Cognition 192 (2019) 103982

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Time and information in perceptual adaptation to speech

Ja Young Choi[a,b], Tyler K. Perrachione[a,*]

[a] Department of Speech, Language, and Hearing Sciences, Boston University, Boston, MA, United States
[b] Program in Speech and Hearing Bioscience and Technology, Harvard University, Cambridge, MA, United States

ARTICLE INFO

Keywords:
Speech perception
Phonetic variability
Categorization
Talker normalization
Adaptation

ABSTRACT

Perceptual adaptation to a talker enables listeners to efficiently resolve the many-to-many mapping between variable speech acoustics and abstract linguistic representations. However, models of speech perception have not delved into the variety or the quantity of information necessary for successful adaptation, nor how adaptation unfolds over time. In three experiments using speeded classification of spoken words, we explored how the quantity (duration), quality (phonetic detail), and temporal continuity of talker-specific context contribute to facilitating perceptual adaptation to speech. In single- and mixed-talker conditions, listeners identified phonetically-confusable target words in isolation or preceded by carrier phrases of varying lengths and phonetic content, spoken by the same talker as the target word. Word identification was always slower in mixed-talker conditions than single-talker ones. However, interference from talker variability decreased as the duration of preceding speech increased but was not affected by the amount of preceding talker-specific phonetic information. Furthermore, efficiency gains from adaptation depended on temporal continuity between preceding speech and the target word. These results suggest that perceptual adaptation to speech may be understood via models of auditory streaming, where perceptual continuity of an auditory object (e.g., a talker) facilitates allocation of attentional resources, resulting in more efficient perceptual processing.

1. Introduction

A core challenge in speech perception is the lack of a one-to-one mapping between acoustic signals and intended linguistic categories (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Talkers differ in their vocal tract anatomy, dialect and speech mannerisms (Johnson, Ladefoged, & Lindau, 1993), resulting in different talkers using remarkably different acoustics to produce the same phoneme, or virtually identical acoustics to produce different phonemes (Hillenbrand, Getty, Clark, & Wheeler, 1995). Because of this variation, listening to speech from multiple or different talkers imposes additional processing costs, resulting in slower and less accurate speech perception than when listening to speech from a single consistent talker (Mullennix & Pisoni, 1990; Magnuson & Nusbaum, 2007). The empirical phenomenon of a talker-specific mode of listening, in which speech is processed faster and more accurately, is called *talker adaptation*, and has been observed across a number of experimental paradigms and for a variety of dependent measures (e.g., Kraljic & Samuel, 2007; Dahan, Drucker, & Scarborough, 2008; Trude & Brown-Schmidt, 2012; Xie, Earle, & Myers, 2018).

A common account of how listeners maintain phonetic constancy across talkers is *talker normalization* (Johnson, 2005; Nusbaum & Magnuson, 1997; Pisoni, 1997), in which listeners use both signal-intrinsic (e.g., Nearey, 1989) and -extrinsic (e.g., Johnson, 1990) information about a talker to establish talker-specific mappings between acoustic signals and abstract phonological representations. Previous studies that have dealt with inter-talker variability mostly asked listeners to decide which of two sounds (e.g., /ba/ vs. /da/; Green, Tomiak, & Kuhl, 1997) or a very small set of isolated words (e.g., Mullennix & Pisoni, 1990; Cutler, Andics, & Fang, 2011) they heard in single- vs. mixed-talker contexts. However, real-world speech rarely occurs in such form. Most of the speech that we encounter comes from one talker at a time and in connected phrases, rather than from mixed talkers in isolated words. Even during conversations with multiple interlocutors, listeners still tend to get a sustained stream of speech from each talker at a time.

Other studies that have investigated how the indexical context affects acoustic-to-phonetic mapping have demonstrated that listeners' perceptual decision of speech can be biased by preceding speech sounds. Manipulation of features in the prior context affects how the listeners perceived the relevant features of following speech signals (Francis, Ciocca, Wong, Leung, & Chu, 2006; Johnson, 1990; Ladefoged & Broadbent, 1957; Leather, 1983). Although these studies shed light on how the mapping between acoustic signals and speech categories is

* Corresponding author at: Department of Speech, Language, and Hearing Sciences, Boston University, 635 Commonwealth Ave., Boston, MA 02215, United States.
E-mail address: tkp@bu.edu (T.K. Perrachione).

https://doi.org/10.1016/j.cognition.2019.05.019
Received 30 October 2018; Received in revised form 11 May 2019; Accepted 25 May 2019

dynamically influenced by the context surrounding the speech sounds, they have focused mainly on how the context affects perceptual decision *outcomes*, not how it affects speech processing *efficiency*. The influence of context on perceptual decisions is clear from such studies, but they tell us little about how much or what kind of information listeners obtain from preceding contexts, nor do they elucidate the time course of using the context information. These limitations are also apparent in recent models of speech processing. For example, Kleinschmidt and Jaeger (2015) proposed a model of speech perception that achieves perceptual constancy through the comparison between encountered acoustic signals and listeners' expectations based on prior experience. Although this model captures the active, dynamic nature of acoustic-to-phonemic mapping and explains why it is harder for listeners to process mixed-talker speech than single-talker speech, ultimately it accounts for only the decision *outcomes* that listeners make, without considering the psychological or biological operations that the perceptual system must undertake in order to reach those decisions, nor how those operations unfold in real time. Pierrehumbert (2016) posited a hybrid model of speech perception in which episodic traces of acoustic details are used in mapping the speech acoustics to an abstract phonemic representation (see also Goldinger, 1998). However, this model also does not describe the mechanistic processes for how information from prior speech encounters is integrated into perceptual decisions.

Overall, current models have thus achieved impressive success in describing the "computational" and "algorithmic" levels of perceptual adaptation to speech, but so far there has been no sustained attempt to account for the "implementational" level (Marr, 1982). Ultimately, our understanding of adaptation to talkers during speech processing still lacks an implementational description of (*i*) how the system operates in real time to arrive at a specific decision outcome among multiple possible interpretations of target speech acoustics, (*ii*) how much and what kinds of information the system uses to achieve such a decision, and (*iii*) the timescale in which the system integrates information about the talker-specific phonetic context of speech to facilitate its decision process. In this paper, we report a preliminary empirical foundation that describes these three key constraints on the implementational level of talker adaptation, and we propose a potential theoretical framework through which talker adaptation can be explored as the integration between domain-general attentional allocation and linguistic representations.

Neuroimaging studies have shown that adaptation to talker-specific speech is associated with reduced physiological cost (Perrachione et al., 2016; Wong, Nusbaum, & Small, 2004; Zhang et al., 2016), indicating that speech processing becomes more physiologically efficient as the listener adapts to a talker. Studies using electroencephalography (EEG) have shown that talker normalization occurs early in speech processing, thus affecting how the listener perceives the speech sound (Kaganovich, Francis, & Melara, 2006; Sjerps, Mitterer, & McQueen, 2011; Zhang et al., 2016). Furthermore, because such physiological adaptation to speech appears dysfunctional in communication disorders like dyslexia (Perrachione et al., 2016), understanding the implementational, mechanistic features of speech adaptation may help identify the psychological and biological etiology of these disorders. However, reduced physiological cost itself *reflects*, rather than *underlies*, the computational implementation of perceptual adaptation, and neuroimaging studies have not yet shown *how* reduced physiological costs reflect efficiency gains in speech processing. Similarly, physiological adaptation alone does not reveal which indexical or phonetic features of real-world speech facilitate early integration of talker information during speech processing. The development of an implementational model of talker adaptation, building upon the rigorous empirical neurobiology of auditory adaptation (e.g., Froemke & Schreiner, 2015; Fritz, Shamma, Elhilali, & Klein, 2003; Jääskeläinen, Ahveninen, Belliveau, Raij, & Sams, 2007; Winkler, Denham, & Nelken, 2009), depends on a better empirical understanding of the psychological contributions of time and information in perceptual adaptation to speech.

Listeners are faster and more accurate at processing speech from a single talker compared to mixed talkers presumably because they learn something about talker-specific idiosyncrasies from previous speech to adapt to each talker, making future speech processing more efficient. In this study, we aimed to further our understanding of how listeners take advantage of preceding speech context to facilitate perceptual decisions about speech. In particular, we wanted to determine how speech processing efficiency is affected by (*i*) the amount of prior information that listeners have about a talker's speech and (*ii*) how much time they have to integrate that information prior to a perceptual decision. These questions are fundamental to establishing an implementational understanding of talker adaptation, as current models of processing talker variability in speech do not elaborate on how and when relevant information about the target talker's speech is ascertained during speech perception (Kleinschmidt & Jaeger, 2015; Pierrehumbert, 2016).

To assess this question, we carried out a series of three experiments that explore the relationship between the amount of information listeners have about the phonetics of a talker's speech, the amount of time they have to process that information before making a perceptual decision, and the efficiency with which they can access speech content. In these experiments, listeners identified whether they heard the word "boot" or "boat" – a challenging speech distinction given the substantial overlap across talkers in the acoustic-phonetic-phonemic realization of the sounds /u/ and /o/ (Choi, Hu, & Perrachione, 2018; Hillenbrand et al., 1995). Because of the enormous potential confusability of these phonemes across talkers, we expected listeners to be much slower to make this decision in mixed-talker conditions, where the trial-by-trial correspondence between speech acoustics and phonemic targets is less stable, compared to single-talker conditions. In each of the three experiments, we manipulated the amount of information that listeners have about the current talker and the amount of time they have to integrate that information prior to identifying the word ("boot"/"boat") by prepending the target words with carrier phrases of various lengths and contents. Specifically, we focused on how the response time to make the word identification changes as a consequence of listening to mixed talkers as opposed to single talker, which we refer to as the *interference effect* of talker variability.

In Experiment 1, we established that speech processing efficiency is impacted by preceding information about a talker and time to process it. By comparing the reduction in interference imparted by shorter vs. longer carrier phrases, we discovered that interference from mixed talkers is reduced as a function of the amount of preceding speech context. In Experiment 2, we examined how the quality of information in the carrier phrase serves to reduce interference. By comparing the reduction in interference made by a phonetically "complex" carrier phrase vs. a phonetically "simple" one, we discovered that the richness of phonetic information conveyed in the carrier phrase does not affect the magnitude of perceptual adaptation when the temporal duration of the carrier phrase is kept constant. In Experiment 3, we investigated how the speech perception system integrates phonetic information over time. By comparing the duration and temporal proximity of the carrier phrases to the target word, we discovered that a sustained stream of information is necessary over the duration of the context for the perceptual system to maximally facilitate adaptation to the talker.

Overall, these experiments reveal (*i*) that the speech perception system appears to need surprisingly little information about a talker's phonetics in order to facilitate efficient speech processing, (*ii*) that the facilitation effect builds up with longer preceding exposure to a talker's speech, but (*iii*) that this gain depends on temporal continuity between adapting speech and word targets. Together, these experiments reveal how the psychological implementation of rapid perceptual adaptation to speech makes use of continuous integration of phonetic information over time to improve speech processing efficiency.

2. Experiment 1: perceptual adaptation to speech depends on preceding speech context

We first investigated how the amount of talker-specific information available before a target word affected the speed with which listeners could identify that word. In Experiment 1, we asked listeners to decide whether they heard the word "boot" or "boat" in either a single- (easy) or mixed- (hard) talker context. Listeners are reliably slower to make perceptual decisions about speech in mixed-talker contexts (e.g., Mullennix & Pisoni, 1990; Choi et al., 2018), and here we measured the extent to which such mixed-talker interference was reduced as a function of the amount of preceding speech context in three conditions: (*i*) no preceding context, (*ii*) a short preceding carrier phrase spoken by the same talker, and (*iii*) a longer preceding carrier phrase spoken by the same talker. The more information a listener has about the current talker, the better their perceptual system should be able to adapt to the particular phonetic-phonemic correspondences of that talker's speech, and the faster they should be able to make perceptual decisions about the speech. Therefore, we hypothesized that the more preceding speech context a listener heard from the current talker, the faster they would be able to recognize speech by that talker, particularly in a challenging mixed-talker listening task.

2.1. Methods

2.1.1. Participants

Native speakers of American English ($N = 24$; 17 female, 7 male; age 19–26 years, mean = 21.4) participated in this study. All participants had a self-reported history free from speech, hearing or language disorders. Participants gave informed, written consent approved and overseen by the Institutional Review Board at Boston University. Additional participants were recruited for this experiment but were excluded from analysis because they had accuracy below 90% in any of the six conditions ($n = 3$).

Our sample size was determined *a priori* via power analysis in combination with the methodological preference for a fully counterbalanced design across conditions (see below). Previous research using this phonemic contrast in a similar behavioral paradigm (Choi et al., 2018) found that processing speech from mixed vs. single talkers incurs a processing cost of +126 ms (17%), an effect size of Cohen's $d = 0.69$. With $N = 24$, we expected to have 95% power to detect talker adaptation effects of at least this magnitude. From the same study, manipulations of target contrast affected talker adaptation by 50 ms (6%; $d = 0.54$); correspondingly, with this sample size we expected to have > 80% power to detect similar magnitudes of difference in the interference effect.

2.1.2. Stimuli

Stimuli included two target words, "boat" and "boot." These target words were chosen because the phonetic-phonemic correspondence of the /o/-/u/ contrast is highly variable across talkers (Hillenbrand et al., 1995) and therefore highly susceptible to interference in a mixed-talker setting (Choi et al., 2018). During the task, these target words were presented either in isolation, preceded by a short carrier phrase ("It's a [boot/boat]"), or preceded by a long carrier phrase ("I owe you a [boot/boat]"). The carrier phrases were chosen so that they contained increasing amounts of information about the speaker's vowel space and vocal tract configuration, presumably offering listeners different amounts of information about how /o/ and /u/ in "boat" and "boot" would sound for a particular talker prior to encountering those words in the sentence (Fig. 1A and D).

Words and carrier phrases were recorded by two male and two female native speakers of American English in a sound-attenuated room with a Shure MX153 earset microphone and Roland Quad Capture sound card sampling at 44.1 kHz and 16bits. Among numerous tokens of the target words and carriers from these speakers, the best quality recordings with similar pitch contours and amplitude envelopes were chosen as the final stimuli set. Then, the selected tokens for each target word for each speaker were concatenated with each carrier phrase, resulting in four sentences created for each speaker. To ensure the naturalness of concatenated sentences, we manipulated pitch, amplitude, and the voice-onset time of the carrier phrase and target words. All the recordings were normalized for RMS amplitude to 65 dB SPL in Praat (Boersma, 2001). Short carrier phrases were 298–382 ms; long carrier phrases were 544–681 ms. Examples of these stimuli are shown in Fig. 1A.

2.1.3. Procedure

Participants had the task of indicating whether they heard "boot" or "boat" on each trial of the experiment. Trials were organized into six separate blocks that factorially varied whether the stimuli were spoken by one talker (*single-talker* conditions) or all four talkers (*mixed-talker* conditions), and whether stimuli were presented in isolation (*no-carrier* conditions), preceded by the carrier phrase "It's a …" (*short-carrier* conditions), or preceded by the carrier phrase "I owe you a …" (*long-carrier* conditions; see Fig. 2). In each block of 48 trials, each target word occurred in 24 trials. Stimulus presentation was pseudo-randomized such that the same target word was not presented for more than three consecutive trials. The order of conditions was counter-balanced across participants using Latin square permutations. A given participant heard the same talker in all three of their single-talker conditions, and the specific talker used in the single-talker conditions was counterbalanced across participants.

Participants were instructed to listen to the stimuli and identify which target word they heard as quickly and accurately as possible by pressing the corresponding number key on a keypad. Trials were presented at a rate of one per 2000 ms. Stimulus delivery was controlled using PsychoPy v.1.8.1 (Peirce, 2007).

2.1.4. Data analysis

Accuracy and response time data were analyzed for each participant in each condition. Accuracy was calculated as the proportion of trials where participants identified the word correctly out of the total number of trials. Response times were measured from the onset of the target word. Response times were log-transformed to ensure normality. Only the response times from correct trials were included in the analysis. Outlier trials that deviated by more than three standard deviations from the mean log response time in each condition were also excluded from the analysis (< 1% of total correct trials). Data were analyzed in R using linear mixed-effects models implemented in the package *lme4*, with response times on each trial as the dependent variable. Fixed factors included *indexical variability* (single-talker, mixed-talker) and *context* (no-carrier, short-carrier, long-carrier). The models also contained random effect terms of within-participant slopes for indexical variability and context and random intercepts for participants (Barr, Levy, Scheepers, & Tily, 2013).[1] 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 differences of least-squares means via *difflsmeans* and by testing specified contrasts on the terms in the linear mixed-effects model relevant to each experiment using the package *lmerTest*. We adopted the significance criterion of $\alpha = 0.05$, with *p*-values based on the Satterthwaite approximation of the degrees of freedom.

2.2. Results

Participants' word identification accuracy was at ceiling

[1] Across experiments, these models took the form, in R notation: log(response time) ~ indexical variability * context + (1 + indexical variability + context | subject).

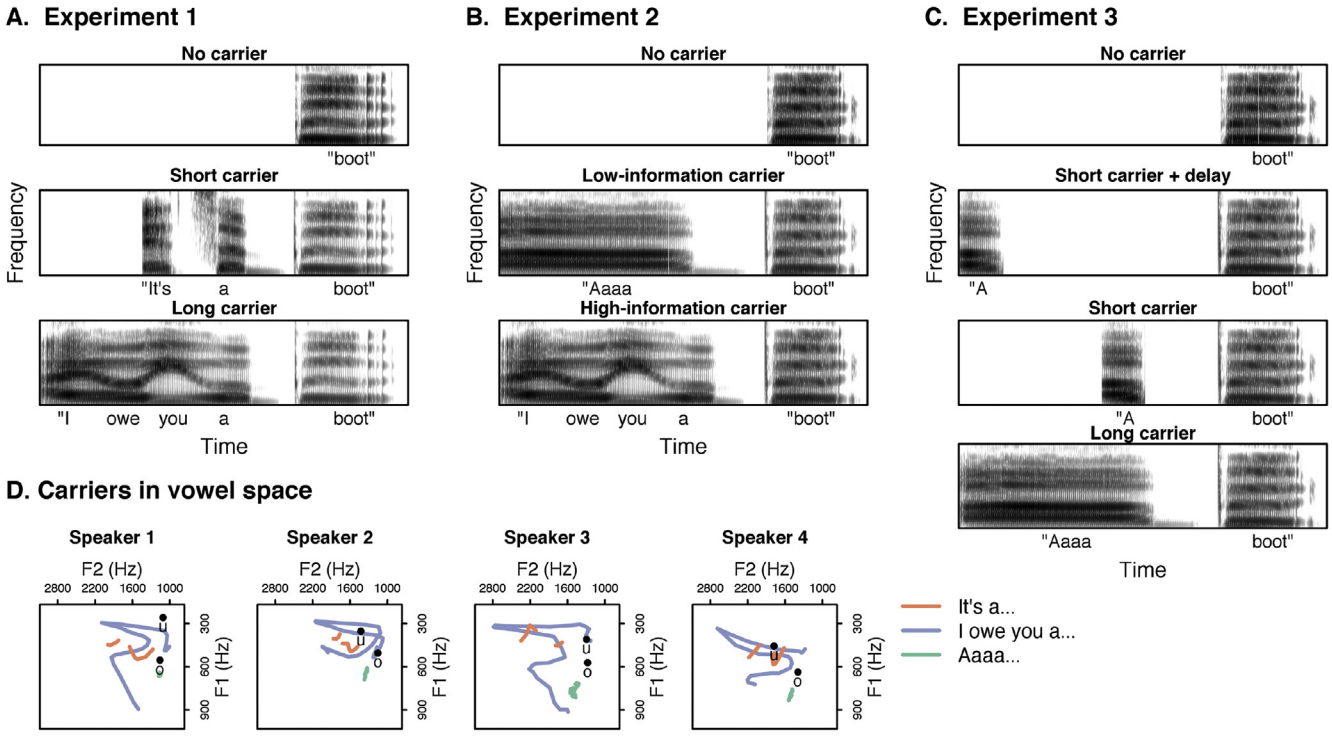

Fig. 1. Stimuli for Experiments 1–3. (A–C) Spectrograms of example stimuli produced by Speaker 2 used in Experiments 1–3 in each condition. (D) Lines indicate the F1-F2 trajectory of all carriers produced by each talker. Black points indicate the F1-F2 position of the /o/ and the /u/ vowels in the target words "boat" and "boot" spoken by each talker. Recordings of all experimental stimuli are available online: https://open.bu.edu/handle/2144/16460.

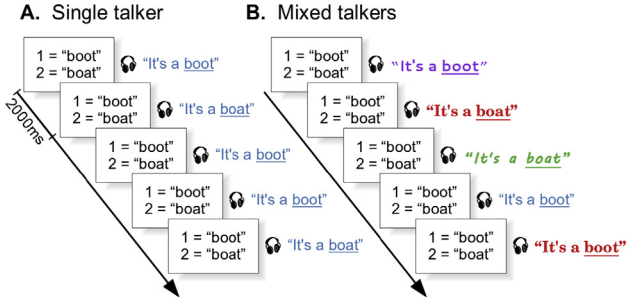

Fig. 2. Task design for all experiments. Participants performed a speeded word identification task while listening to speech produced by either (A) a single talker or (B) mixed talkers. The short-carrier condition for Experiment 1 is shown.

Table 1
Mean ± s.d. response time (ms) in each condition in Experiment 1.

	No carrier	Short carrier	Long carrier
Single talker	698 ± 85	666 ± 78	672 ± 50
Mixed talkers	792 ± 86	736 ± 91	711 ± 70
Differences	95 ± 63	70 ± 56	40 ± 46

(mean = 98% ± 2%). Consequently, the dependent measure for this experiment was response time (Table 1), as is usual for studies of perceptual adaptation in speech perception (e.g., Choi et al., 2018; Magnuson & Nusbaum, 2007; McLennan & Luce, 2005). Participants' response times in each condition are shown in Fig. 3.

There was a significant main effect of *indexical variability* ($F(1, 23) = 109.01$; $p \ll 0.0001$). Post-hoc pairwise testing revealed that response times in the mixed-talker condition were significantly slower than in the single-talker condition overall ($\beta = 0.090$, *s.e.* = 0.009, $t = 10.44$, $p \ll 0.0001$). There was also a main effect of *context* ($F(1, 23) = 11.62$; $p < 0.0004$). Post-hoc pairwise testing revealed that response times in the no-carrier condition were significantly slower than those in either the short-carrier ($\beta = 0.058$, *s.e.* = 0.017, $t = 3.40$, $p < 0.003$) or long-carrier ($\beta = 0.067$, *s.e.* = 0.014, $t = 4.66$, $p < 0.0002$) conditions, but overall response times in the two carrier conditions did not differ ($\beta = 0.009$, *s.e.* = 0.016, $t = 0.56$, $p = 0.58$).

Importantly, there was a significant *indexical variability* × *context* interaction ($F(2, 6659) = 27.52$; $p \ll 0.0001$). Pairwise tests revealed that the effect of indexical variability was significant for each of the three context conditions independently (Table 1; no-carrier single- vs. mixed-talkers: $\beta = 0.124$, *s.e.* = 0.010, $t = 11.96$, $p \ll 0.0001$; short-carrier single- vs. mixed-talker: $\beta = 0.096$, *s.e.* = 0.010, $t = 9.21$, $p \ll 0.0001$; long-carrier single- vs. mixed-talker: $\beta = 0.050$, *s.e.* = 0.010, $t = 4.84$, $p \ll 0.0001$), with the difference appearing to decrease as a function of carrier length. Contrasts on the linear model for successive differences between levels of the speech context factor revealed significant interactions between indexical variability and speech context for both levels of carrier length (short-carrier vs. no-carrier: $\beta = 0.014$, *s.e.* = 0.005, $t = 2.86$, $p < 0.005$; long-carrier vs. short-carrier: $\beta = 0.023$, *s.e.* = 0.005, $t = 4.51$, $p \ll 0.0001$). That is, the interference effect of mixed talkers became successively smaller as the context increased from no carrier to a short carrier to a long carrier. Pairwise comparisons of least squares means revealed this effect to be driven primarily by decreasing response time in the mixed-talker conditions with increasing carrier length (Supplemental Table 1).

Together, this pattern of results indicates that listening to speech from multiple talkers incurred a significant processing cost compared to listening to speech from a single talker, but that the magnitude of this interference was attenuated with larger amounts of preceding talker-specific speech detail, and thus opportunity to perceptually adapt to the target talker, preceding the target word.

2.3. Discussion

The results from the first experiment show that the availability of

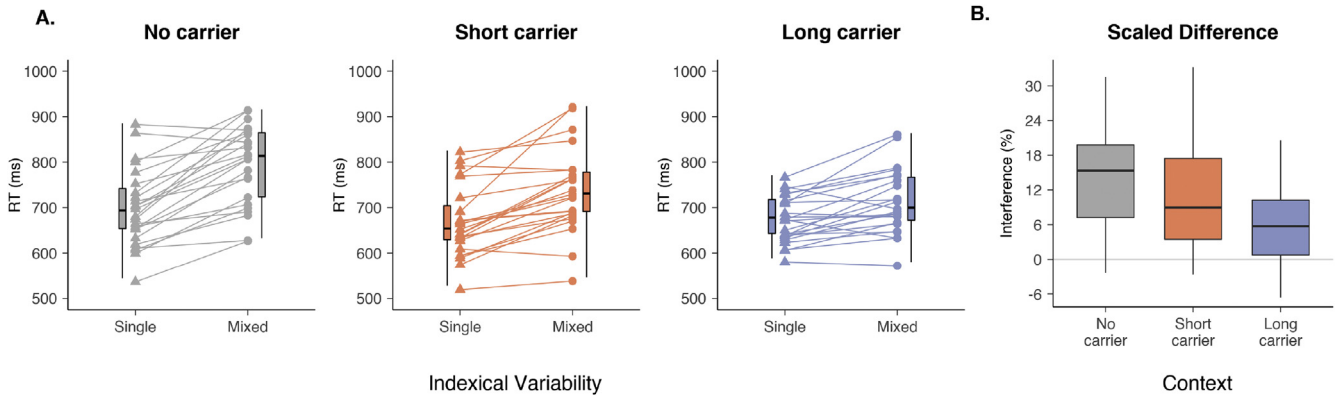

Fig. 3. Results for Experiment 1. Effects of talker variability and context across talkers on response times. (A) Connected points show the change in response times for individual participants between the single- and mixed-talker conditions across three levels of context. Box plots in each panel show the distribution (median, interquartile range, extrema) for each variability-by-context condition. (B) The interference effect of indexical variability is shown for each level of context. The distribution of differences in response time between the mixed- and single-talker conditions is shown, scaled within participant to their response time in the single-talker condition: ((mixed − single)/single) × 100. Significant interference was observed for every level of context. The long-carrier condition showed a significantly smaller interference effect than either the no-carrier or the short-carrier condition.

immediately preceding connected speech from a talker reduces the processing cost associated with speech perception in a multi-talker context. This result extends the observation that the outcomes of perceptual decisions in speech are affected by preceding speech context (Johnson, 1990; Laing, Liu, Lotto, & Holt, 2012) to show that speech context also affects the efficiency of processing a talker's speech, and does so as a function of the amount of speech context available. We observed quantitative differences in the amount of speech processing efficiency gain as a function of time and information in the preceding speech context: Compared to when there is no preceding context, a short ~300 ms speech carrier reduces the processing cost of speech perception in a multi-talker context from 14% to 11%, and a longer, ~600 ms carrier reduces this cost to just 6%. This observation establishes that listeners rapidly adapt to a talker's speech, becoming increasingly efficient at speech perception on the order of hundreds of milliseconds as they accumulate talker-specific information about talkers' speech production.

Although the results from this experiment reveal that increasing the amount of preceding connected speech context from a talker facilitates speech perception for that talker, it remains unresolved precisely why the longer carrier afforded greater perceptual adaptation to speech. In Experiment 1, the long and short carrier conditions differed in at least three ways. First, the two carriers had different total durations: The average duration of the short carrier phrase ("It's a …") was 340 ms, whereas that of the long carrier phrase ("I owe you a …") was 615 ms. Second, they contained different amount of information about a talker's vocal tract and articulation: the short carrier phrase encompassed two vowels (/ɪ/, /ʌ/) that varied primarily in F2, while the long carrier phrase contained at least five distinct vowel targets (/a/, /ɪ/, /o/, /u/, /ʌ/) and effectively sampled the entire vowel space (Fig. 1A and D). Third, the long carrier contained the vowels present in the target words (/o/ and /u/), whereas the short carrier did not. That is, compared to the short carrier, the long carrier both contained richer and more relevant talker-specific detail and provided listeners with more time to adapt to the talker.

In order to ascertain the unique contribution of time and information on perceptual adaptation to speech, we conducted a second experiment in which the duration of the carrier phrases was held constant while the amount of phonetic information conveyed by each carrier was manipulated.

3. Experiment 2: perceptual adaptation in high- and low-information contexts

In this experiment, we assessed the question of whether perceptual adaptation to speech context depends principally on the *quantity* of talker-specific information versus the *duration* (amount of time) available for perceptual adaptation to adjust listeners' phonetic-phonemic correspondences. As in Experiment 1, we used a speeded lexical classification paradigm in which listeners identified words preceded by varying speech contexts. In Experiment 2, we manipulated the carrier phrases so that they were fixed in their durations but differed in the amount of detail they revealed about the talker's vowel space and other articulatory characteristics (Fig. 1B and D): a *high-information* carrier phrase contained a richer amount of information that revealed the extent of each talker's vowel space, whereas a *low-information* carrier phrase revealed talkers' vocal source characteristics, but served only as a spectrotemporal "snapshot" of their vocal tract, with minimal time-varying articulatory information. If perceptual adaptation to speech depends on the amount of talker-specific information available, then the interference effect of mixed-talker speech should be lower in the high-information carrier phrase than the low-information carrier

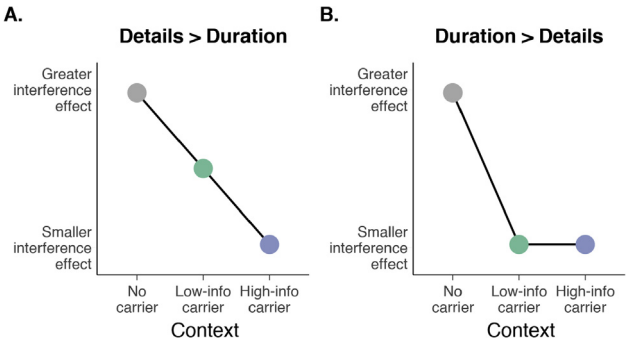

Fig. 4. Hypothesized patterns of results for Experiment 2. Potential patterns for the interference effect of talker variability across the three experimental conditions, as predicted by the two different hypotheses about contextual effects on talker adaptation. (A) If the amount of talker-specific phonetic details in a carrier contributes more to talker adaptation than the duration of the carrier, the interference effect will be lower in the high-information carrier condition than in the low-information carrier condition. (B) If the duration of a carrier contributes more to talker adaptation than the richness of its phonetic details, the interference effect will not differ between the low- and the high-information carriers, as their durations are matched.

(Fig. 4A). However, if perceptual adaptation depends principally on the amount of time available to recalibrate the phonetic-phonemic correspondences computed by the speech perception system – not the amount of information needed to recalculate those correspondences – then the duration-matched high- and low-information carriers should equally reduce the amount of interference in mixed-talker conditions (Fig. 4B).

3.1. Methods

3.1.1. Participants

A new sample of native speakers of American English ($N = 24$; 21 female, 3 male; age 18–26 years, mean = 21.3) participated in this study. All participants had a self-reported history free from speech, hearing or language disorders. Participants gave informed, written consent approved and overseen by the Institutional Review Board at Boston University. Additional participants were recruited for this experiment but were excluded because they had accuracy below 90% in any of the six conditions ($n = 1$). No participant in Experiment 2 had also been in Experiment 1. The sample size in Experiment 2 was determined based on the same paradigm and power-analysis criteria as Experiment 1. In Experiment 1, we found that, between the long- and short-carrier conditions, there was a difference of mixed-talker processing cost on the order of 30 ms (5%; $d = 0.60$). We determined that we would have 80% power to detect effects of a similar magnitude in Experiment 2.

3.1.2. Stimuli

Stimuli included the same two target words “boat” and “boot” from Experiment 1. During the task, these words were presented either in isolation, preceded by the same *high-information* carrier phrase as in Experiment 1 (i.e., “I owe you a [boot/boat]”), or preceded by a *low-information* carrier phrase, in which the vowel /ʌ/ (as the “a” pronounced in “a boat”) was sustained for the length of the high-information carrier (i.e., “Aaaaa [boot/boat]”). Words and carrier phrases were recorded using the same two male and two female native American English speakers and with the same recording procedural parameters as in Experiment 1. Among numerous tokens of the words and carriers from these speakers, the best quality recordings with similar pitch contours and amplitude envelopes were chosen as the final stimuli set. For the low-information carrier, each speaker was recorded briefly sustaining the word “a” (/ʌ/) before saying the target word. The carrier was isolated from the target word, and its duration was adjusted using the *pitch synchronous overlap-and-add* algorithm (PSOLA; Moulines & Charpentier, 1990) implemented in the software Praat so that it matched the duration of the high-information carrier phrase recorded by the same speaker. After choosing the best tokens of each word and carrier, the carriers and targets were concatenated so that they resembled natural speech as in Experiment 1. All the recordings were normalized for RMS amplitude to 65 dB in Praat (Boersma, 2001). Examples of these stimuli are shown in Fig. 1B.

3.1.3. Procedure

Participants had the task of indicating whether they heard “boot” or “boat” on each trial of the experiment. Trials were organized into six blocks that factorially varied whether the stimuli were spoken by one talker (*single-talker* conditions) or all four talkers (*mixed-talker* conditions), and whether stimuli were presented in isolation (*no-carrier* conditions), preceded by the duration-matched carrier, “a…” (*low-information carrier* conditions), or preceded by the carrier phrase, “I owe you a…” (*high-information carrier* conditions). In each block of 48 trials, each target word occurred in 24 trials. Stimulus presentation was pseudo-randomized such that the same target word was not presented for more than three sequential trials.

Participants were instructed to listen to the stimuli and identify which target word they heard as quickly and accurately as possible by pressing the corresponding number key on a keypad. Trials were presented at a rate of one per 2000 ms. Stimulus delivery was controlled using PsychoPy v.1.8.1 (Peirce, 2007). The order of conditions was counter-balanced across participants using Latin square permutations. The same talker’s recordings were used in all single-talker conditions for a given participant, with the specific talker used in single-talker conditions counterbalanced across participants.

3.1.4. Data analysis

As in Experiment 1, accuracy and log-transformed response time data were analyzed for each participant in each condition, with the same operationalization and quality control for these data (< 1% of trials excluded). Data were analyzed in R using the same algorithms, statistical thresholds, and random effect structure as before. Fixed factors in the linear mixed-effects models included *indexical variability* (single-talker, mixed-talker) and *speech context* (no carrier, low-information carrier, high-information carrier).

3.2. Results

Participants' word identification accuracy was again at ceiling (98% ± 2%), and so the dependent measure for this experiment was also response time (Table 2). Participants' response times in each condition are shown in Fig. 5.

As in Experiment 1, there was a significant main effect of *indexical variability* ($F(1,23) = 66.66$, $p \ll 0.0001$), with post-hoc pairwise tests revealing slower response times in mixed- than single-talker conditions overall ($\beta = 0.068$, *s.e.* = 0.008, $t = 8.16$, $p \ll 0.0001$). Again, there was also a main effect of *context* ($F(1, 23) = 7.99$; $p < 0.003$). Post-hoc pairwise testing revealed that response times in the no-carrier condition were significantly slower than those in either the low-information ($\beta = 0.056$, *s.e.* = 0.016, $t = 3.49$, $p < 0.002$) or high-information carrier ($\beta = 0.082$, *s.e.* = 0.021, $t = 3.88$, $p < 0.0008$) conditions. In this experiment, too, response times in the two carrier conditions did not differ significantly (low- vs. high-information carrier: $\beta = 0.026$, *s.e.* = 0.014, $t = 1.80$, $p = 0.085$).

In this experiment, we again observed the significant *indexical variability* × *context* interaction ($F(2, 6627) = 24.96$; $p \ll 0.0001$). Pairwise tests revealed that the effect of indexical variability was significant for each of the three context conditions independently (Table 2; no carrier single- vs. mixed-talkers: $\beta = 0.107$, *s.e.* = 0.010, $t = 10.67$, $p \ll 0.0001$; low-information carrier single- vs. mixed-talker: $\beta = 0.053$, *s.e.* = 0.010, $t = 5.26$, $p \ll 0.0001$; high-information carrier single- vs. mixed-talker: $\beta = 0.046$, *s.e.* = 0.010, $t = 4.63$, $p \ll 0.0001$).

As in Experiment 1, listening to speech in all mixed-talker contexts in Experiment 2 had a deleterious effect on listeners' ability to make perceptual decisions about speech content, even when preceded by talker-specific phonetic information from the carriers, with the difference again appearing to be greatest in the no-carrier condition. However, in Experiment 2, the interference effect of mixed talkers appeared to be roughly equal for the two speech carriers. Contrasts on the linear model for successive differences between levels of the speech context factor revealed a significant interaction between indexical variability and speech context for no carrier vs. the low-information carrier ($\beta = 0.027$, *s.e.* = 0.005, $t = 5.76$, $p \ll 0.0001$); however, the variability-by-context interaction did not differ between the low- and high-information carriers ($\beta = 0.003$, *s.e.* = 0.005, $t = 0.67$, $p = 0.50$).

Table 2
Mean ± s.d. response time (ms) in each condition in Experiment 2.

	No carrier	Low-information carrier	High-information carrier
Single talker	705 ± 128	679 ± 84	662 ± 78
Mixed talkers	784 ± 125	716 ± 87	697 ± 84
Differences	79 ± 54	37 ± 43	35 ± 50

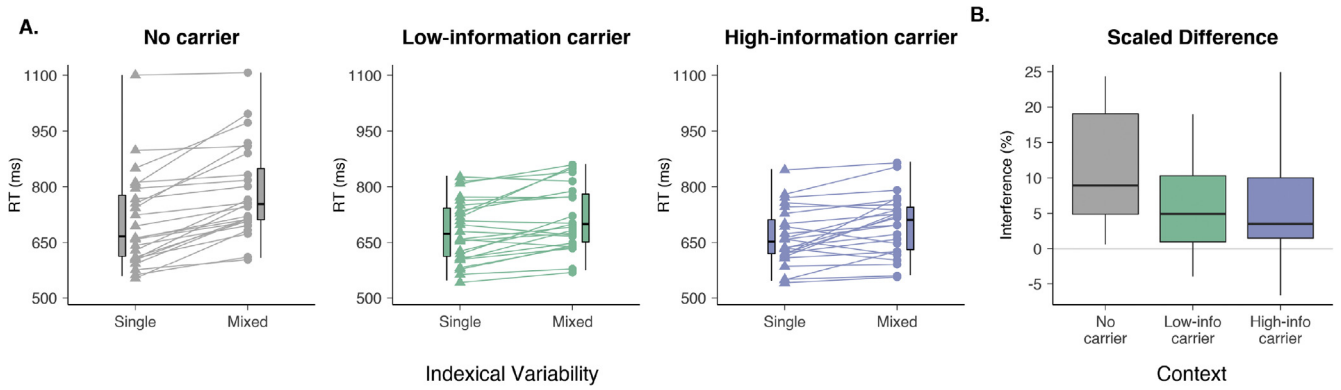

Fig. 5. Results for Experiment 2. Effects of talker variability and context on response times. (A) Connected points show the response times in the single- and mixed-talker conditions across three levels of context for individual participants. Box plots in each panel show the distribution (median, interquartile range, extrema) for each variability-by-context condition. (B) The interference effect of indexical variability is shown for each level of context. The distribution of differences in response time between the mixed- and single-talker conditions is shown, scaled within participant to their response time in the single-talker condition: ((mixed − single)/single) × 100. Significant interference was observed for every level of context. Both the low-information and the high-information carrier conditions showed a significantly smaller interference effect than the no-carrier condition. There was no significant difference in the interference effect between the low-information and high-information carrier conditions. The pattern of results is consistent with what is expected when the duration of carrier is more important factor than the amount of talker-specific phonetic details (Fig. 4B).

As in Experiment 1, these interactions arose primarily due to the carrier phrases' facilitatory effect on listeners' response times in mixed-talker conditions than differences in single-talker conditions (Supplemental Table 2).

This pattern of results replicates the observation from Experiment 1 that speech context facilitates perceptual adaptation to a talker compared to no context. However, when the duration of the preceding context is matched, the amount of talker-specific perceptual adaptation appears to be equivalent regardless of the amount of articulatory-phonetic information available from the talker.

3.3. Discussion

The results from Experiment 2 refine our understanding of the temporal dimension of auditory adaptation to talkers and the source of information that facilitates this adaptation. As in Experiment 1, the interference effect of talker variability was greatest in the no-carrier condition where listeners were not given any preceding speech context, and the effect was reduced in both the low- and high-information carrier conditions where the brief preceding speech context allowed listeners to adapt to the talker on each trial. Surprisingly, Experiment 2 revealed that the increase in processing efficiency afforded by a carrier phrase in multi-talker speech contexts did not differ as a function of the amount or relevance of phonetic information available in the speech carrier. The high-information carrier phrase, highly dynamic in terms of time-frequency information about a talker's vocal tract and articulation and containing in part the same vowels as the target words, yielded no more adaptation than the low-information carrier phrase of the same duration, which was essentially a spectrotemporally-invariant snapshot of the talker. This observation suggests that auditory adaptation requires time to unfold but does not depend on the availability of rich details about the phonetics of a talker's speech.

Previous models of speech perception that assume an abstract representation of a talker's vowel space acknowledge that listeners use their prior experience of a talker to create this representation and use it to understand speech (Kleinschmidt & Jaeger, 2015; Pierrehumbert, 2016). However, these models do not describe the implementational level of these computations; that is, they do not elaborate what kind of or how much talker-specific information is needed to affect perceptual outcomes, nor do they account for how or when the information must be integrated by listeners in order for them to utilize it for the perception of upcoming speech. The results from our experiment show that a carrier phrase that thoroughly samples the talker's vowel space is no more facilitatory than a much more impoverished form of carrier speech, suggesting that the amount of talker-specific information necessary to make speech processing more efficient is, in fact, minimal. It is possible that because inter-talker variability in the acoustic realization of speech is not completely random but rather structured regarding talkers' socio-indexical characteristics (Kleinschmidt, 2018), talker-specific cues with minimal phonetic information might have sufficiently facilitated talker adaptation in Experiment 2.

Coupled with the results of Experiment 1 where a longer carrier phrase afforded greater facilitation of speech processing efficiency than a shorter carrier, the results of Experiment 2 also suggest that the speech perception system requires a sufficient amount of time to integrate talker-specific information to facilitate the processing of future speech content. This raises the question of how the timecourse of such integration unfolds. Some authors have claimed that episodic models of speech processing – in which reactivation of listeners' memories of prior speech experiences guides future speech processing – can account for talker normalization / adaptation phenomena (Goldinger, 1998). Contemporary computational models have explicitly incorporated these mnemonic mechanisms into their perceptual decision processes (Kleinschmidt & Jaeger, 2015; Pierrehumbert, 2016): When a listener hears speech from a particular talker, the speech processing system will implicitly recognize that talker, re-activate related memories of their speech, and integrate them into perceptual processing in order to guide talker-specific interpretation of upcoming speech sounds. However, memory reactivation is a time-dependent process. Consequently, one implication of an episodic account of talker adaptation is that integration of talker specific information will be *ballistic*; that is, once a new talker is encountered, memories of that talker's speech automatically tune the speech perception system to facilitate processing that talker's speech, but a certain amount of time is required for the auditory system to reactivate the relevant memories underlying its perceptual recalibration.

Alternatively, rather than the time-dependent reactivation of memories of similar speech as predicted by episodic/mnemonic models of speech processing, the integration of talker-specific information may depend on continuous integration of a talker's speech over time, akin to auditory streaming and auditory object formation (Shinn-Cunningham, 2008; Winkler et al., 2009). In this account, continuous exposure to a talker's speech facilitates attentional orientation to the relevant auditory features associated with that talker, such that there is a facilitatory effect of not only the length of an adapting speech context, but also its temporal proximity to a speech target. To adjudicate between a

mnemonic/ballistic model of talker adaptation and an object continuity/streaming model, we therefore undertook a third experiment in which we varied both the *duration* of the adapting speech context and its *continuity* with respect to the target word.

4. Experiment 3: effects of temporal proximity and duration on perceptual adaptation

In Experiment 2, we discovered that the amount of time that listeners have to perceptually adapt to a target talker is at least as important as the quantity of information they have about that talker's speech. This observation raises new questions about the original results from Experiment 1: Was the short carrier less effective at reducing interference from the mixed-talker condition because listeners had less time to reactivate talker-specific memories to guide perception of the upcoming word via episodic speech processing (Kleinschmidt & Jaeger, 2015)? Or because they required more time to orient their attention to the relevant talker-specific features via auditory streaming and auditory object formation (Shinn-Cunningham, 2008)? In Experiment 3, we evaluate whether the facilitatory effects of speech adaptation simply require a certain amount of time after an adapting stimulus to take effect, or whether they depend on the continuous integration of talker-specific information over time. That is, we explore whether the processes supporting perceptual adaptation in speech are, in effect, "ballistic," such that exposure to speech from a given talker automatically effects (i.e., primes) changes in listeners' perceptual processing of upcoming speech, or whether adaptation is better understood as "streaming," in which continuous, consistent information proximal to target speech is required for perceptual adaptation.

To evaluate these possibilities, we developed four variations of the carrier phrase manipulation from Experiments 1 and 2. We again utilized the *no-carrier* condition as a baseline for maximal interference and the *long- (low-information) carrier* condition to effect maximal adaptation. In addition, in Experiment 3 we added two new conditions: a *short-carrier without delay* condition, in which listeners heard a short, sustained vowel "a" (/ʌː/) immediately before the target word, and a *short-carrier with delay* condition, in which listeners heard a vowel of the same brief duration, but its onset displaced in time from the target word with a duration equal to that of the long-carrier condition (Fig. 1C).

The mnemonic/ballistic and the object-continuity/streaming models of talker adaptation predict different patterns of facilitation effected by these carrier-phrase conditions in the mixed-talker context. If talker adaptation is ballistic, then once speech is encountered and talker-specific memories are reactivated, we should expect equal amounts of facilitation by the long-carrier and short-carrier-with-delay conditions. Because the onset of speech in these conditions occurs equidistant from the target lexical item, listeners will have had equal time to re-activate talker-specific memories. Correspondingly, both of those conditions should offer greater facilitation than the short-carrier-without-delay, in which speech onset occurs closer to the target word and thus affords less time for activation and integration of talker-specific memories (Fig. 6A). Alternatively, if talker adaptation depends on attentional reorientation via auditory streaming across time, then the pattern of results should be markedly different (Fig. 6B): the long-carrier should offer the greatest facilitation, as it affords the maximum amount of continuous information about a target talker's speech, followed by the short-carrier-without-delay, which has a shorter duration but which ends with equal temporal proximity to the target word, and finally with the least facilitation effected by the short-carrier-with-delay, which not only offers less speech to adapt from, but which also interrupts the continuity of the talker-specific auditory stream.

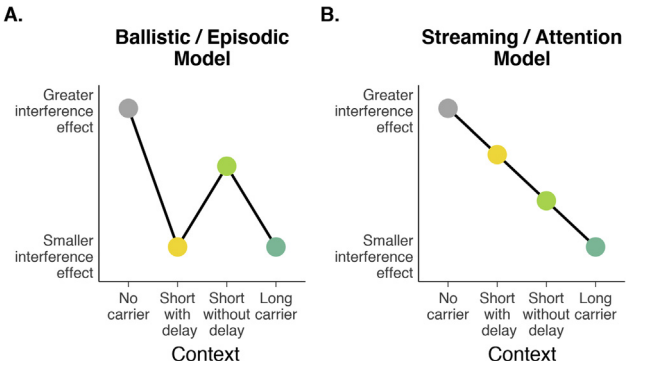

Fig. 6. Hypothesized patterns of results for Experiment 3. Potential patterns for the interference effect of talker variability across the four experimental conditions, as predicted by the two different hypotheses of the contribution of temporal continuity of context. (A) The predicted pattern from an episodic account of speech perception. Due to having the greatest time available to reactivate talker-specific memories, the long-carrier and short-carrier-with-delay conditions should have the smallest (and equal) interference effects. The short-carrier-without-delay has less time to access memories, and so should have a larger interference effect than either of the other carriers. (B) The predicted pattern from an attention/streaming model of speech perception. In contrast to the episodic account, this model predicts a greater interference effect in the short-carrier-with-delay condition than either the short-carrier-without-delay condition or the long-carrier condition. In these latter two conditions, the temporal proximity between the adapting speech and the target word should facilitate the emergence of a talker-specific auditory object and improve processing efficiency.

4.1. Methods

4.1.1. Participants

Another new sample of native speakers of American English (N = 24; 18 female, 6 male; age 18–26 years, mean = 19.8) participated in this study. All participants had a self-reported history free from speech, hearing or language disorders. Participants gave informed, written consent overseen by the Institutional Review Board at Boston University. Additional participants recruited for this experiment (n = 3) were excluded for having accuracy below 90% in any of the eight conditions. None of the participants in Experiment 3 had previously participated in either Experiments 1 or 2.

4.1.2. Stimuli

Stimuli again included the two target words "boat" and "boot." During the task, these words were presented in isolation or preceded by a short-duration carrier ("a boot"), a short-duration carrier with an intervening pause ("a … boot") or a long-duration carrier phrase ("aaaaa boot") (Fig. 1C). Words and carriers were recorded by the same two male and two female native American English speakers as Experiment 1. The long-duration carriers were the same as the low-information carriers used in Experiment 2. The short-duration carriers were resynthesized from each speaker's long-duration carrier, reducing the duration of the pre-closure sustained vowel to 20% of that of the long-carrier. The closure (formant transitions into the closure and silent period before the burst in the /b/ of the target word) were not resynthesized; they remained the same as in the natural recordings. The resulting short carriers (vowel plus /b/ closure) were on average 215 ms in duration. We ensured that the total duration of each speaker's short-duration carriers plus the intervening pause matched the duration of that speaker's long-duration carrier. Each speaker's three carrier phrases were then concatenated with the target words spoken by the same speaker to produce natural-sounding recordings.

4.1.3. Procedure

Participants had the task of indicating whether they heard "boot" or "boat" on each trial of the experiment. Trials were organized into eight blocks that factorially varied whether the stimuli were spoken by one talker (*single-talker* conditions) or all four talkers (*mixed-talker* conditions), and whether stimuli were presented in isolation (*no-carrier*

conditions), preceded immediately by the short-duration carrier "a" (*short-duration carrier without delay* conditions), preceded by the short-duration carrier with an intervening pause (*short-duration carrier with delay* conditions), or preceded by the long-duration carrier "aaaaa" (*long-duration carrier* conditions). In each block of 48 trials, each target word occurred in 24 trials. Stimulus presentation was pseudo-randomized such that the same target word was not presented for more than three sequential trials. The order of conditions was counter-balanced across participants using Latin square permutations. The talker heard by a listener was constant across all four of their single-talker conditions, and which of the four talkers appeared in the single-talker condition was counterbalanced across participants.

Participants were instructed to listen to the stimuli and identify which target word they heard as quickly and accurately as possible by pressing the corresponding number key on a keypad. Trials were presented at a rate of one per 2000 ms. Stimulus delivery was controlled using PsychoPy v.1.8.1 (Peirce, 2007).

4.1.4. Data analysis

Like Experiments 1 and 2, accuracy and response time data were analyzed for each participant in each condition, with the same operationalization and quality control for these data (< 1% of trials excluded). Data were again analyzed in R using the same algorithms, statistical thresholds, and random effect structure as before. Fixed factors in the linear mixed-effects models included *indexical variability* (single-talker, mixed-talker) and *speech context* (no carrier, short-duration carrier with delay, short-duration carrier without delay, long-duration carrier).

4.2. Results

Participants' word identification accuracy was again at ceiling (99% ± 2%), and so as in Experiments 1 and 2, the dependent measure for Experiment 3 was response time (Table 3). Participants' response times in each condition are shown in Fig. 7.

As before, there was a significant main effect of *indexical variability* ($F(1, 23) = 71.89$; $p \ll 0.0001$), with post-hoc tests revealing slower response times in mixed- than single-talker conditions overall ($\beta = 0.077$, *s.e.* = 0.009, $t = 8.47$, $p \ll 0.0001$). Likewise, there was a significant main effect of *context* ($F(3, 23) = 22.38$; $p \ll 0.0001$). Post-hoc pairwise testing revealed that response times were faster for all carrier phrases compared to words in isolation, but that there were minimal differences between the carrier phrase conditions (Supplemental Table 3).

The interaction between *indexical variability* × *speech context* was again significant in Experiment 3 ($F(3, 8891) = 16.55$; $p \ll 0.0001$). Pairwise tests revealed that the effect of indexical variability was significant for each of the four context conditions independently (Table 3; no carrier single- vs. mixed-talkers: $\beta = 0.111$, *s.e.* = 0.011, $t = 10.35$, $p \ll 0.0001$; short-carrier-with-delay: $\beta = 0.083$, *s.e.* = 0.011, $t = 7.74$, $p \ll 0.0001$; short-carrier-without-delay: $\beta = 0.065$, *s.e.* = 0.011, $t = 6.07$, $p \ll 0.0001$; long-carrier: $\beta = 0.047$, *s.e.* = 0.011, $t = 4.42$, $p \ll 0.0001$). Like Experiments 1 and 2, listening to speech in every mixed-talker context in Experiment 3 imposed a processing cost on listeners' ability to make perceptual decisions about speech content, notwithstanding the type or proximity of the carrier phrase.

Table 3
Mean ± s.d. response time (ms) in each condition in Experiment 3.

	No carrier	Short carrier with delay	Short carrier without delay	Long carrier
Single talker	670 ± 72	649 ± 60	651 ± 72	640 ± 71
Mixed talkers	754 ± 85	706 ± 67	698 ± 77	671 ± 67
Differences	84 ± 56	57 ± 53	47 ± 44	31 ± 54

The pairwise differences between response times to single- vs. mixed-talker speech across the four levels of speech context appear to suggest that the interference effect of mixed-talker speech decreases as a function of carrier length and proximity to the target word – a pattern consistent with the attention/streaming model. To test whether the overall pattern of carrier-phrase facilitation was more consistent with the predictions of an episodic model (Fig. 6A) or a streaming model (Fig. 6B), we conducted a series of polynomial contrasts on our mixed-effects model. Based on the ordering of levels of the speech context factor shown in Fig. 6, the episodic model predicts a pattern of variability × context interactions following the trajectory of a cubic function (i.e., down-up-down, with the greatest interference for no carrier, less interference for a short carrier with delay matched in onset time to the long carrier, more interference again for a short carrier without delay, and finally least interference for a long carrier), whereas the attentional model predicts a pattern of interaction magnitudes that is strictly linear (i.e., sequentially smaller for those levels in that order). The polynomial contrast on the interaction effects in the mixed effects model was significant for the linear term ($\beta = 0.023$, *s.e.* = 0.003, $t = 7.00$, $p \ll 0.0001$), but not for the cubic term ($\beta = 0.001$, *s.e.* = 0.003, $t = 0.32$, $p = 0.75$) – a pattern consistent with an attentional model, but not an episodic model.

It is worth noting that the ordering of levels of the context factor is arbitrary, and that re-ordering the levels produces a different shape to the predictions of the two models. If the order of the two short carrier terms were swapped (i.e., no-carrier, short-carrier-without-delay, short-carrier-with-delay, long-carrier), the streaming model would now predict a cubic fit for the data across conditions (i.e., a down-up-down pattern to the difference in interference). If the factors are ordered in this way, the cubic term for the interaction effect is now significant ($\beta = 0.013$, *s.e.* = 0.003, $t = 3.93$, $p \ll 0.0001$) – a pattern consistent with the streaming model, but not one predicted by the episodic model.

Finally, we analyzed the linear mixed-effects model using a contrast matrix to compare successive pairwise differences of the levels of the speech context factor. These contrasts revealed that the variability-by-context interaction was significantly greater for no carrier vs. the short carrier with delay ($\beta = 0.014$, *s.e.* = 0.005, $t = 2.94$, $p < 0.004$), and trended in the direction suggested by the streaming model for the short carrier with delay vs. the short carrier without delay ($\beta = 0.009$, *s.e.* = 0.005, $t = 1.90$, $p = 0.057$), and for the short carrier without delay vs. long carrier ($\beta = 0.009$, *s.e.* = 0.005, $t = 1.89$, $p = 0.059$). Re-ordering the levels of this factor, as above, revealed that the variability-by-context interaction was significantly greater for no carrier vs. the short carrier without delay ($\beta = 0.023$, *s.e.* = 0.005, $t = 4.85$, $p \ll 0.0001$) and for the short carrier with delay vs. the long carrier ($\beta = 0.018$, *s.e.* = 0.005, $t = 3.79$, $p < 0.0002$). As in Experiments 1 and 2, these interactions tended to be driven by shorter response times in the mixed-talker conditions as the carrier became longer and adjacent to the target word, rather than differences in response times in the single-talker conditions (Supplemental Table 4). Overall, the relative differences between these speech context levels follow the pattern predicted by the streaming model in Fig. 6B, but not the episodic model in Fig. 6A.

4.3. Discussion

The pattern of results from the third experiment appears to be primarily consistent with the predictions made by an object continuity/streaming model of talker adaptation, but inconsistent with those made by a mnemonic/ballistic model. Processing interference due to mixed talkers was reduced most by a long carrier, less by a short carrier immediately adjacent to the target word, and least by a short carrier temporally separated from the target word. These results follow the pattern expected if listeners are continuously integrating talker-specific features over time as they adapt to a talker's speech (Fig. 6B), rather than the time required to re-activate memories of a talker once

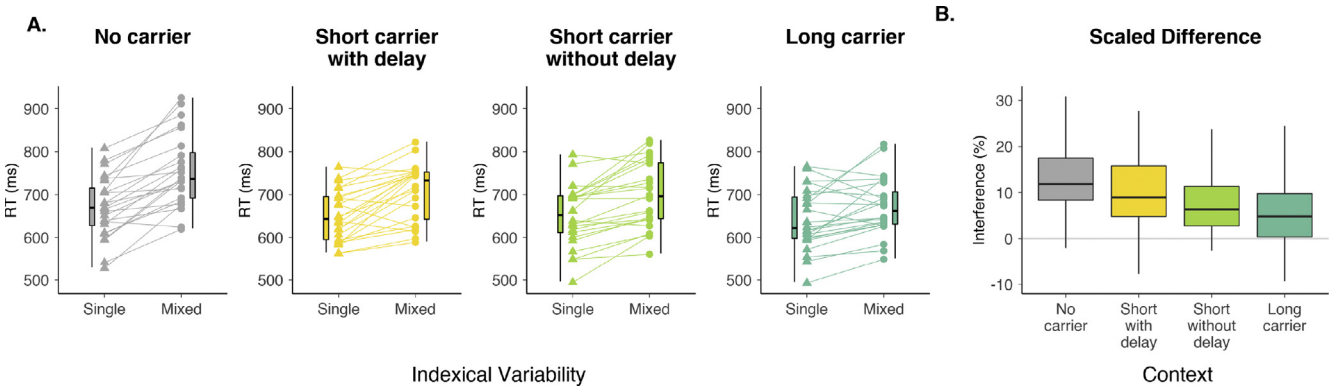

Fig. 7. Results for Experiment 3. Effects of talker variability and context across talkers on response times. (A) Connected points show the change in response times for individual participants between the single- and mixed-talker conditions across four levels of context. Box plots in each panel show the distribution (median, interquartile range, extrema) for each variability-by-context condition. (B) The interference effect of indexical variability is shown for each level of context. The distribution of differences in response time between the mixed- and single-talker conditions is shown, scaled within participant to their response time in the single-talker condition: ((mixed − single)/single) × 100. Significant interference was observed for every level of context. The duration of the carrier phrase and its temporal proximity (continuity) to the target speech both contributed to reducing the processing cost on speech perception associated with mixed talkers. This pattern of results is consistent with what the streaming/attention model predicts (Fig. 6B).

encountered (Fig. 6A). An episodic model of talker adaptation would predict similar amounts of reduction in interference by the short carrier with delay and by the long carrier, as both offer an equal amount of time to reactivate talker-specific memories. A memory-based model would also predict a greater reduction in interference by a short carrier with delay than one immediately adjacent to the target speech. However, we observed the opposite of both of these predictions in our data: the short carrier with delay was least effective in facilitating talker adaptation.

It has been shown that temporal continuity is an important feature that allows perceptual object formation and auditory streaming (Best, Ozmeral, Kopčo, & Shinn-Cunningham, 2008; Bressler, Masud, Bharadwaj, & Shinn-Cunningham, 2014; Woods & McDermott, 2015). Thus, both the temporal continuity and the duration of the incoming speech signal are important factors that allow listeners to integrate a set of acoustic signals as a single auditory object (here, a talker), focus their attention on it, and ultimately process it more efficiently. In the context of this experiment, the long-carrier and short-carrier-with-delay conditions provided listeners with the same temporal duration to adapt to the talker but differed in temporal continuity. Ultimately, the lack of temporal continuity in speech resulted in a reduced facilitatory effect of talker adaptation when compared to either a time-matched continuous signal or a quantity-matched adjacent signal. The long-carrier condition may have provided listeners with more time to build an auditory stream that involves the carrier and the target word than the short-carrier-without-delay conditions, although these two carriers did not differ in terms of continuity with the target word. In the short-carrier-with-delay conditions, the facilitatory effect yielded by the carrier was significantly smaller than the effect yielded by the long carrier even though both conditions should have provided the listeners with the same amount of time to adapt to the talker. However, in the short-carrier-with-delay condition, the build-up of a coherent auditory stream over time may have been hindered by the temporal gap between the carrier and the target word, leading to less facilitation compared to the short-carrier-without-delay condition.

5. General discussion

In this study, we explored how listeners utilize preceding speech context to adapt to different talkers, making acoustic-to-linguistic mappings more efficient despite cross-talker variability in the acoustic realization of speech sounds. Across three experiments that factorially manipulated the duration, richness of phonetic detail, and temporal continuity of carrier phrases, participants' speech processing in a mixed-talker context was always more efficient when they heard target words preceded by a speech carrier than when they heard the words in isolation. This suggests that the perceptual system incorporates preceding speech context in a way that not only biases the perceptual outcomes of speech perception (e.g., Johnson, 1990), but also makes speech perception more efficient. Moreover, based on the findings from Experiment 1, we found that the interference from mixed talkers was reduced as a function of the amount of preceding speech context from each talker, even for as little as 300–600 ms of preceding information.

Interestingly, in Experiment 2, we observed that a prior speech context consisting of only a single sustained vowel had just as much facilitatory effect as another context that fully sampled each talker's entire vowel space, provided the preceding speech samples had the same duration. Thus, the gradient effect of carrier length on perceptual adaptation observed in Experiment 1 may be ascribable to the varying durations of the short and the long carriers, rather than the difference in the amount of information that each carrier entailed. Following up on these results, in Experiment 3, we explored how the perceptual adaptation process unfolds in time. The results from Experiment 3 indicated that, rather than the time preceding the target speech, it appears to be the combination of the speech context's duration and temporal continuity with respect to the target speech that underlies the facilitatory effect of preceding context. Together, the findings from these three experiments provide a preliminary empirical foundation for developing an implementational-level understanding of how perceptual adaptation to speech occurs in real time. Further, when evaluated in the context of theoretical frameworks that invoke either memory or attention as the mechanism underlying efficiency gains in perceptual adaptation to speech, these results suggest that a model of speech adaptation that accounts for temporal effects on adaptation appears compellingly similar to one that relies on domain-general attentional processes for auditory object-continuity and streaming.

5.1. Extension and refinement of prior models of talker adaptation

5.1.1. Contextual tuning models

Previous studies exploring the impact of extrinsic cues on the perception of following target speech have primarily emphasized the role of context as a frame of reference against which the target speech can be compared to affect the outcomes of perceptual decisions. For example, variation in the F1 of an introductory sentence can bias perceptual decisions for following, acoustically identical, speech sounds (Ladefoged & Broadbent, 1957). This biasing effect of context is consistent with *contextual tuning theory*, which proposes that preceding

speech provides talker-specific context (i.e., the talker's vocal characteristics) for interpreting the following speech target (Nusbaum & Morin, 1992). Contemporary models have formalized such propositions for determining perceptual outcomes for speech, as in the *ideal adapter framework* (Kleinschmidt & Jaeger, 2015). However, context does more than just provide a reference for weighting perceptual decisions about speech categories; preceding speech also allows listeners to process target speech contrasts more efficiently. The present results raise the possibility that the mechanisms underlying this efficiency gain may be analogous to those important for feedforward allocation of auditory attention in perceptual streaming, namely, the duration and temporal continuity of the preceding content.

Surprisingly, the amount of phonetic information does not appear to be a critical factor in the efficiency gains associated with talker adaptation, suggesting that early models of talker normalization as explicit perceptual modeling of speakers' vocal tracts (e.g., Joos, 1948; Ladefoged & Broadbent, 1957) may not accurately capture the perceptual mechanisms of adaptation, which instead appear to be more akin to automatic, bottom-up allocation of attentional resources (e.g., Bressler, Masud, Bharadwaj, & Shinn-Cunningham, 2014; Choi et al., 2018). This observation also raises the question of what kinds of information *are* necessary or sufficient for auditory object formation for a given talker. In this study, we found that a sustained, neutral vowel was sufficient for listeners to successfully "adapt" to a talker, reducing perceptual interference from listening to speech in a mixed-talker setting. Others have shown that similarly little – even nonlinguistic – information in a preceding auditory stream can bias perceptual decisions (e.g., Laing et al., 2012), and that listeners can successfully build auditory streams about highly variable sources of speech, provided the information is temporally contiguous (Woods & McDermott, 2018).

5.1.2. Active control process models

The facilitatory effect of context on perceptual adaptation has been explained with models that treat speech perception as an active process of building possible hypotheses and testing them against the incoming signal. Such models often propose an active control mechanism (e.g., Magnuson & Nusbaum, 2007; Wong & Diehl, 2003), by which some cognitive process monitors incoming speech and initiates the computations underlying perceptual adaptation (e.g., Nearey, 1989) in the presence or expectation of talker variability. According to such an account, the perceptual interference induced by mixed-talker speech (e.g., Assmann, Nearey, & Hogan, 1982; Green et al., 1997; Mullennix & Pisoni, 1990; Morton, Sommers, & Lulich, 2015; Choi et al., 2018) can be interpreted as the cognitive cost of engaging the active control mechanism when talker variability is detected (or even just assumed; cf. Magnuson & Nusbaum, 2007). Under an active control process account, listeners can engage this active control mechanism when they encounter each new talker's carrier phrase in mixed talker conditions. Consequently, the perceptual system will not need to expend as much cognitive effort to map the incoming acoustics of the subsequent target word to its intended linguistic representation. The present results offer a potential extension to the active control framework by suggesting that one possible mechanism underlying this control process – the cognitive process effecting efficiency gains in speech perception – may be the successful, feedforward allocation of attention for auditory streaming and auditory object formation. The observation from Experiment 3 that there is less talker adaptation in the short-carrier-with-delay condition than in either the short-carrier-without-delay or the long-carrier conditions also seems to imply that any active control process needs to operate over a sustained, temporally continuous auditory signal. This reduces the likelihood that the cognitive process underlying speech processing efficiency gains reflects top-down effects of episodic memory such as priming. The operationalization of this cognitive process as one of attentional allocation is further supported by the observation that, while the long carrier provides no additional phonetic information compared to the short carrier with delay, it still affords greater adaptation to the target talker. This demonstrates that an active control process is unlikely to be one that builds a sophisticated phonetic model of a talker's speech and/or vocal tract to operate as a transformation function for normalizing incoming speech signals, but instead may be one that is involved in identifying a coherent auditory object in the environment to which to allocate attention (Shinn-Cunningham, 2008). An extensive literature in the fields of perception and attention has shown that attentional allocation enhances perceptual sensitivity and decreases the cognitive cost for perceptual identification (e.g., Best, Ozmeral, & Shinn-Cunningham, 2007; Kidd, Arbogast, Mason, & Gallun, 2005; Alain & Arnott, 2000).

5.1.3. Episodic memory models

An alternative account of talker-specific speech processing that has been invoked to explain efficiency gains under talker adaptation is an episodic model of speech perception (e.g., Goldinger, 1998). In episodic models, memories of encountered speech contain rich details about the speech, such as who was speaking, rather than just storing its abstract phonetic content. An episodic account of speech perception could plausibly have been advanced as an explanation for the results seen in Experiments 1 and 2. Under such an account, when the listener obtains a cue to the talker they will hear, they can retrieve the appropriate talker-specific exemplars of the target words, even when the amount of talker specific information is seriously limited in its duration or phonetic content (e.g., Bachorowski & Owren, 1999), as in the short-carrier from our Experiment 1 or the low-information carrier from Experiment 2, respectively. Memory retrieval is not an instantaneous process; having more time to match an auditory prime against talker-specific memories (as in the long-carrier of Experiment 1, or either carrier in Experiment 2) would improve the likelihood that an appropriate episode could be retrieved. Correspondingly, under an episodic model, we would predict the same pattern of facilitation as what we observed in Experiments 1 and 2 – carriers with longer durations having more facilitatory effect than a short carrier, regardless of the amount of their phonetic contents. However, the results of Experiment 3 provide an important caveat, in that they do not appear to conform to the predictions of a mnemonic account of talker-specific efficiency gains in speech processing. Under an episodic account, we would have expected the amount of facilitation afforded by the short-carrier-with-delay and long-carrier conditions to be similar, since these two conditions provide listeners with the same amount of time and phonetic information from which to retrieve relevant talker-specific exemplars. What we instead observed in Experiment 3 was the opposite of this prediction; there were greater efficiency gains from a long carrier and a temporally contiguous short carrier than from a short carrier with delay.

These empirical data also offer the opportunity to consider how more recent, formal models of talker-specific speech processing may be extended to incorporate a temporal (implementational) level of explanation of talker adaptation. The highly influential ideal adapter framework of Kleinschmidt and Jaeger (2015) has formalized the episodic view of talker-specific speech processing. Specifically, this model posits that the perceptual decision outcomes in speech are the result of recognizing an internal model of a talker that has a similar cue distribution as the incoming signal, thus correctly matching internal models of speech to incoming speech acoustics: When the number of potential models is large, validating the correct model will be slower and less accurate. However, when the number of models is smaller – such as when a listener can limit model selection to a single talker – speech recognition will be faster. The computation underlying this internal model selection is described as an inference that draws not only on bottom-up evidence from the speech signal but also top-down expectation from signal-extrinsic cues such as visual or phonetic cues (Kleinschmidt & Jaeger, 2015, pp. 180–182). While this framework has been highly successful in its account of the perceptual outcomes of speech processing, future work in this area should also consider what kinds of information the perceptual system needs in order to choose the

correct talker model and, critically, how the perceptual system incorporates such cues over time. The present study suggests a novel extension to the empirical and theoretical framework for understanding the implementational-level mechanisms of short-term perceptual adaptation to a talker's speech: In examining the patterns of how talker adaptation unfolds over time, our results raise the possibility that the efficient, feedforward allocation of auditory attention involved in streaming/object formation may operate as the active cognitive process underlying talker adaptation.

5.2. Auditory attention and streaming as a candidate implementational-level explanation for talker adaptation

Explaining the findings from Experiment 3, in which the duration of speech context and its temporal continuity with the target speech afforded maximal talker adaptation, requires us to consider possible mechanisms by which talker-specific information can be continuously integrated over time to improve perception. One domain where such a mechanism already exists it that of auditory scene analysis, in which the attentional selection of auditory objects occurs via streaming (Shinn-Cunningham, 2008; Winkler et al., 2009). Successfully deploying attention to an auditory object relies heavily on temporal continuity (Best et al., 2008; Shinn-Cunningham, 2008), occurs automatically when there is featural continuity (Bressler et al., 2014; Woods & McDermott, 2015; Lim, Shinn-Cunningham, & Perrachione, in press), and enhances the efficiency of perceptual processing of an auditory source (Shinn-Cunningham & Best, 2008; Duncan, 2006; Cusack, Decks, Aikman, & Carlyon, 2004). Under this framework, the delay between the carrier and the target word in the short-carrier-with-delay condition of Experiment 3 disrupts the integration of the carrier and the target word into a coherent auditory object, resulting in less talker adaptation and a greater interference effect in mixed-talker environments compared to the other carrier phrases, which were temporally continuous with the target speech.

Findings from neuroimaging studies on perception and attention provide additional evidence for a role of attentional allocation in the efficiency gains attributed to talker adaptation. Prior expectation modulates the magnitude of neural adaptation to repeated stimuli (Summerfield & Egner, 2009; Todorovic & de Lange, 2012), and auditory feature-specific attention affects neurophysiological adaptation, as measured by fMRI (Alho, Rinne, Herron, & Woods, 2014; Altmann, Henning, Döring, & Kaiser, 2008; Da Costa, van der Zwaag, Miller, Clarke, & Saenz, 2013). These findings that top-down attention and expectation drive neural adaptation further support the idea that attention mediates neural adaptation to talkers, as well. Correspondingly, studies have consistently reported reduced neural responses to the speech of a single, consistent talker compared to mixed or changing talkers (Belin & Zatorre, 2003; Chandrasekaran, Chan, & Wong, 2011; Perrachione et al., 2016), and some have speculated that stream discontinuity from changing talkers may incur a continuous attentional cost reflected in activation of structures supporting domain-general attention (Wong et al., 2004). Indeed, Zhang et al. (2016) reported that a talker change induced a reduction in the P300—an electrophysiological marker of attention—when subjects performed a phonetic task without explicitly attending to talker identity. This provides further evidence that adaptation to a talker is related to more efficient allocation of auditory attention. In this vein, systems neuroscience studies have also shown that neural representations of sounds are enhanced by prior expectation and attention in animals over short timescales (e.g., Jaramillo & Zador, 2011; Fritz, Elhilali, David, & Shamma, 2007; Zhou, de Villers-Sidani, Panizzutti, & Merzenich, 2010). The informational content of neural responses also rapidly becomes attuned to the spectrotemporal structure of an attended talker and suppresses the speech of unattended talkers (Ding & Simon, 2012; Mesgarani & Chang, 2012; Zion Golumbic, Poeppel, & Schroeder, 2012), with such neural tracking of attended speech improving over the course of a single sentence (Zion Golumbic et al., 2013). These results, indicating a temporal evolution of talker-specific tuning, are consistent with the findings from our study that talker adaptation unfolds with continued stimulation over time. Taken together, neural studies of humans and animals provide converging lines of evidence for a model in which talker adaptation in speech processing occurs as the auditory system forms a continuous auditory object via effective allocation of attention.

A streaming/attention model of talker adaptation also provides testable, falsifiable predictions about when and how talker adaptation is likely to occur. From the assumption that talker adaptation depends on attentional allocation to a continuous auditory object follows the prediction that disruption of the attention will disproportionately reduce or eliminate the processing gains afforded by talker adaptation in mixed-talker contexts compared to single-talker ones. For instance, a brief attentional disruption when listening to a single, continuous talker might incur the same inefficiencies in speech perception as listening to mixed-talker speech. Likewise, an increase in cognitive load by adding secondary tasks (e.g., Fairnie, Moore, & Remington, 2016) will reduce the amount of attentional resources that can be allocated to talker-specific speech processing and thus may have a disproportionately deleterious effect on speech processing in single-talker contexts compared to mixed-talker ones.

5.3. Limitations and directions for future work

Across three experiments, we parametrically varied the length, content, and contiguity of speech context preceding target words to investigate how context facilitates speech processing. The pattern of results across these three experiments both sheds light on the temporal and informational factors underlying talker adaptation and emphasizes the potential contributions of domain-general attention and auditory streaming in talker adaptation. However, considerable future work remains to both replicate and extend the predictions made by this framework. In particular, our observations are based on a limited set of carefully-controlled laboratory stimuli – two words spoken by four talkers in the absence of any auditory distractor. While we chose these particular stimuli to optimize the processing interference from multiple talkers (Choi et al., 2018), it will be important to show that these results generalize to contrasts that are less confusable across talkers. Furthermore, the repetitious identification of either of two target words is a *de minimis* case of speech perception, whereas real world utterances are highly heterogeneous and depend on a larger variety of contextual cues. Likewise, auditory streaming has traditionally been explored in contexts where multiple sound sources compete for attention and perceptual organization, whereas the present experiments involved multiple sequential, rather than simultaneous, sound sources. While the suggestion that talker adaptation involves feedforward auditory streaming also offers an opportunity to bridge previously disparate work on talker variability and source selection in speech processing, a model approaching speech adaptation as a process of building auditory objects will need to be further studied in more complex auditory scenes. Future work involving open-set stimuli, more ecological utterances, and real-world listening environments – such as conversations – will be needed to better understand how talker adaptation entails auditory attention. Finally, a major requirement of future work will be to reconcile the predictions and results of processes-based measures of talker adaptation (e.g., differences in response time to single vs. mixedtalkers with vs. without carrier phrases) with outcome-based measurements (e.g., differences in perceptual biases for ambiguous vowels or consonants based on contextual information, as in Johnson, 1990).

6. Conclusions

The results from this study show that speech processing is made more efficient via the perceptual adaptation to a talker arising from preceding speech context. The pattern of results suggests that talker

adaptation is facilitated by exposure to preceding speech from a talker that is brief (but not too brief), that is temporally continuous with the target speech, and that needs contain only minimal phonetic content. Together, these patterns of temporal and informational effects on talker adaptation raise the possibility that the efficiency gains in speech perception associated with talker adaptation may reflect the successful allocation of auditory attention to a target auditory object (i.e., a talker).

7. Open-source dataset

The complete set of stimuli, paradigm scripts, data, and data analysis scripts associated with these studies are available for download from our institutional archive: https://open.bu.edu/handle/2144/16460.

Acknowledgments

We thank Sara Dougherty, Elly Hu, Emily Thurston, Terri Scott, and Lauren Gustainis for their assistance, and Sung-Joo Lim and Barbara Shinn-Cunningham for helpful discussion. Research reported in this article was supported by the NIDCD of the National Institutes of Health under award number R03DC014045. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2019.05.019.

References

Alain, C., & Arnott, S. R. (2000). Selectively attending to auditory objects. *Frontiers in Bioscience, 5*, D202–D212.
Alho, K., Rinne, T., Herron, T. J., & Woods, D. L. (2014). Stimulus-dependent activations and attention-related modulations in the auditory cortex: A meta-analysis of fMRI studies. *Hearing Research, 307*, 29–41.
Altmann, C. F., Henning, M., Döring, M. K., & Kaiser, J. (2008). Effects of feature-selective attention on auditory pattern and location processing. *NeuroImage, 41*(1), 69–79.
Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual, and acoustic aspects. *The Journal of the Acoustical Society of America, 71*(4), 975–989.
Bachorowski, J.-A., & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *The Journal of the Acoustical Society of America, 106*(2), 1054.
Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*, 255–278.
Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *NeuroReport, 14*(16), 2105–2109.
Best, V., Ozmeral, E. J., Kopčo, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences, 105*(35), 13174–13178.
Best, V., Ozmeral, E. J., & Shinn-Cunningham, B. G. (2007). Visually-guided attention enhances target identification in a complex auditory scene. *Journal for the Association for Research in Otolaryngology, 8*(2), 294–304.
Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*, 341–345.
Bressler, S., Masud, S., Bharadwaj, H., & Shinn-Cunningham, B. (2014). Bottom-up influences of voice continuity in focusing selective auditory attention. *Psychological Research Psychologische Forschung, 78*(3), 349–360.
Chandrasekaran, B., Chan, A. H. D., & 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*(3), 784–797.
Cusack, R., Decks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology: Human Perception and Performance, 30*(4), 643–656.
Cutler, A., Andics, A., & Fang, Z. (2011). Inter-dependent categorization of voices and segments. In 17th meeting of the International Congress of Phonetic Sciences, Hong Kong.
Da Costa, S., van der Zwaag, W., Miller, L. M., Clarke, S., & Saenz, M. (2013). Tuning in to sound: Frequency-selective attentional filter in human primary auditory cortex. *Journal of Neuroscience, 33*(5), 1858–1863.
Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception: Adjusting the signal or the representations? *Cognition, 108*(3), 710–718.
Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences, 109*(29), 11854–11859.
Duncan, J. (2006). EPS Mid-Career Award 2004: Brain mechanisms of attention. *The Quarterly Journal of Experimental Psychology, 59*(1), 2–27.
Fairnie, J., Moore, B. C., & Remington, A. (2016). Missing a trick: Auditory load modulates conscious awareness in audition. *Journal of experimental psychology: Human perception and performance, 42*(7), 930.
Francis, A. L., Ciocca, V., Wong, N. K. Y., Leung, W. H. Y., & Chu, P. C. Y. (2006). Extrinsic context affects perceptual normalization of lexical tone. *The Journal of the Acoustical Society of America, 119*(3), 1712–1726.
Fritz, J., Shamma, S., Elhilali, M., & Klein, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nature Neuroscience, 6*(11), 1216.
Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention — focusing the searchlight on sound. *Current Opinion in Neurobiology, 17*(4), 437–455.
Froemke, R. C., & Schreiner, C. E. (2015). Synaptic plasticity as a cortical coding scheme. *Current opinion in neurobiology, 35*, 185–199.
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.
Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251.
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.
Jääskeläinen, I. P., Ahveninen, J., Belliveau, J. W., Raij, T., & Sams, M. (2007). Short-term plasticity in auditory cognition. *Trends in Neurosciences, 30*(12), 653–661.
Jaramillo, S., & Zador, A. M. (2011). The auditory cortex mediates the perceptual effects of acoustic temporal expectation. *Nature Neuroscience, 14*(2), 246.
Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni, & R. Remez (Eds.). *The handbook of speech perception* (pp. 363–389). Malden, MA: Blackwell Publishers.
Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *The Journal of the Acoustical Society of America, 88*(2), 642–654.
Johnson, K., Ladefoged, P., & Lindau, M. (1993). Individual differences in vowel production. *The Journal of the Acoustical Society of America, 94*(2), 701–714.
Joos, M. (1948). Acoustic phonetics. *Language Monographs, 23*, 136.
Kaganovich, N., Francis, A. L., & Melara, R. D. (2006). Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. *Brain Research, 1114*, 161–172.
Kidd, G., Jr, Arbogast, T. L., Mason, C. R., & Gallun, F. J. (2005). The advantage of knowing where to listen. *The Journal of the Acoustical Society of America, 118*(6), 3804–3815.
Kleinschmidt, D. F. (2018). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience,* 1–26.
Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review, 122*(2), 148.
Kraljic, T., & Samuel, A. G. (2007). Perceptual adjustments to multiple speakers. *Journal of Memory and Language, 56*(1), 1–15.
Ladefoged, & Broadbent (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America, 29*, 98–104.
Laing, E. J., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a tune: Talker normalization via general auditory processes. *Frontiers in Psychology, 3*, 203.
Leather, J. (1983). Speaker normalization in perception of lexical tone. *Journal of Phonetics*.
Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review, 74*(6), 431.
Lim, S.-J., Shinn-Cunningham, B.G., & Perrachione, T.K. (in press). Effects of talker continuity and speech rate on auditory working memory. Attention, Perception, and Psychophysics. https://doi.org/10.3758/s13414-019-01684-w.
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.
Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information.* New York, NY: Henry Holt and Co., Inc.
McLennan, C. T., & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology-Learning, Memory, & Cognition, 31*, 306–321.
Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature, 485*(7397), 233.
Morton, J. R., Sommers, M. S., & Lulich, S. M. (2015). The effect of exposure to a single vowel on talker normalization for vowels. *The Journal of the Acoustical Society of America, 137*(3), 1443–1451.
Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9*(5–6), 453–467.
Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47*(4), 379–390.
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. Johnson, & J. W. Mullennix (Eds.). *Talker variability in speech processing* (pp. 109–132). San Diego, CA: Academic Press.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In Y. Tohkura, Y. Sagisaka, & E. Vatikiotis-Bateson (Eds.). *Speech perception, production, and linguistic structure* (pp. 113–134). Tokyo: Ohmasha Publishing.
Peirce, J. W. (2007). PsychoPy: Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8–13.
Perrachione, T. K., Del Tufo, S. N., Winter, R., Murtagh, J., Cyr, A., Chang, P., ... Gabrieli, J. D. E. (2016). Dysfunction of rapid neural adaptation in dyslexia. *Neuron, 92*, 1383–1397.
Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics, 2*, 33–52.
Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson, & J. W. Mullennix (Eds.). *Talker variability in speech processing* (pp. 9–32). San Diego, CA: Academic Press.
Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences, 12*(5), 182–186.
Shinn-Cunningham, B. G., & Best, V. (2008). Selective attention in normal and impaired hearing. *Trends in Amplification, 12*(4), 283–299.
Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011). Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics. *Neuropsychologia, 49*, 3831–3846.
Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences, 13*(9), 403–409.
Todorovic, A., & de Lange, F. P. (2012). Repetition suppression and expectation suppression are dissociable in time in early auditory evoked fields. *Journal of Neuroscience, 32*(39), 13389–13395.
Trude, A. M., & Brown-Schmidt, S. (2012). Talker-specific perceptual adaptation during online speech perception. *Language and Cognitive Processes, 27*(7–8), 979–1001.
Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: Predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences, 13*(12).
Woods, K. J. P., & McDermott, J. H. (2015). Attentive tracking of sound sources. *Current Biology, 25*(17), 2238–2246.
Wong, P. C., & Diehl, R. L. (2003). Perceptual normalization for inter-and intratalker variation in Cantonese level tones. *Journal of Speech, Language, and Hearing Research, 46*(2), 413–421.
Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience, 16*(7), 1173–1184.
Xie, X., Earle, F. S., & Myers, E. B. (2018). Sleep facilitates generalisation of accent adaptation to a new talker. *Language, Cognition and Neuroscience, 33*(2), 196–210.
Zhang, C., Pugh, K. R., Mencl, W. E., Molfese, P. J., Frost, S. J., Magnuson, J. S., ... Wang, W. S.-Y. (2016). Functionally integrated neural processing of linguistic and talker information: An event-related fMRI and ERP study. *Neuroimage, 124*, 536–549.
Zhou, X., de Villers-Sidani, E., Panizzutti, R., & Merzenich, M. M. (2010). Successive-signal biasing for a learned sound sequence. *Proceedings of the National Academy of Sciences of the United States of America, 107*(33), 14839–14844.
Zion Golumbic, E. M., Poeppel, D., & Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective. *Brain and language, 122*(3), 151–161.
Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., ... Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a 'cocktail party'. *Neuron, 77*(5), 980–991.