Time and information in perceptual adaptation to speech

Title:
Time and information in perceptual adaptation to speech

Authors:
Ja Young Choi[1,2] & Tyler K. Perrachione[1]*

[1]Department of Speech, Language, and Hearing Sciences
Boston University
Boston, Massachusetts

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

***Correspondence:**
Tyler K. Perrachione, Ph.D.
Department of Speech, Language, and Hearing Sciences, Boston University
635 Commonwealth Ave.
Boston, MA 02215
Phone: +1.617.358.7410
Email: tkp@bu.edu

Word Count:	10057 words
Number of figures:	7 figures
Number of tables:	3 tables
Open source dataset:	https://open.bu.edu/handle/2144/16460

Abstract

Perceptual adaptation to a talker enables listeners to resolve the many-to-many mapping between variable speech acoustics and abstract linguistic representations more efficiently. 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 – but not the amount of talker-specific phonetic information it contained – increased. 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.

Keywords: speech perception; phonetic variability; categorization; talker normalization; adaptation

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 et al., 1967). Talkers differ in their vocal tract anatomy, dialect and speech mannerisms (Johnson et al., 1993), resulting in different talkers using remarkably different acoustics to produce the same phoneme, or virtually identical acoustics to produce different phonemes (Hillenbrand et al., 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, 1997).

Recent models of speech processing have made forays into describing how the speech perception system might resolve acoustic-phonetic ambiguity across talkers while maintaining a stable phonology. 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 only accounts for 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, or 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 talker adaptation in speech processing still lacks an implementational description of how the system (*i*) 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 indexical and phonetic context of speech to facilitate its decision process. In this paper, we provide an empirical foundation that describes 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.

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 et al., 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 of a speech stream affects speech processing have focused on the perceptual decision *outcomes* of that processing (Ladefoged & Boradbent, 1957; Johnson, 1990; Leather, 1983; Francis et al., 2006). However, none have yet examined how the indexical context affects the *efficiency* with which listeners process talkers' speech – the hallmark effect of perceptual adaptation – nor have they considered how much or what kind of information listeners obtain from preceding contexts in order to maximize the efficiency of their perceptual decisions. How much time does it take, then, for a listener to become adapted to a talker's speech in this kind of talker-switching context? And what kinds of details about talkers' speech do listeners need in order to more efficiently map the acoustic-phonemic composition of upcoming speech sounds?

Neuroimaging studies have shown that adaptation to talker-specific speech is associated with reduced physiological cost (Wong, Nusbaum & Small, 2004; Zhang et al., 2016; Perrachione 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 et al., 2006; Sjerps et al., 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 et al., 2015; Fritz et al., 2003; Jääskeläinen et al., 2007; Winkler et al., 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 are silent as to how and when relevant information about the target talker's speech is ascertained during speech perception (Pierrehumbert, 2016; Kleinschmidt & Jaeger, 2015).

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 /o/ and /u/ (Hillenbrand et al., 1995; Choi, Hu, & Perrachione, 2018). 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 showing reduction in the interference effect as the combined length and information of the carrier phrase increased. In Experiment 2, we examined how the quantity 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, revealing 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, revealing 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, Hu, & Perrachione, 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 counter-balanced design across conditions (see below). Previous research using this phonemic contrast in a similar behavioral paradigm (Choi, Hu, & Perrachione, 2018) found that processing speech from mixed vs. single talkers incurs a processing cost of +126ms (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 50ms (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 and therefore highly susceptible to interference in a mixed-talker setting (Choi, Hu, & Perrachione, 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,D**).

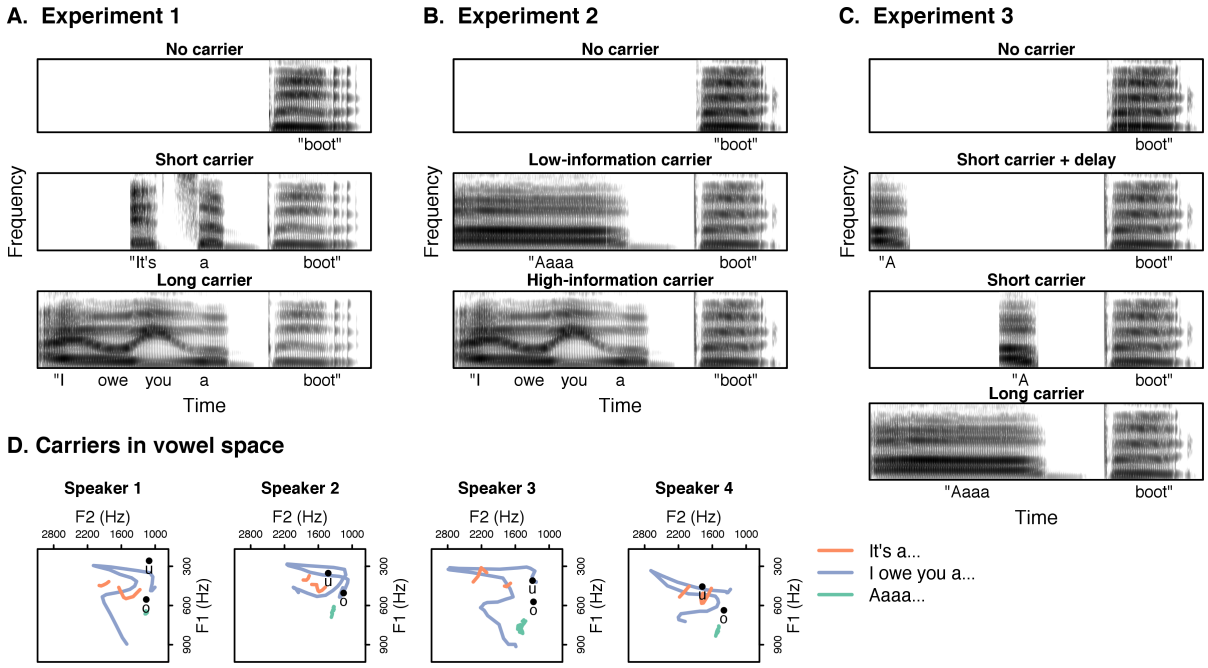

Figure 1: Stimuli for Experiments 1-3. (A,B,C) Spectrograms of example stimuli produced by Speaker 2 used in Experiments 1-3 in each condition. **(D)** Lines indicate the F1-F2 trajectory of all carriers produced by each talker. Black points indicate the F1-F2 position of the /o/ and the /u/ vowels in 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

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 44.1kHz 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. Pitch and amplitude of the carrier phrase and the target word, as well as the voice-onset time between the end of the carrier phrase on the onset of target word were manipulated so that the concatenated sentences were indistinguishable from natural speech. 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 sequential trials. The order of conditions was counter-balanced across participants using Latin square permutations.

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 2000ms. Stimulus delivery was controlled using PsychoPy v.1.8.1 (Peirce, 2007).

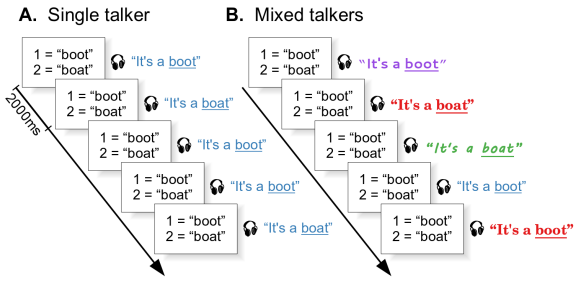

Figure 2: Task design for all experiments. Participants performed a speeded word identification task while listening to speech produced by either **(A)** a single talker or **(B)** mixed talkers. The short-carrier condition for Experiment 1 is shown.

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 more closely approximate a normal distribution, as expected by the models. 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 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 et al., 2013). Significance of main

effects and interactions was determined by adopting significance criterion of $\alpha = 0.05$, with *p*-values for model terms based on the Satterthwaite approximation of the degrees of freedom obtained from the package *lmerTest*.

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

2.2. Results

Participants' word identification accuracy was at ceiling (mean = 98% ± 2%). Consequently, the dependent measure for this experiment was response time (**Table 1**), as is usual for studies of perceptual adaption in speech perception (e.g., Choi, Hu & Perrachione, 2018; Magnuson & Nusbaum, 2007; McLennan & Luce, 2005). Participants' response times in each condition are shown in **Figure 3**.

Compared to the single-talker conditions, response times in the mixed-talker conditions were significantly slower overall (single 679 ms vs. mixed 747 ms; $\beta = 0.12$, *s.e.* = 0.011, $t = 11.84$, $p < 3.7 \times 10^{-11}$). For each of the three carrier conditions independently, we observed significantly faster response times in the single-talker condition than in the mixed-talker condition (**Table 1**): *no carrier* single-talker 698 ms *vs*. mixed-talker 792 ms ($\beta = 0.12$, *s.e.* = 0.017, $t = 7.47$, $p < 1.4 \times 10^{-7}$); *short-carrier* single-talker 666 ms *vs*. mixed-talker 736 ms ($\beta = 0.096$, *s.e.* = 0.015, $t = 6.28$, $p < 2.1 \times 10^{-6}$); long-*carrier* single-talker 672 ms *vs*. mixed-talker 711 ms ($\beta = 0.050$, *s.e.* = 0.013, $t = 3.91$, $p < 7.2 \times 10^{-4}$). These results indicate that listening to speech in a mixed talker context had a consistent, deleterious effect on listeners' ability to make

perceptual decisions about speech content, even when target speech was preceded by additional talker-specific phonetic information.

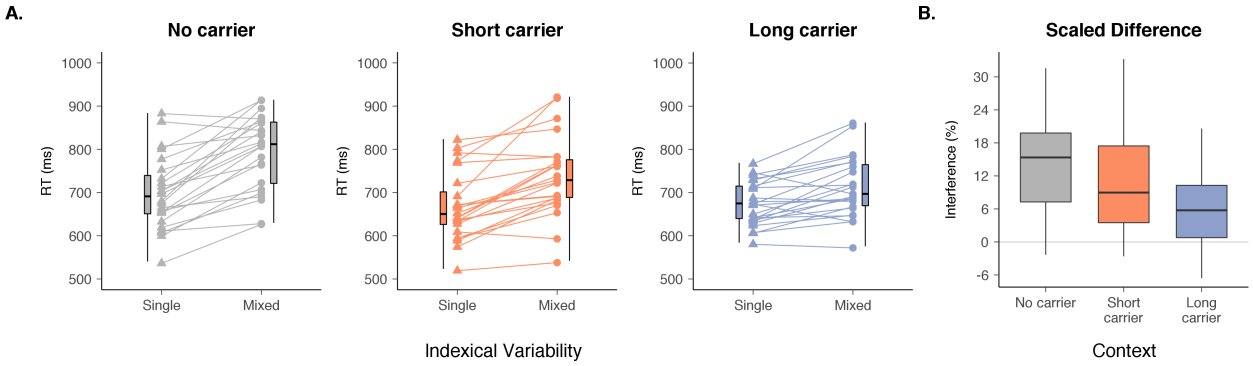

Figure 3: Results for Experiment 1. Effects of talker variability and context across talkers on response times. **(A)** Connected points show the change in response times for individual participants between the single- and mixed-talker conditions across three levels of context. Box plots in each panel 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.

In a model including all three carriers simultaneously, significant carrier × variability interactions were observed, indicating that the magnitude of perceptual adaptation between the single- and mixed-talker conditions differed depending on the type of carrier phrase that preceded the target word. Listeners exhibited significantly more interference from the mixed-talker condition (versus the single-talker condition) in the *no-carrier* condition (+95 ms / 14%) than in either the *short-carrier* (+70 ms / 11%; $\beta = 0.045$, *s.e.* = 0.010, $t = 4.52$, $p < 0.0002$) or *long-carrier* (+40 ms / 6%; $\beta = 0.074$, *s.e.* = 0.010, $t = 7.37$, $p < 1.7 \times 10^{-7}$) conditions. Likewise, the amount of interference listeners experienced in the *short-carrier* condition was significantly greater than in the *long-carrier* one ($\beta = 0.029$, *s.e.* = 0.010, $t = 7.37$, $p < 0.01$). 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 immediately preceding connected speech from a talker reduces the processing cost associated with speech perception in a multi-talker context. This result provides a temporal, process-based explanation for prior reports that the outcomes of perceptual decisions in speech are affected by preceding speech context (Johnson, 1990; Laing et al., 2012). We also 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 ~300ms speech carrier reduces the processing cost of speech perception in a multi-talker context from 14% to 11%, and a longer, ~600ms 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 listeners 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, there are two unresolved possibilities for why the longer carrier afforded greater perceptual adaptation to speech. In Experiment 1, the long and short carrier conditions differed in two ways. First, the two carriers had different total durations: The average duration of the short carrier phrase ("It's a ...") was 340ms, whereas that of the long carrier phrase ("I owe you a ...") was 615ms. Second, they contained different amount of talker-specific phonetic information the adapting speech revealed

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/, /i/, /o/, /u/, /ʌ/) and effectively sampled the entire vowel space (**Fig. 1A,D**). That is, the long carrier not only contained more talker-specific detail about his/her speech production, but it also 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 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,D**): a *high-information* carrier phrase contained a richer amount of information that reveals the extent of each talker's vowel space, whereas a *low-information* carrier phrase revealed talkers' 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 high-information carrier phrase should result in a greater reduction of the interference effect of

mixed-talker speech the low-information carrier (**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**).

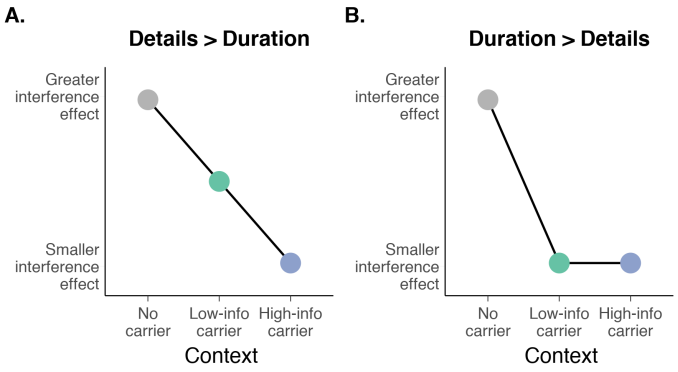

Figure 4: Hypothesized patterns of results for Experiment 2. Potential patterns for the interference effect of talker variability across the three experimental conditions, as predicted by the two different hypotheses about contextual effects on talker adaptation. **(A)** If the amount of talker-specific phonetic details in a carrier contributes more to talker adaptation than the duration of the carrier, the interference effect will be greater in the low-information carrier condition than in the high-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-information and the high-information carriers, as their durations are matched.

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 30ms (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 2000ms. 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.

3.1.4. Data Analysis

As in Experiment 1, 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 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).

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

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 **Figure 5**.

As in Experiment 1, response times in the mixed-talker conditions were significantly slower than those in the single-talker conditions overall (single 682 ms *vs*. mixed 732 ms; β = 0.046, *s.e.* = 0.010, t = 4.61, $p < 0.0002$). For all three speech context conditions independently, we again observed significantly faster response times in the single-talker condition than in the mixed-talker condition (**Table 2**): *no carrier* single-talker 705 ms *vs*. mixed-talker 784 ms (β = 0.11, *s.e.* = 0.014, t = 7.47, $p < 1.36 \times 10^{-7}$); *low-information carrier* single-talker 679 ms *vs*. mixed-talker 716 ms (β = 0.059, *s.e.* = 0.012, t = 4.28, $p < 2.9 \times 10^{-4}$); *high-information carrier* single-talker 662 ms *vs*. mixed-talker 697 ms (β = 0.046, *s.e.* = 0.015, t = 3.18, $p < 4.2 \times 10^{-3}$). Like in Experiment 1, listening to speech in all mixed-talker contexts in Experiment 2 had deleterious effect on listeners' ability to make perceptual decisions about speech content, even when preceded by talker-specific phonetic information from the carriers.

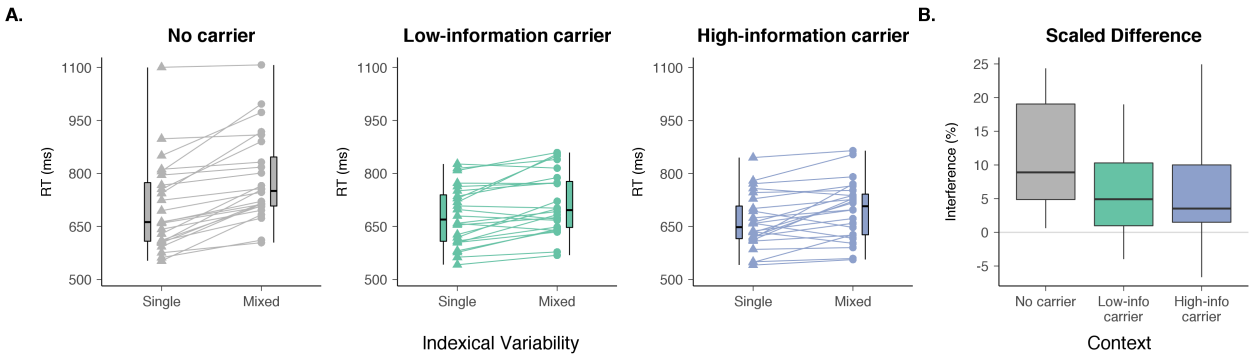

Figure 5: Results for Experiment 2. Effects of talker variability and context across talkers on response times. **(A)** Connected points show the change in response times for individual participants between the single- and mixed-talker conditions across three levels of context. Box plots in each panel 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. 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**).

In a model including all three speech contexts simultaneously, we observed an interesting – and surprising – pattern of significant context × variability interactions, indicating different effects on the magnitude of perceptual adaptation between the single- and mixed-talker conditions across speech contexts in Experiment 2. Listeners exhibited significantly more interference from the mixed-talker condition (versus the single-talker condition) in the *no-carrier* condition (+79 ms / 12%) than in both the *low-information* (+37 ms / 6%; $\beta = 0.029$, *s.e.* = 0.010, $t = 7.37$, $p < 0.01$) and *high-information* (+35 ms / 5%; $\beta = 0.074$, *s.e.* = 0.010, $t = 7.37$, $p < 1.7 \times 10^{-7}$) carrier conditions. Surprisingly, however, the amount of interference between the *low-* and *high-information* carriers was essentially identical ($\beta = 0.0063$, *s.e.* = 0.0094, $t = 0.67$, $p = 0.51$). This pattern of results replicates the observation from Experiment 1 that speech context facilitates the 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 the second experiment 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 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, 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 talker-specific representation of vowel space and use it to understand their 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. Inter-talker variability in the acoustic realization of speech is not completely random but rather structured regarding talkers' socio-indexical characteristics (Kleinschmidt, 2018), which may contribute to how talker-specific cues with minimal phonetic information sufficiently facilitate talker adaptation.

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 in 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 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" (/ʌ:/) 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, since the onset of speech in these conditions occurs equidistant from the target lexical item. Correspondingly, both the long-carrier and short-carrier-

with-delay conditions should offer greater facilitation than the short-carrier-without delay, in which speech onset occurs closer to the target word and affords less time for activation and integration of talker-specific memories (**Fig. 6A**). Alternatively, if talker adaptation depends on attentional reorientation via auditory streaming, 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.

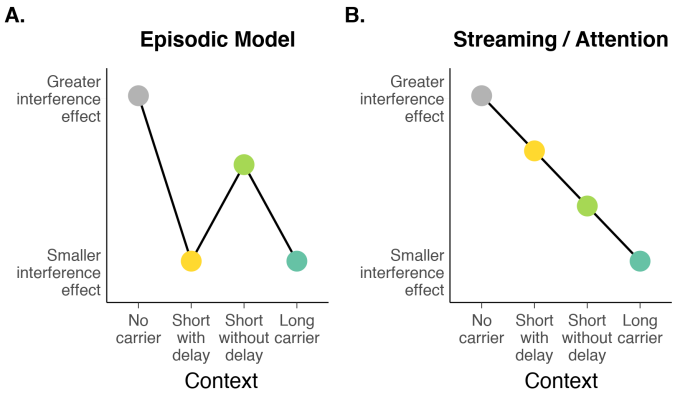

Figure 6: Hypothesized patterns of results for Experiment 3. Potential patterns for the interference effect of talker variability across the four experimental conditions, as predicted by the two different hypotheses of the contribution of temporal continuity of context. **(A)** In an episodic account of speech perception, due to the time available to reactivate talker-specific memories, such a model predicts a greater interference effect in the short-carrier-without-delay condition than either the short-carrier-with-delay condition or the long-carrier condition, which should be equally facilitatory. **(B)** In contrast, an attention/streaming model of speech perception 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, due to the ease in developing a talker-specific auditory object resulting from temporal proximity between the adapting speech and target word.

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 Experiments 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 their voiced duration to 20% of that of the long-carrier (average 215 ms). For the short-duration carriers with an intervening pause, the duration of the pause was calculated so that the duration of each speaker's short carrier plus the 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.

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 2000ms. 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).

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

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 **Figure 7**.

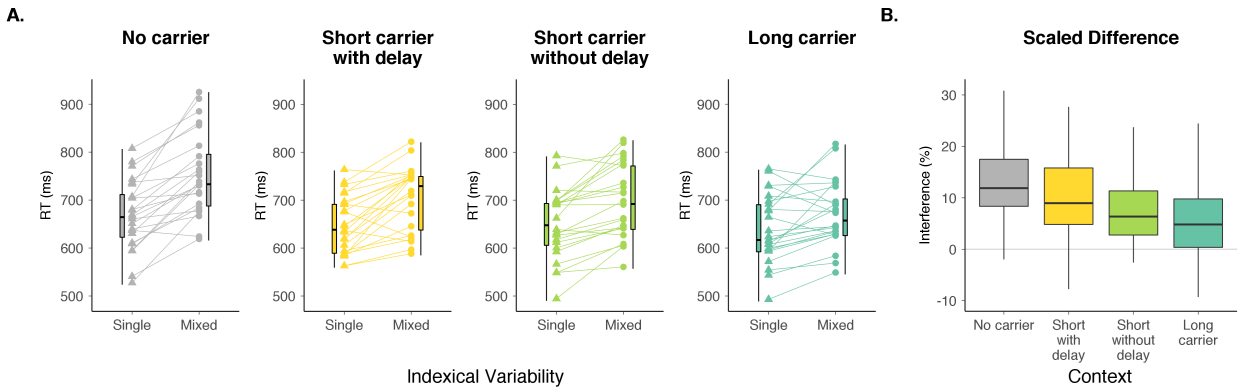

Figure 7: Results for Experiment 3. Effects of talker variability and context across talkers on response times. **(A)** Connected points show the change in response times for individual participants between the single- and mixed-talker conditions across four levels of context. Box plots in each panel 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 result is consistent with what the streaming/attention model predicts (**Fig. 6B**).

Compared to the single-talker conditions, response times in the mixed-talker conditions were significantly slower overall (single 652 ms vs. mixed 707 ms; $\beta = 0.047$, *s.e.* = 0.011, $t = 4.42$, $p < 6.2 \times 10^{-5}$). For all four carrier conditions independently, we observed significantly faster response times in the single-talker condition than in the mixed-talker condition (**Table 3**):

no carrier single-talker 670 ms *vs*. mixed-talker 754 ms ($\beta = 0.11$, *s.e.* = 0.014, $t = 7.66$, $p < 8.9 \times 10^{-8}$); *short-carrier with delay* single-talker 649 ms *vs*. mixed-talker 706 ms ($\beta = 0.084$, *s.e.* = 0.015, $t = 5.43$, $p < 1.7 \times 10^{-5}$); *short-carrier without delay* single-talker 651 ms *vs*. mixed-talker 698 ms ($\beta = 0.065$, *s.e.* = 0.013, $t = 5.10$, $p < 3.7 \times 10^{-5}$); *long-carrier* single-talker 640 ms *vs*. mixed-talker 671 ms ($\beta = 0.048$, *s.e.* = 0.016, $t = 2.99$, $p < 6.6 \times 10^{-3}$). 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.

The model including all four carriers simultaneously revealed a telling pattern of significant carrier × variability interactions, providing key insight into how the magnitude of perceptual adaptation between the single- and mixed-talker conditions is affected by the length and proximity of the adapting carrier phrase preceding the target word. Listeners exhibited significantly more interference from the mixed-talker condition (versus the single-talker condition) in the *no-carrier* condition (+84 ms / 12%) than in the *short-carrier with delay* (+57 ms / 9%; $\beta = 0.028$, *s.e.* = 9.5×10^{-3}, $t = 2.94$, $p < 3.4\times 10^{-3}$), *short-carrier without delay* (+47 ms / 7%; $\beta = 0.046$, *s.e.* = 9.4×10^{-3}, $t = 4.85$, $p < 1.3\times 10^{-6}$), or *long-carrier* (+31 ms / 5%; $\beta = 0.064$, *s.e.* = 9.4×10^{-3}, $t = 6.74$, $p < 1.7\times 10^{-11}$) conditions.

Interference was significantly greater in the *short-carrier with delay* condition than the *long-carrier* condition ($\beta = 0.036$, *s.e.* = 9.5×10^{-3}, $t = 3.79$, $p < 1.6 \times 10^{-4}$). The difference in interference between the two short-carrier conditions trended towards greater interference in the *short-carrier-with-delay* than *short-carrier without delay* condition ($\beta = 0.018$, *s.e.* = 9.5×10^{-3}, $t = 1.90$, $p = 0.057$). Finally, the difference in interference between the *short-carrier without delay*

condition and *long-carrier* condition was marginally significant and trended towards greater interference in the *short-carrier* condition ($\beta = 0.018$, *s.e.* $= 9.5 \times 10^{-3}$, $t = 1.89$, $p = 0.059$).

4.3. Discussion

The results from the third experiment are consistent with the predictions made by an object continuity/streaming model of talker adaptation, but inconsistent with those made by a mnemonic/ballistic model. The 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 encountered (**Fig. 6A**). A model of talker adaptation that depends on episodic access predicts an equally large facilitation yielded by the short carrier with delay and by the long carrier, and more adaptation to a short carrier with delay than one without. This is the opposite of what we found in this experiment; 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 et al., 2008; Bressler et al., 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 on talker adaptation when compared to either a time-matched continuous signal or a quantity-matched adjacent signal. The long-carrier condition 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 they 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 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 is 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 all 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 established that the perceptual system incorporates preceding speech context not only to bias the perceptual outcomes of speech perception (e.g., Johnson, 1990), but also to make speech perception more efficient. Moreover, based on the findings from Experiment 1, we found that the interference effect of processing speech from multiple talkers was reduced as a function of the

amount of preceding speech context from each talker, even for as little as 300-600ms of preceding information.

Interestingly, in Experiment 2, we observed that prior speech context comprising 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 can be ascribed 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 real time. The results from Experiment 3 revealed that it is not simply the time preceding the target speech but rather 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 comprehensive empirical foundation for an implementational-level understanding of how perceptual adaptation to speech occurs in real time. Further, when evaluated in the context of two potential theoretical frameworks for explaining the pattern of efficiency gains in perceptual adaptation to speech, these results convergently lend support to a model of speech adaptation that bears striking similarity to domain-general attentional processes for auditory object-continuity and streaming.

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 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, and the mechanisms by which this efficiency gain are obtained appear to be the same as those involved in allocating attention in perceptual streaming, namely, the duration and temporal continuity of the preceding content. Interestingly (and to us, 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 et al., 2014; Choi, Hu, & Perrachione, 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 to successfully orient listeners' attention to a target auditory stream (talker) and reduce 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).

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, 1983) 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, Tomiak, & Kuhl, 1997; Mullennix & Pisoni, 1990; Morton, Sommers, & Lulich, 2015; Choi, Hu, & Perrachione, 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, when listeners in our study had access to preceding context in mixed-talker conditions, they would have been able to engage this active control mechanism as soon as they encounter a carrier spoken by a new talker. Because the carrier already activated this process, when listeners encounter the target word, the perceptual system does not need to expend as many cognitive resources to map the incoming acoustics to the intended linguistic representation. However, the present results go further in identifying the likely mechanism underlying this control process and therefore refining the theoretical framework under which talker adaptation can be understood; namely, the cognitive process effecting efficiency gains in speech perception appears to be the successful allocation of attention for auditory streaming and auditory object formation. Just as the evidence from Experiment 3 is at odds with a mnemonic/ballistic model of talker normalization (cf. Goldinger, 1998), so too does the observation 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 suggest that any active control process needs to operate over a sustained, temporally continuous auditory signal. The operationalization of this cognitive process as one of attentional allocation is further validated by the observation that the long-carrier provides no additional phonetic information compared to the short-carrier-with-delay, but still affords greater adaptation to the target talker. This demonstrates that an active control process cannot merely be building a sophisticated phonetic model of a talker's speech and/or vocal tract, but instead must be picking out (streaming) an 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 et al., 2005; Alain & Arnott, 2000).

An alternative account of talker-specific speech processing that is sometimes 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 be 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 & Owen, 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 explicitly reject a mnemonic account of talker adaptation-based efficiency gains in speech processing. Under an episodic account, we should expect the facilitation afforded by the short-carrier-with-delay and long-carrier conditions to be equal, 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 actually 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 revisit more recent, formal models of speech adaptation and extend them into the implementational level of explanation. The highly influential ideal adapter framework of Kleinschmidt & 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 is slower and less accurate, whereas when the number of models is smaller – such as when a listener can limit model selection to a single talker – the recognition of speech is 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). However, this model, although highly successful in its algorithmic-level account of speech processing, is limited in that it does not consider the implementational level – i.e., it does not specify what kinds of information that the perceptual system needs in order to choose the correct model nor, critically, how the perceptual system incorporates such cues over time, which are crucial aspects of how a biopsychological system achieves a computational process. The present study provides an empirical and theoretical framework for understanding the implementational-level mechanisms short-term perceptual adaptation to a talker's speech: Namely, by showing how talker adaptation unfolds over time, these results implicate the that the active cognitive process of adapting to the talker that unfolds over time is likely the efficient allocation of auditory attention involved in streaming / object formation as the active cognitive process underlying 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 identify a mechanism by which talker-specific information is continuously integrated over time to improve perception. Such a mechanism is readily available in the domain of auditory scene analysis as the attentional selection of auditory objects 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., 2015; 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 et al., 2004). In the short-carrier-with-delay condition of Experiment 3, the delay between the carrier and the target word interrupts 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 than the other carrier phrases which were temporally continuous with the target speech.

Findings from neuroimaging studies on perception and attention provide converging evidence for talker adaptation as an efficiency gain resulting from attentional allocation. Prior expectation modulates the magnitude of neural adaptation to repeated stimuli (Summerfield & Egner, 2009; Todorovic et al., 2011), and auditory feature-specific attention affects neurophysiological adaptation, as measured by fMRI (Altmann et al., 2008; Alho et al., 2014; Da Costa et al., 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 (Wong, Nusbaum, & Small, 2004; Chandrasekaran, Chan, & Wong, 2011; Belin & Zatorre, 2003; Perrachione et al., 2016). 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 the result of more efficient allocation of auditory attention. Consistent with this account, 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 et al., 2007; Zhou et al., 2010). The informational content of neural responses also rapidly become attuned to the spectrotemporal structure of an attended talker and suppress the speech of unattended talkers (Zion Golumbic et al., 2012; Mesgarani & Chang, 2012; Ding & Simon, 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 consistently suggest that talker adaptation in speech processing is likely to occur as the auditory system forms a continuous auditory object via effective allocation of attention.

A further advantage of the streaming/attention model of talker adaptation over prior accounts of talker normalization is that this model 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 reduce or eliminate the processing gains afforded by talker adaptation. For instance, a brief attentional disruption when listening to a single, continuous talker is likely 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.

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. Moreover, the mechanistic (implementational) explanation for this adaptation appears to be the successful allocation of auditory attention that is facilitated by exposure across a sufficient duration to temporally continuous speech from a talker. Together, these data suggest that the efficiency gains in speech

perception associated with talker adaptation likely reflect the successful allocation of auditory attention to a target auditory object (i.e., a talker).

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.

References

Alain, C., & Arnott, S. R. (2000). Selectively attending to auditory objects. *Front. Biosci*, *5*, D202-D212.

Alain, C., Snyder, J. S., He, Y., & Reinke, K. S. (2006). Changes in auditory cortex parallel rapid perceptual learning. *Cerebral Cortex*, *17*(5), 1074-1084.

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*, *78*(3), 349–360.

Cai, S., Beal, D. S., Ghosh, S. S., Guenther, F. H., & Perkell, J. S. (2014). Impaired timing adjustments in response to time-varying auditory perturbation during connected speech production in persons who stutter. *Brain and language*, *129*, 24-29.

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. 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.

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How Do Expectations Shape Perception? *Trends in Cognitive Sciences*, *22*(9), 764–779.

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.

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

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 Pisoni, D.B. & Remez, R. (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.

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 Jr, G., 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.
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. (under review). "Effects of talker continuity and speech rate on auditory working memory."
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., Halverson, K., Ghosh, S.S., Christodoulou, J.A. & 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.

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.

Summerfield, C., Wyart, V., Johnen, V. M., & de Gardelle, V. (2011). Human Scalp Electroencephalography Reveals that Repetition Suppression Varies with Expectation. *Frontiers in Human Neuroscience*, *5*, 67.

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.

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.

Zhang, C., Pugh, K. R., Mencl, W. E., Molfese, P. J., Frost, S. J., Magnuson, J. S., Peng, G., & 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., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., & Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a 'cocktail party'. *Neuron*, *77*(5), 980–991.