker-related inefficiencies in speech perception, and thus makes a different prediction about how much benefit listeners can extract from talker continuity in a mixed-talker situation. In this model, when faced with potential uncertainty about the acoustic composition of speech sounds—such as in listening contexts involving multiple talkers—listeners allocate cognitive resources in anticipation of the need to resolve that acoustic-phonetic uncertainty (Nusbaum & Magnuson, 1997; Nusbaum & Schwab, 1986). Under this account, top-down expectation of talker variability—triggered either by detection of a novel talker or by listeners' situational knowledge—engages additional cognitive resources for determining the phonetic content of speech (Heald & Nusbaum, 2014; Magnuson & Nusbaum, 2007). This allows listeners' perceptual system to be flexible in accommodating variability in the mapping between incoming acoustic signals and internal phonetic representations, but at the cost of increased processing time compared with listening contexts where variability is not expected. Thus, while short-term talker continuity may refocus the listener's auditory attention and make speech processing more efficient (Choi & Perrachione, 2019a), this feedforward process may not ultimately be sufficient to completely ameliorate the cognitive costs of talker variability when listeners are expecting to hear multiple talkers and are thus subjecting the speech signal to additional top-down analysis in anticipation of phonetic ambiguity from the ongoing variability (Heald & Nusbaum, 2014). The active control model thus predicts that, faced with uncertainty from potential talker changes, speech processing will *always* be less efficient in a mixed-talker context compared with a single-talker context.

In this study, we aimed to understand how talker adaptation unfolds over time. Listeners identified spoken words that were preceded by various durations of continuous speech from the same talker in blocks where they either expected to hear speech from multiple different talkers (mixed-talker contexts) versus one single talker (single-talker contexts). By parametrically varying the duration of continuous speech from the talker on each trial, we investigated how the processing costs associated with talker variability change as listeners rapidly adapt to the new talker on each trial. In particular, we were interested to ascertain what duration of preceding speech, if any, would allow participants' word identification in a mixed-talker condition to be as efficient as in a single-talker condition.

Methods

Participants

Native speakers of American English (N = 24; 20 females, four males; mean age = 19.8 years, range: 18–22 years) successfully completed this study. Additional participants were

recruited for this study but excluded from analysis (n = 5) because they performed below our a priori inclusion criterion, requiring >90% accuracy in each condition (Choi & Perrachione, 2019a). All participants reported a history free from speech, language, or hearing disorders. No participant had previously participated in a similar experiment in our laboratory or had prior experience with the talkers. Participants provided written informed consent, approved and overseen by the Institutional Review Board at Boston University.

Stimuli

The naturally spoken English words "boot" and "boat" were recorded by eight native speakers of American English (four females, four males). These words were chosen because their minimally contrastive vowels (/u/ vs. /o/) have the greatest potential acoustic-phonemic ambiguity across talkers (Choi et al., 2018; Hillenbrand et al., 1995).

In addition to the target words, speakers were also recorded producing a brief, sustained "uh" before the words "boot" and "boat" ([ʌ:but], [ʌ:bot]). These recordings were spliced at the end of the silent portion between the closure and the release burst of /b/ so that the sustained /ʌ:/ could be used as a carrier to elicit talker adaptation/attentional reorientation (Choi & Perrachione, 2019a; Johnson, 1990). Among numerous recordings of the carrier, the token with the most stable formant frequencies, amplitude, and fundamental frequency was selected for each talker. Then, using the *pitch synchronous overlap-and-add* algorithm (PSOLA; Moulines & Charpentier, 1990), implemented in the software Praat (Boersma, 2001), the duration of the voiced part of the carrier was adjusted so that the total duration of the carrier equaled 300, 600, 900, 1,200, and 1,500 ms. These carriers were then prepended to the target words. This carrier was chosen because an isolated vowel carrier (e.g., "A…") has been shown to induce as much adaptation as phonetically rich carrier phrases (e.g., "I owe you a…") of equivalent duration (Choi & Perrachione, 2019a).

Recordings were made in a sound-attenuated booth using a Shure MX153 microphone and Roland Quad Capture sound card, sampled at 44.1 kHz and 16-bit resolution. Stimuli were RMS amplitude normalized to 65 dB sound pressure level (SPL) in Praat.

Procedure

Participants performed a speeded word recognition task in which they identified the target word as quickly and accurately as possible by pressing a corresponding number on a keypad. Participants received verbal instructions at the beginning of the experiment, and written directions assigning a number to each target word were displayed on the screen throughout the experiment. Stimuli were presented with a 1,500-ms interval between the onset of the target word and the onset of the following stimulus (see Fig. 1), and stimulus delivery was controlled using PsychoPy2 (Version 1.83.03; Peirce, 2007) via Sennheiser HD-380 Pro headphones.

The task was divided into separate blocks, parametrically varying in talker variability (single vs. mixed) and carrier duration (0, 300, 600, 900, 1,200, 1,500 ms). Each block was 48 trials long, with each target word presented 24 times per block. Stimuli were presented in a pseudorandom order such that the same target word did not repeat in more than three consecutive trials and the same talker did not repeat in adjacent trials in mixed-talker conditions. Participants heard all talkers during the mixed talker conditions. The talker heard during each single talker condition was counterbalanced across participants and carrier lengths. The order of conditions was counterbalanced across participants.

Results

Accuracy and response time (RT) data were collected on each trial. Accuracy in each condition was calculated as the proportion of trials where the participant responded correctly out of all trials. Because our a priori inclusion criteria required participants to have accuracy above 90% in every condition, our planned analyses focused on differences in RT alone, as in Choi and Perrachione (2019a). Analogous tasks have shown limited dynamic range for analyses of accuracy (Kapadia & Perrachione, 2020), and participants' word identification in the present study was at ceiling (mean accuracy 99% ± 1% across participants). Correspondingly, the dependent measure of interest in the present study was RT, which serves as a metric of speech processing efficiency.

RT was measured as the delay between the onset of the target word and the participant's keypad response. RT was log-transformed to more closely approximate a normal distribution. Only RTs from correct trials were included in the analysis. Outlier trials with log-transformed RTs exceeding three standard deviations from the participant's mean for that condition were also excluded from RT analysis (0.8% of all trials). Data were analyzed in R using linear mixed-effects models implemented in the package *lme4*, with log-transformed RTs as the dependent variable. Categorical fixed factors included *talker variability* (single vs. mixed talker) and *carrier duration* (0, 300, … , 1,500 ms). The model also contained random effect terms for within-participant slopes for talker variability and carrier duration and random intercepts for participants (Barr et al., 2013). In the model design matrix, deviation-coded contrasts were applied to the talker variability factor, and contrasts for successive differences (i.e., 0 vs. 300, 300 vs. 600, etc.) were applied to the carrier duration factor.

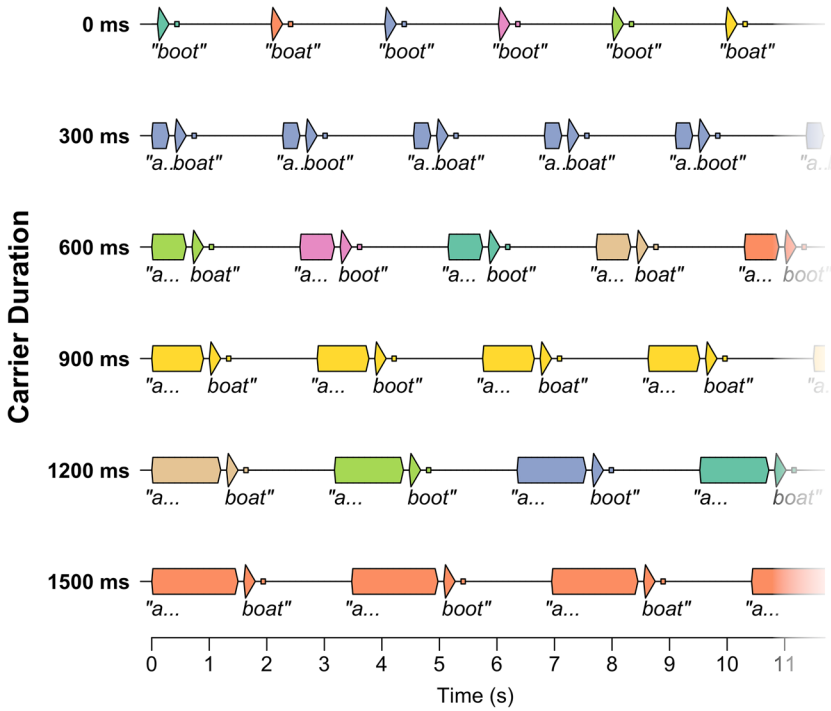

Fig. 1 Schematic of the task design. A stylized version of the task acoustics is shown, depicting the speech waveforms from different talkers in different colors. Participants identified the target word ("boot" or "boat") on each trial. Target words were presented in isolation (0-ms carrier) or preceded by a sustained vowel /ʌː/ from the same talker. Trials were blocked by carrier length (varying parametrically from 0 to 1,500 ms) and by the single-talker or mixed-talker condition. *Mixed-talker* conditions are shown for 0-ms, 600-ms, and 1,200-ms carriers; and *single-talker* conditions are shown for 300-ms, 900-ms, and 1,500-ms carriers

Significance of fixed factors was determined in a Type III analysis of variance (ANOVA). Significant effects from the ANOVA were followed by post hoc pairwise analyses using difference of least squares means implemented in the package *lsmeans* in R and testing contrasts on the terms of linear mixed-effects model using the package *lmerTest* in R. Significance of main effects and interactions was determined by adopting the significance criterion of $\alpha = 0.05$, with *p*-values based on the Satterthwaite approximation of degrees of freedom.

The ANOVA of the linear mixed effects model of RT revealed a main effect of *talker variability* such that RTs in the mixed-talker conditions were significantly slower than those in the single-talker conditions overall, $F(1, 23) = 76.41$, $p \ll .0001$. Post hoc pairwise analysis showed that, within every level of carrier duration, RTs in the mixed-talker condition were significantly slower than the corresponding single-talker condition (see Table 1 and Fig. 2a). Carrier duration had no significant effect on overall RT, $F(5, 23) = 1.10$, $p = .39$.

There was a significant interaction between *talker variability* and *carrier duration,* $F(5, 13084) = 10.08$, $p \ll .0001$). Contrast terms on the linear mixed effect model showed that this interaction was significant between the 0-ms and 300-ms conditions ($\beta = -0.013$, $SE = 0.0059$, $t = -2.25$, $p < .025$) and between the 300-ms and 600-ms conditions ($\beta = -0.021$, $SE = 0.0059$, $t = -3.56$, $p < .0004$), but not for any of the conditions with longer carriers (all $|\beta| < 0.004$, all $|t| < 0.62$, all $p > .53$; see Table 2 and Fig. 2b). That is, the difference in RT between the mixed-talker and single-talker conditions decreased as the carrier duration increased from none to 300 ms, and again from 300 ms to 600 ms, but further increases of carrier duration did not reduce the effect of talker variability, which remained significant for all carrier lengths through 1,500 ms.

Discussion

In this study, we explored to what extent immediately preceding speech from a talker can ameliorate the processing costs that talker variability imposes on word identification. We found that the RT difference between single-talker and mixed-talker conditions steadily decreased as the duration of the preceding speech context increased from 0 to 600 ms. Beyond 600 ms, however, additional exposure to continuous speech from each talker in the mixed-talker condition did not further facilitate speech processing efficiency compared with the single-talker condition. Processing speed in the mixed-talker condition was always slower than in the single-talker condition, even when the target word was preceded by 1,500 ms of continuous speech from the same talker. This piecewise pattern of results—a continuous reduction in

Table 1 Response time differences between talker variability conditions (by carrier duration)

Carrier duration (ms)	RT (mean ± *SD* ms)			Difference of least square means			
	Single talker	Mixed talker	Difference	*β*	*SE*	*t*	*p*
0	674 ± 131	752 ± 103	78 ± 70	0.115	0.011	10.40	≪.0001
300	668 ± 108	734 ± 118	66 ± 67	0.089	0.011	8.00	≪.0001
600	685 ± 108	720 ± 121	35 ± 43	0.047	0.011	4.22	≪.0001
900	691 ± 143	724 ± 122	33 ± 58	0.051	0.011	4.62	≪.0001
1,200	699 ± 118	742 ± 126	44 ± 59	0.058	0.011	5.27	≪.0001
1,500	699 ± 117	742 ± 114	42 ± 57	0.061	0.011	5.54	≪.0001

mixed-talker interference for connected speech contexts up to 600 ms followed by a constant difference thereafter—is inconsistent with an account of talker adaptation based on a single mechanism. Instead, these data suggest that at least two independent mechanisms are in play: One for rapid adaptation, which continuously improves speech processing efficiency up to ~600 ms of exposure, and a second for sustained expectation of talker variability that operates over longer timescales, such as at least the length of one of the experimental blocks.

The first mechanism supporting talker adaptation appears to be a stimulus-driven reorientation of auditory attention that unfolds up to ~600 ms. Within this time frame, listeners experience continuous improvements in speech processing efficiency after encountering a new talker. This is in line with other recent observations that word recognition is facilitated immediately after hearing speech from the same talker (Carter et al., 2019; Choi & Perrachione, 2019a; Kapadia & Perrachione, 2020; Lim et al., 2019a; Morton et al., 2015). There are several reasons to think that these efficiency gains are the result of stimulus-driven reorientation of auditory attention: First, temporal discontinuity between the adapting speech and the target word disrupts this effect (Choi & Perrachione, 2019a), consistent with other evidence that temporal discontinuity interrupts attention to speech (Best et al., 2008; Bressler et al., 2014). Second, this process appears to depend on continuity in the auditory modality, as nonmatching or nonauditory primes do not facilitate auditory word recognition (Morton et al., 2015). Third, this process appears to be engaged automatically and independent of listeners' top-down expectations about who the talker will be (Carter et al., 2019). Neurophysiological correlates of speech processing under talker variability also lend support to the idea that a feedforward, attention-based mechanism partially underlies talker adaptation: Abrupt talker discontinuity alters evoked neural responses to auditory onsets and

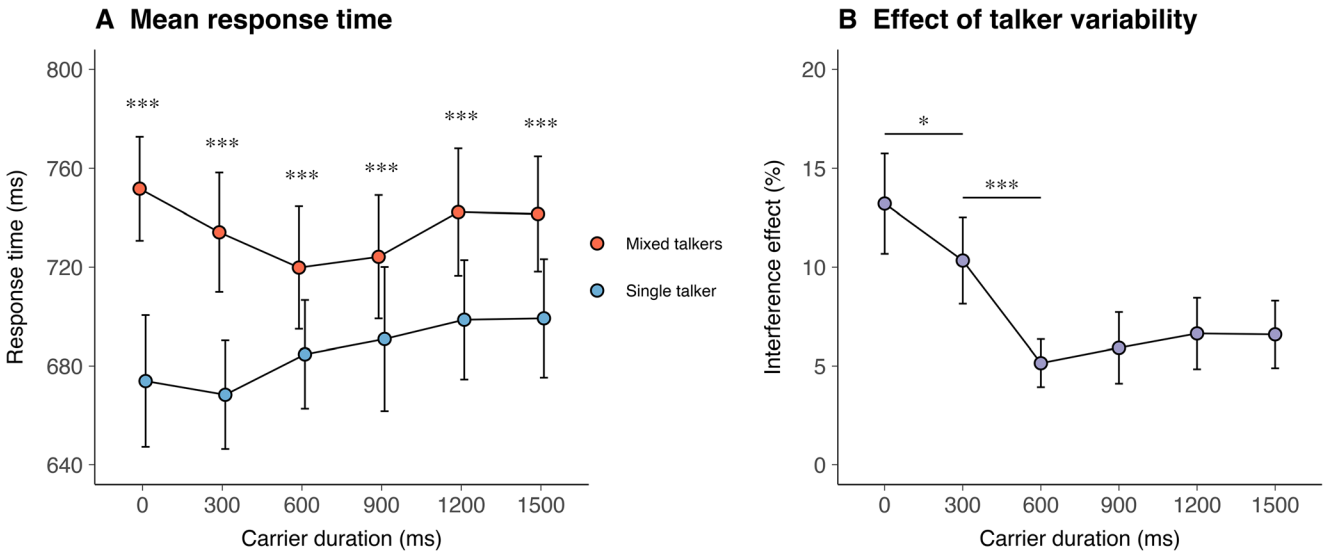

Fig. 2 Effects of talker variability and carrier duration across talkers on response times. **a** Mean response times were faster in the single-talker condition than the corresponding mixed-talker condition at every carrier length. **b** The effect of talker variability is shown for each level of carrier duration. The mean response time difference between the mixed-talker and single-talker conditions is shown, scaled within participant: ((mixed − single)/single) × 100. The effect of talker variability was significantly reduced as the carrier duration increased from 0 to 300 ms, and 300 to 600 ms, but remained constant for all longer carriers. Error bars indicate ±1 *SEM* across participants. *$p < .05$; ***$p < .0005$

Table 2 Interactions between talker variability and carrier duration on log response time

Contrast	Interaction with talker variability			
	β	*SE*	*t*	*p*
0 vs. 300 ms	−0.013	0.006	−2.251	<.025
300 vs. 600 ms	−0.021	0.006	−3.559	<.0004
600 vs. 900 ms	0.002	0.006	0.373	.709
900 vs. 1,200 ms	0.004	0.006	0.616	.538
1,200 vs. 1,500 ms	0.001	0.006	0.249	.803

desynchronizes attention-related neural alpha oscillatory power (Mehrai et al., 2018), and talker discontinuities evoke greater pupil dilation responses and larger late cortical potentials associated with distractor suppression (Lim et al., 2021). Similarly, noninvasive electrical stimulation of left temporal lobe disrupts the behavioral facilitation associated specifically with local talker continuity in global mixed-talker contexts (Choi & Perrachione, 2019b).

A second mechanism supporting talker adaptation appears to involve changes to the mental computations that support speech processing, which are realized over longer timescales than those involved in feedforward attentional reorientation. It is only during sustained periods of listening to one talker, free from the possibility of having to hear another talker, when listeners seem able to maximize their speech processing efficiency. One proposed difference in speech processing that fits this timeframe is in the extent to which top-down cognitive resources are deployed in anticipation of the processing demands associated with talker variability—a mechanism described in the active control model of speech processing (Heald et al., 2015). According to this framework, speech perception is a cognitively active process, in which incoming speech signals are compared against their various possible interpretations, the cognitive-computational demands of which increase in contexts where there is greater acoustic-phonemic uncertainty (Nusbaum & Schwab, 1986; see also Kleinschmidt & Jaeger, 2015).

Several converging lines of evidence suggest that this active control mechanism is best characterized as a difference in mental states, in which listeners either expect to hear speech from a single talker (minimizing the computational demands of speech processing) or from more than one talker (triggering a more computationally demanding, and therefore less efficient, mode of speech processing; Heald & Nusbaum, 2014; Magnuson & Nusbaum, 2007). First, this mechanism operates over a relatively long timeframe and seems to be insensitive to short-term expectations about the talker: Knowing that speech will alternate predictably between two talkers does not improve word recognition efficiency compared with speech from random talkers (Kapadia & Perrachione, 2020). Priming the identity of the upcoming talker in mixed-talker contexts, whether via short auditory (Choi & Perrachione, 2019a) or visual cues (Morton et al., 2015), also does not allow speech to be processed as efficiently as in a single-talker context. Second, this mechanism appears to operate in a categorical fashion: The additional cognitive demands of speech processing are the same regardless of how many different talkers there are beyond one (Kapadia & Perrachione, 2020; Mullennix & Pisoni, 1990). Furthermore, when listeners are performing a task that involves the possibility of talker variability, they remain slower to process speech even during brief periods of talker continuity: Within a mixed-talker context, word recognition during 10-s blocks of speech from a continuous talker remained less efficient than word recognition in longer, single-talker contexts (Stilp & Theodore, 2020), and in a context where the talker could change randomly on any trial, even 12-s spans of speech from a single talker did not offer additional improvement in speech processing efficiency versus a 4-s span (Lim et al., 2019a). Although trial-by-trial talker continuity in an otherwise mixed-talker condition facilitates word recognition, it does not facilitate it as much as listening in a single-talker context (Choi & Perrachione, 2019a; Kapadia & Perrachione, 2020; Morton et al., 2015; Stilp & Theodore, 2020). Generally, the idea that two dissociable cognitive mechanisms might underlie talker adaptation over different timescales parallels the idea that multiple distinct sensory/perceptual mechanisms underlie different aspects of short-term auditory normalization (reviewed in Sjerps et al., 2019), together underscoring the computational complexity of ecological speech processing.

Of these two mechanisms, the shorter-timescale one appears clearly related to stimulus-driven reorientation of auditory attention (reviewed above and in Lim et al., 2021). However, it remains an open question what process or circuit accomplishes the cognitive resource allocation postulated by the active control model. The patterns of mixed versus single talker interference discussed above are reminiscent of questions in the cognitive control and task switching literature (reviewed in Kiesel et al., 2010), suggesting by analogy that domain-general cognitive resources may be implicated in the additional processing demands of mixed-talker contexts. Indeed, there is some evidence that talker variability poses additional demands on working memory resources (Antoniou & Wong, 2015; Lim et al., 2019b; Nusbaum & Morin, 1992). However, brain imaging studies that compare neural activation during speech recognition in single-talker versus mixed-talker contexts find differences almost exclusively in bilateral temporal areas associated with speech processing (Belin & Zatorre, 2003; Chandrasekaran et al., 2011; Perrachione et al., 2016; Wong et al., 2004), not lateral prefrontal areas associated with domain-general cognitive operations (e.g., Fedorenko et al., 2013). This observation may not actually exclude working memory as a mechanism, as there is

a growing body of research to suggest that working memory for speech itself may rely on the same superior temporal circuits that carry out speech recognition (Jacquemot & Scott, 2006; Koenigs et al., 2011; Leff et al., 2009; Majerus, 2013; Perrachione et al., 2017; Scott & Perrachione, 2019), in contrast to the strict dissociation between verbal working memory and phonological processing posited in classical theories (Baddeley, 1986, 2003). However, while transcranial electrical stimulation of left superior temporal lobe during word recognition disrupts the facilitatory effect of short-term talker continuity that is associated specifically with stimulus-driven attentional reorientation, such stimulation does not affect the longer-term differences between listening in sustained single-talker versus mixed-talker blocks that should depend on both short-term and long-term adaptation mechanisms (Choi & Perrachione, 2019b). This may suggest there is a hemispheric dissociation between the short (stimulus-driven attentional reorientation) and long (state-driven working memory allocation) timescales of talker adaptation—a possibility consistent with other evidence of talker-specific speech processing in right temporal areas (Luthra, 2021; Myers & Theodore, 2017). Ultimately, both the algorithmic and implementational (Marr, 1982) nature of the cognitive resources associated with talker adaptation over longer timescales remains an important area for future work.

Conclusions

Talker adaptation in speech processing appears to be a multi-component process, supported by at least two dissociable cognitive mechanisms: a stimulus-driven, feedforward process that operates on continuous speech up to approximately 600 ms to reorient listeners' auditory attention to a new speech source; and a top-down, feedback process that operates on longer timescales to allocate additional cognitive resources to the increased processing demands of resolving phonetic ambiguity across talkers.

Acknowledgments We thank Virginia Best, Sung-Joo Lim, Yaminah Carter, Jessica Tin, Grace Mecha, Michelle Njoroge, Kamilah Harunna, Chinazo Otiono, Maya Saupe, Yijing Lin, Amabel Antwi, Nicole Chen, and Barbara Shinn-Cunningham. This work was supported in part by the National Institute on Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health under awards R01 DC004545 (to Gerald Kidd).

References

Antoniou, M., & Wong, P. C. M. (2015). Poor phonetic perceivers are affected by cognitive load when resolving talker variability. *Journal of the Acoustical Society of America*, *138*, 571–574. https://doi.org/10.1121/1.4923362

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*(4), 975–989. https://doi.org/10.1121/1.387579

Baddeley, A. D. (1986). *Working memory*. : Clarendon Press.

Baddeley, A. D. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience, 4*(10), 829–839. https://doi.org/10.1038/nrn1201

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*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

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

Best, V., Ozmerol, E. J., Kopčo, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences of the United States of America, 105,* 13174–13178.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*(9/10), 341–345.

Bregman, A. S. (1990). *Auditory scene analysis*. MIT Press.

Bressler, S., Masud, S., Bharadwaj, H., & Shinn-Cunningham, B. (2014). Bottom-up influences of voice continuity in focusing selective auditory attention. *Psychological Research*, *78*, 349–360.

Carter, Y. D., Lim, S.-J., & Perrachione, T. K. (2019). *Talker continuity facilitates speech processing independent of listeners' expectations*. Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia.

Chandrasekaran, B., Chan, A., & Wong, P.C.M. (2011). Neural processing of what and who information in speech. *Journal of Cognitive Neuroscience*, *23*(10), 2690–2700.

Choi, J. Y., Hu, E. R., & Perrachione, T. K. (2018). Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing. *Attention, Perception, & Psychophysics*, *80*, 784–797.

Choi, J. Y., & Perrachione, T. K. (2019a). Time and information in perceptual adaptation to speech. *Cognition*, *192*, Article 103982.

Choi, J. Y., & Perrachione, T. K. (2019b). Noninvasive neurostimulation of left temporal lobe disrupts rapid talker adaptation in speech processing. *Brain and Language*, *196*, Article 104655.

Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions of frontal and parietal cortex. *Proceedings of the National Academy of Sciences, 110*(41), 16616–16621.

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

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. L. M., Klos, S., & Nusbaum, H. C. (2015). Understanding speech in the context of variability. In G. Hickok & S. Small (Eds.), *Neurobiology of language* (pp. 195–206). Academic Press.

Heald, S. L. M., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, *8*, 35.

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.

Jacquemot, C., & Scott, S.K. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Science, 10,* 480–486.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America, 88,* 642–654.

Johnson, K., & Mullennix, J. W. (Eds.). (1997). *Talker variability in speech processing*. Academic Press.

Kapadia, A. M., & Perrachione, T. K. (2020). Selecting among competing models of talker adaptation: Attention, cognition, and memory in speech processing efficiency. *Cognition*, *204*, Article 104393.

Kiesel, A., Steinhauer, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin, 136,* 849–874.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognizing the familiar, generalizing to the similar, and adapting to the novel. *Psychological Review*, *122*(2), 148.

Koenigs, M., Acheson, D. J., Barbey, A. K., Solomon, J., Postle, B. R., & Grafman, J. (2011). Areas of left perisylvian cortex mediate auditory-verbal short-term memory. *Neuropsychologia, 49*(13), 3612–3619.

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). https://doi.org/10.3389/fpsyg.2012.00203

Leff, A.P., Schofield, T. M., Crinion, J. T., Seghier, M. L., Grogan, A., Green, D. W., & Price, C. J. (2009). The left superior temporal gyrus is a shared substrate for auditory short-term memory and speech comprehension: Evidence from 210 patients with stroke. *Brain, 132,* 3401–3410.

Lim, S.-J., Carter, Y. D., Njoroge, J. M., Shinn-Cunningham, B. G., & Perrachione, T. K. (2021). Talker discontinuity disrupts attention to speech: Evidence from EEG and pupillometry. *Brain & Language, 221,* Article 104996.

Lim, S.-J., Qu, A., Tin, J.A.A., & Perrachione, T.K. (2019a). *Attentional reorientation explains processing costs associated with talker variability*. 19th International Congress of Phonetic Sciences (Melbourne, August 2019).

Lim, S.-J., Shinn-Cunningham, B. G., & Perrachione, T. K. (2019b). Effects of talker continuity and speech rate on auditory working memory. *Attention, Perception, & Psychophysics, 81,* 1167–1177.

Luthra, S. (2021). The role of the right hemisphere in processing phonetic variability between talkers. *Neurobiology of Language*, *2*(1), 138–151.

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*(2), 391–409.

Magnuson, J.S., Nusbaum, H.C., Akahane-Yamada, R., & Saltzman, D. (2021). Talker familiarity and the accommodation of talker variability. *Attention, Perception, & Psychophysics, 83*, 1842–1860. https://doi.org/10.3758/s13414-020-02203-y

Majerus, S. (2013). Language repetition and short-term memory: an integrative framework. *Frontiers in Human Neuroscience,* 7, 357.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. Henry Holt and Co.

Mehrai, G., Shinn-Cunningham, B., & Dau, T. (2018). Influence of talker discontinuity on cortical dynamics of auditory spatial attention. *NeuroImage, 179,* 548–556. https://doi.org/10.1016/j.neuroimage.2018.06.067

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5/6), 453–467.

Morton, J. R., Somers, 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., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*(4), 379–390.

Myers, E. B., & Theodore, R. M. (2017). Voice-sensitive brain networks encode talker-specific phonetic detail. *Brain and Language, 165,* 33–44.

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

Nusbaum, H. C., & Magnuson, J. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. A. Johnson & J. W. Mullennix (Eds.), *Talker variability and speech processing* (pp. 109–132). Academic Press.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, E. Vatikiotis-Bateson, & Y. Sagisaka (Eds.), *Speech perception production and linguistic structure* (pp. 113–134). IOS Press.

Nusbaum, H. C., & Schwab, E. C. (1986). The role of attention and active processing in speech perception. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern recognition by humans and machines: Speech perception* (Vol. 1, pp. 113–157). Academic Press.

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., Halverson, K., Ghosh, S. S., Christodoulou, J. A., & Gabrieli, J. D. E. (2016). Dysfunction of rapid neural adaptation in dyslexia. *Neuron, 92,* 1383–1397.

Perrachione, T. K., Ghosh, S. S., Ostrovskaya, I., Gabrieli, J. D. E., & Kovelman, I. (2017). Phonological working memory for words and nonwords in cerebral cortex. *Journal of Speech, Language, and Hearing Research, 60,* 1959–1979.

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *Journal of the Acoustical Society of America, 130,* 461–472. https://doi.org/10.1121/1.3593366

Potter, R. K., & Steinberg, J. C. (1950). Toward the specification of speech. *Journal of the Acoustical Society of America, 22,* 807–820.

Scott, T. L., & Perrachione, T. K. (2019). Common cortical architectures for phonological working memory identified in individual brains. *NeuroImage, 202,* Article 116096.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186.

Sjerps, M. J., Fox, N. P., Johnson, K., & Chang, E. F. (2019). Speaker-normalized sound representations in the human auditory cortex. *Nature Communication*, 10, 2465. https://doi.org/10.1038/s41467-019-10365-z

Stilp, C. E., & Theodore, R. M. (2020). Talker normalization is mediated by structured indexical information. *Attention, Perception, & Psychophysics*, *82,* 2237–22431. https://doi.org/10.3758/s13414-020-01971-x

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

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

Open practices statement

The raw data, analysis code, stimuli, and stimulus delivery scripts from this project are available online via our institutional repository (https://open.bu.edu/handle/2144/16460). This study was not preregistered.