

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

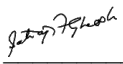


DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the
Division of Medical Sciences
Speech and Hearing Bioscience and Technology
have examined a dissertation entitled

*The Cognitive and Neural Bases of Processing Talker Variability in
Speech Perception*

presented by Ja Young Choi
candidate for the degree of Doctor of Philosophy and hereby
certify that it is worthy of acceptance.

Signature: 

Typed Name: Dr. Satrajit Ghosh

Signature: 
Julie Arenberg (May 4, 2022 12:33 EDT)

Typed Name: Dr. Julie Arenberg

Signature: 
Sofia Vallila Rohter (May 10, 2022 10:53 EDT)

Typed Name: Dr. Sofia Vallila Rohter

Signature: 
Lori Holt (May 4, 2022 11:13 EDT)

Typed Name: Dr. Lori Holt

Date: May 04, 2022

The Cognitive and Neural Bases of Processing Talker Variability in Speech Perception

A dissertation presented

by

Ja Young Choi

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Speech and Hearing Bioscience and Technology

Harvard University

Cambridge, Massachusetts

May 2022

© 2022 Ja Young Choi

All rights reserved.

The Cognitive and Neural Bases of Processing Talker Variability in Speech Perception

Abstract

Talker variability is the principal source of phonetic variability in speech signals, resulting in a lack of consistency in acoustic-to-phonetic mapping. Previous studies have repeatedly demonstrated that listeners incur additional processing costs in order to successfully extract phonetic information in the face of such significant variability across talkers. These costs manifest as lower accuracy and slower response time in speech perception tasks and increased neural response in auditory brain regions in mixed- relative to single-talker settings. However, it is unknown how talker adaptation processes unfold over time and how they interact with the amount and nature of information preceding the target speech. Moreover, neuroimaging results alone cannot establish causal roles for auditory regions in talker adaptation. The first series of behavioral experiments investigates the effect of preceding context on talker adaptation by comparing response times in auditory word identification tasks between single- and mixed-talker conditions while manipulating the following attributes of carrier speech preceding the target word: its duration, its amount of talker-specific phonetic detail, and its temporal continuity to the target word. Results indicate that the duration of the carrier, but not the richness of detail within it, has a significant effect on talker adaptation, and that temporal continuity between the context and the target word facilitates talker adaptation. Extending these findings, the next study examines how processing talker variability improves as a function of the duration of continuous speech from a talker. Results demonstrate that the facilitatory effect of immediately preceding

speech on talker adaptation linearly increases up to 600 ms, but longer exposure to continuous speech from a talker had no further facilitatory effect on processing in mixed-talker contexts. The last study investigates the causal role of left superior temporal gyrus (STG) in processing talker variability by applying high-definition transcranial direct current stimulation (HD-tDCS) to left STG while participants performed an auditory word identification task. Neurostimulation of left STG selectively decreased the facilitatory effect of immediately preceding context on talker adaptation. Discussed in light of models of speech perception, these studies together suggest a role for stimulus-driven auditory attention in behavioral phenomena known as talker adaptation.

Table of Contents

Title Page	i
Copyright	ii
Abstract	iii
Acknowledgments	vii
List of Tables and Figures	ix
Chapter 1. Introduction	1
1.1 Phonetic variability and speech perception	1
1.2 The processing costs of talker variability	2
1.3 Models of speech perception	3
1.4 Neural bases of talker adaptation	6
1.5 Current dissertation	7
Chapter 2. Time and information in talker adaptation	9
2.1 Introduction	9
2.2 Experiment 1: Perceptual adaptation to speech depends on preceding speech context	14
2.2.1 Methods	15
2.2.2 Results	19
2.2.3 Discussion	21
2.3 Experiment 2: Perceptual adaptation in high- and low-information contexts	23
2.3.1 Methods	24
2.3.2 Results	27
2.3.3 Discussion	29
2.4 Experiment 3: Effects of temporal proximity and duration in perceptual adaptation	31
2.4.1 Methods	34

2.4.2 Results	36
2.4.3 Discussion	39
2.5 General Discussion	40
Chapter 3. Two distinct mechanisms of processing talker variability	51
3.1 Introduction	51
3.2 Methods	55
3.3 Results	58
3.4 Discussion	61
Chapter 4. Neurostimulation of left temporal lobe disrupts rapid talker adaptation	66
4.1 Introduction	66
4.2 Methods	70
4.3 Results	75
4.4 Discussion	78
4.5 Conclusions	84
Chapter 5. Conclusions	85
5.1 Summary	85
5.2 Future directions	88
References	91

Acknowledgments

First and foremost, I thank my advisor, Tyler Perrachione, for mentoring me for all these years. Tyler welcomed me into this world of research and has encouraged me with his infectious optimism and enthusiasm at every step of the way. It has been an incredible privilege to pursue research with Tyler's guidance and support. I would also like to thank my Dissertation Advisory Committee — Satra Ghosh, Stefanie Shattuck-Hufnagel, and David Gow — for their invaluable time and insights. Discussions with them made my thoughts richer and wider and the process more exciting and fulfilling.

I've received continued support from SHBT professors — Gwen Géléoc, Bertrand Delgutte and Heidi Nakajima. They graciously lent me their time and support when I didn't know where I was headed or how to move forward.

I am lucky to have had all my colleagues in the Communication Neuroscience Research Laboratory, especially Terri Scott, Alex Kapadia, Gabrielle Torre, Sung-Joo Lim, Jessica Tin, Yaminah Carter, and Deirdre McLaughlin. They have been inspiring colleagues with whom I am excited to work and discuss science, and they've also been great friends who have shown me beautiful acts of friendship and kindness.

Also on my side throughout the graduate school journey has been my fellow cohort SHBT class of 2016 — Jeanne Gallée, Jan Iyer, Adrian Cho, John Lee and Steve McInturff. I am thankful to have learned so much with them and from them.

Looking back to how I started my graduate school career, I would like to thank the mentors that I met at Rice University, Tatiana Schnur and Özge Gürcanlı. They welcomed me into their labs and gave me the first taste of research that encouraged me to embark on this journey.

For me, working towards a doctorate degree involved not only rigorous academic work but also a lot of introspection to understand myself and my mental health. In a global pandemic, in a world that shifts more rapidly than ever, I have been fortunate enough to have my therapist, Marianne Cook, who has helped me navigate through the world of academia and beyond.

I want to thank all my people who have made my life in Boston warmer and richer. All the laughs and tears that I have shared with Yon Soo Park, Boram Lee and Jaewon Yoon have helped me find the strength that I didn't know I had in me. My friendship with Hyunsuk Yun and Seohyun Lee has been a firm reminder that in Cambridge there is a little pocket of warmth no matter what. I was able to endure some of the hardest times of my graduate school thanks to all the delightful shots of espresso and even more delightful companionship that Gea Hyun Shin gave me.

Finally, with all my heart, I thank my family. My parents have always gone out of their ways to make sure that I have the best learning experience all my life, and my brother has shown me an example of continuously building a life path that suits oneself. I am deeply grateful for their unwavering love and encouragement from the other side of the globe.

List of Tables and Figures

Tables

Table 2.1. Mean \pm s.d. response time (ms) in each condition in Experiment 1.....	20
Table 2.2. Mean \pm s.d. response time (ms) in each condition in Experiment 2.....	28
Table 2.3. Mean \pm s.d. response time (ms) in each condition in Experiment 3.....	37
Table 3.1. Response time differences between talker variability conditions (by carrier duration)	59
Table 3.2. Interactions between talker variability and carrier duration on log response time.....	60
Table 4.1. Mean \pm s.d. response time (ms) in each condition.....	75

Figures

Figure 2.1. Stimuli for Experiments 1-3.....	16
Figure 2.2. Task design for all experiments.....	18
Figure 2.3. Results for Experiment 1.....	20
Figure 2.4. Hypothesized patterns of results for Experiment 2.....	24
Figure 2.5. Results for Experiment 2.....	28
Figure 2.6. Hypothesized patterns of results for Experiment 3.....	34
Figure 2.7. Results for Experiment 3.....	37
Figure 3.1. Schematic of the task design.....	57
Figure 3.2. Effects of talker variability and carrier duration across talkers on response times....	60
Figure 4.1. Stimulus variability across talkers and task design.....	71
Figure 4.2. tDCS paradigm.....	73
Figure 4.3. Processing cost of talker variability by speech context and stimulation condition....	76

Chapter 1. Introduction

1.1 Phonetic variability and speech perception

Although people without speech, hearing or language disorders experience everyday speech comprehension as effortless, speech perception actually involves a non-trivial computational process that rapidly and accurately extracts meaningful linguistic messages from highly variable acoustic speech signals. Listeners have a stable phonetic percept of a word regardless of where or from whom they hear it, but a single word may never be heard in the same way, as the acoustic realization of speech is highly variable due to a variety of factors including cross-talker differences in vocal tract anatomy and sociocultural influences on their speech production; even the same talker's production of the same word can vary because of coarticulatory effects, the environment in which their speech takes place, prosody, their age, and so on. Such immense variability in the physical properties of speech results in a significant overlap in phonological categories across talkers, as seen in acoustic measurements of English vowels (Peterson & Barney, 1952; Hillenbrand et al., 1995) and consonants (Koenig, 2000).

Among various factors, the major source of acoustic variability in speech is differences among talkers (Kleinschmidt, 2019). Anatomical differences in the length, volume and shape of the oral, pharyngeal, and nasal cavities across talkers result in variability in vocal tract resonance. These differences are, in consequence, perceived as a wide range of voice quality and formant frequencies. For example, males tend to have longer vocal tracts than females, which results in a gender difference in the distribution of frequency resonances. In addition, talkers' pattern of speech production is largely influenced by the accent and dialect that they learned in their childhood, which is often determined by regional and sociocultural factors. The phonetics of talkers' native language also has a big impact on how they produce speech in a certain

language (Flege, 1981). As listeners, we constantly encounter speakers who vary in all these dimensions and rapidly accommodate the variability in order to extract phonetic information from their speech.

Coarticulatory effects, where the articulation of a conceptually isolated speech sound prior to or following the speech sound affects the articulation of another speech sound, also have a substantial influence on the acoustics of speech production. The place and manner of articulation of a sound are influenced by not only the sounds that immediately precede or follow it but also the ones that are rather distant from it within the word or the utterance. As most of the speech that we encounter occurs in a stream of speech sounds rather than one speech sound in isolation, we are constantly experiencing the effect of coarticulation.

Moreover, talkers may use different ways of speech to compensate for the acoustics of the environment, and their suprasegmental elements such as intonation, rhythm, and stress may differ depending on the message they are trying to convey or their emotional state.

1.2 The processing costs of talker variability

Due to the lack of invariance in acoustic-phonemic mapping caused by aforementioned factors, additional processing costs are incurred in speech perception when listeners are subjected to talker variability. This cost of talker variability has been demonstrated repeatedly in previous studies showing that processing the speech of multiple talkers is slower and/or less accurate than attending to one talker's speech in a variety of behavioral tasks. For example, a classic study by Mullennix and Pisoni (1990) showed that, when listeners heard recordings of words spoken by multiple talkers, they were slower to identify the initial phoneme in those words, compared to when they heard the words spoken by only one talker. This cost of talker

variability persisted even when there was no potential confusability between the target phonemes that the listeners were identifying (Choi, Hu & Perrachione, 2018) and when the listeners were already familiar with the voices (Magnuson et al., 2021). In speech memory tasks, listeners were less accurate and slower when the phonetic properties of words are dissimilar between the stages of encoding and recognition (Palmeri, Goldinger, & Pisoni, 1993). Word recognition in noise was less accurate when they encountered words produced by unfamiliar talker (Nygaard, Sommers, & Pisoni, 1994); this appears to show the comparative advantage listeners can get from familiar talkers because they can better recover the masked part of the sound if they are already familiar with the talker-specific idiosyncrasies (Holmes & Johnsrude, 2020), which is also consistent with findings that attending selectively to one of two competing talkers is easier when listeners are familiar with the talker's voice (Newman & Evers, 2007). Regional accent differences result in a temporary disruption in speech processing as the listener adapts to the accent (Floccia, Goslin, Girard, & Konopczynski, 2006). Interestingly, manipulations that do not affect the phonetic features of speech do not have the effect of variability. For example, changing the talker or rate of speech reduces response accuracy, but changing the amplitude of speech sound does not (Bradlow, Nygaard, & Pisoni, 1999). The remarkable reproducibility of the costs of talker variability observed across countless studies using various behavioral assays appears to be because talker variability is the principal source of variability in phonetic properties of speech (Kleinschmidt, 2019).

1.3 Models of speech perception

One influential branch of understanding speech perception tends to treat acoustic variation in speech as a noise that needs to be reduced and thus “normalized” so that the acoustic

cues are scaled to invariable representations of meaningful speech sounds. *Intrinsic talker normalization* is where acoustic cues are scaled relative to other simultaneous acoustic cues within the target speech sound (e.g., Johnson, 1990; Syrdal & Gopal, 1986; Nearey, 1989). For example, Syrdal and Gopal (1986) showed that rescaling the first and second formant frequencies relative to the frequency of F3 and F0 reduced talker variation and increased the accuracy of vowel classification in the high/low and front/back dimensions. *Extrinsic talker normalization*, on the other hand, uses phonetic information from the surrounding context speech produced by the talker in order to scale the acoustic cues of the other speech sound. Evidence for this mechanism comes from studies showing that manipulating phonological features in the preceding context biased listeners' perception of the relevant features of following speech (Ladefoged & Broadbent, 1957; Francis, Ciocca, Wong, Leung, & Chu, 2006; Gerstman, 1968; Nusbaum & Morin, 1992; Wong & Diehl, 2003; Laing et al., 2012; Holt, 2006; Sjerps et al., 2011). Such a biasing effect of context is consistent with *contextual tuning theory*, which posits that preceding speech provides talker-specific context for interpreting the following speech (Nusbaum & Morin, 1992).

An alternative to talker normalization models is the *episodic model* of speech perception (Goldinger, 1998; Nygaard & Pisoni, 1998; Pierrehumbert, 2002; Johnson, 2005). According to this model, similarity to stored episodic memory of speech that statistically samples the space of talker characteristics provides a sufficient basis for phonetic constancy without explicit normalization. An incoming speech signal activates acoustically similar episodes in memory, and the incoming speech is recognized based upon the set of activated episodic traces. Depending on the specific branch of this perspective, the phonetic information and the indexical information are encoded as an integral whole episode rather than being analyzed into features and talker-

specific characteristics (e.g., Goldinger, 1998), or the episodes are analyzed so that phonetically relevant features are extracted from them (e.g., Pierrehumbert, 2002). The key assumption is of the episodic model is that achieving phonetic constancy is not based on normalization but rather based on episodic memories of speech that listeners have encountered. Supporting this notion, studies have shown that processing phonetic and indexical information cannot be clearly dissociated (Mullennix & Pisoni, 1990), and that training on indexical information facilitates phonetic learning (Nygaard & Pisoni, 1998). Kleinschmidt and Jaeger (2015) formalized this notion that phonetic processing is intricately linked to talker information with their *ideal adapter framework*, a theoretical account of speech perception positing that listeners have distinct beliefs (internal models) about how different talkers produce their speech sounds. Evidence for the link between processing phonetic information and talker information can also be found in studies showing that, just as talker variability interferes with speech comprehension, unfamiliar phonological forms – variability in phonology of language – interferes with talker identification (McLaughlin et al., 2019).

A more recent explanation of talker adaptation comes from a *feedforward attentional model*, positing that adapting to changing talkers is a manifestation of listeners successfully and quickly orienting their attention toward the talker, similar to the formation of an auditory stream (Bregman, 1990). The cognitive costs of processing talker variability are significantly reduced when stimuli are presented in blocks of the same talker (Stilp & Theodore, 2020). Choi and Perrachione (2019) showed that the duration and temporal proximity of preceding speech, not the richness of its phonetic details, affect word recognition efficiency. Talker continuity facilitates adaptation to different talkers automatically and immediately (Bressler et al., 2014; Kapadia & Perrachione, 2020; Lim et al., 2019a; Morton et al., 2015), while talker discontinuity

immediately incurs processing costs by disrupting listeners' attention and forcing them to reorient their attention to the new talker (Lim et al., 2021; Lim et al., 2019b; Mehrai et al., 2018; Wong et al., 2004). This set of evidence suggests that the efficiency afforded by talker continuity is a type of feedforward auditory attention (Shinn-Cunningham, 2008).

1.4 Neural bases of talker adaptation

Speech signals include both phonetic information and information about the speaker's voice. Traditionally, as right hemisphere stroke did not seem to cause deficit in comprehending the message from the speech signal (Blumstein & Myers, 2014), mainly the left hemisphere was implicated in speech processing. The left superior temporal region, in particular, was found to be critical in processing the details of speech sounds and categorizing them into meaningful linguistic units (e.g., Desai et al., 2008; Myers, 2007; Mesgarani et al., 2014; Yi et al., 2019). Processing vocal properties and identity, on the other hand, has been shown to be the role of primarily the right hemisphere (Stevens, 2004).

Although those studies have established at least partial independence of phonetic and vocal information, the interaction between the two dimensions is evident not only from the aforementioned behavioral studies but also from other neuroimaging studies. Electrophysiological evidence shows an interaction between talker information and linguistic information in speech processing (Zhang, Peng, & Wang, 2013; Kaganovich, Francis, & Melara, 2006). Left STG has been shown to increase in activity when listeners are processing mixed talkers relative to when processing single talker (Belin & Zatorre, 2003; Chandrasekaran, Chan, & Wong, 2011; Perrachione et al., 2016; Wong, Nusbaum, & Small, 2004; Zhang et al., 2016; von Kriegstein et al., 2010). The relative contributions of different brain regions at different

timepoints, and how different regions interact with each other to support processing talker variability, remains to be investigated.

1.5 Current dissertation

Chapter 2 presents a series of behavioral experiments that parametrically manipulated different attributes of the speech context that immediately preceded the target speech: the *duration* of the context, the *amount of detail* about each talker's speech articulation embedded within the context, and the *temporal proximity* of the context with the target speech. Each participant's speech classification speed was compared between mixed- and single-talker conditions.

The work from Chapter 2 motivated the investigation of the time course of talker adaptation in Chapter 3. Presented in Chapter 3 is an experiment where listeners performed rapid identification of target words spoken by single or mixed talkers, presented in isolation or preceded by a single-vowel carrier of varying durations spoken by the talker of target word. This work aimed to evaluate the hypothesis that a carrier vowel of sufficient duration can facilitate talker adaptation to the degree that word identification in the mixed-talker blocks is as efficient as in the single-talker blocks.

Chapter 4 presents a neurostimulation study where participants received non-invasive brain stimulation to their left superior temporal region while they performed an auditory word identification task with and without context speech preceding the target word, in single- and mixed-talker blocks. The response time difference between the single-talker and the mixed-talker conditions in each stimulation condition and each speech context condition is analyzed, and we

specifically investigate how the neurostimulation of left STG influences the interaction between talker variability and the availability of preceding speech context.

Chapter 2. Time and information in talker adaptation

Reproduced from

Choi, J. Y., & Perrachione, T. K. (2019). Time and information in perceptual adaptation to speech. *Cognition*, 192. <https://doi.org/10.1016/j.cognition.2019.05.019>

2.1 Introduction

A core challenge in speech perception is the lack of a one-to-one mapping between acoustic signals and intended linguistic categories (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Talkers differ in their vocal tract anatomy, dialect and speech mannerisms (Johnson, Ladefoged, & Lindau, 1993), resulting in different talkers using remarkably different acoustics to produce the same phoneme, or virtually identical acoustics to produce different phonemes (Hillenbrand, Getty, Clark, & Wheeler, 1995). Because of this variation, listening to speech from multiple or different talkers imposes additional processing costs, resulting in slower and less accurate speech perception than when listening to speech from a single consistent talker (Mullennix & Pisoni, 1990; Magnuson & Nusbaum, 2007). The empirical phenomenon of a talker-specific mode of listening, in which speech is processed faster and more accurately, is called talker adaptation, and has been observed across a number of experimental paradigms and for a variety of dependent measures (e.g., Kraljic & Samuel, 2007; Dahan, Drucker, & Scarborough, 2008; Trude & Brown-Schmidt, 2012; Xie, Earle, & Myers, 2018). A common account of how listeners maintain phonetic constancy across talkers is talker normalization (Johnson, 2005; Nusbaum & Magnuson, 1997; Pisoni, 1997), in which listeners use both signal-intrinsic (e.g., Nearey, 1989) and extrinsic (e.g., Johnson, 1990) information about a talker to

establish talker-specific mappings between acoustic signals and abstract phonological representations. Previous studies that have dealt with inter-talker variability mostly asked listeners to decide which of two sounds (e.g., /ba/ vs. /da/; Green, Tomiak, & Kuhl, 1997) or a very small set of isolated words (e.g., Mullennix & Pisoni, 1990; Cutler, Andics, & Fang, 2011) they heard in single- vs. mixed-talker contexts. However, real-world speech rarely occurs in such form. Most of the speech that we encounter comes from one talker at a time and in connected phrases, rather than from mixed talkers in isolated words. Even during conversations with multiple interlocutors, listeners still tend to get a sustained stream of speech from each talker at a time. Other studies that have investigated how the indexical context affects acoustic-to-phonetic mapping have demonstrated that listeners' perceptual decision of speech can be biased by preceding speech sounds. Manipulation of features in the prior context affects how the listeners perceived the relevant features of following speech signals (Francis, Ciocca, Wong, Leung, & Chu, 2006; Johnson, 1990; Ladefoged & Broadbent, 1957; Leather, 1983). Although these studies shed light on how the mapping between acoustic signals and speech categories is dynamically influenced by the context surrounding the speech sounds, they have focused mainly on how the context affects perceptual decision outcomes, not how it affects speech processing efficiency. The influence of context on perceptual decisions is clear from such studies, but they tell us little about how much or what kind of information listeners obtain from preceding contexts, nor do they elucidate the time course of using the context information. These limitations are also apparent in recent models of speech processing. For example, Kleinschmidt and Jaeger (2015) proposed a model of speech perception that achieves perceptual constancy through the comparison between encountered acoustic signals and listeners' expectations based on prior experience. Although this model captures the active, dynamic nature of acoustic-to-

phonemic mapping and explains why it is harder for listeners to process mixed-talker speech than single-talker speech, ultimately it accounts for only the decision outcomes that listeners make, without considering the psychological or biological operations that the perceptual system must undertake in order to reach those decisions, nor how those operations unfold in real time. Pierrehumbert (2016) posited a hybrid model of speech perception in which episodic traces of acoustic details are used in mapping the speech acoustics to an abstract phonemic representation (see also Goldinger, 1998). However, this model also does not describe the mechanistic processes for how information from prior speech encounters is integrated into perceptual decisions.

Overall, current models have thus achieved impressive success in describing the “computational” and “algorithmic” levels of perceptual adaptation to speech, but so far there has been no sustained attempt to account for the “implementational” level (Marr, 1982). Ultimately, our understanding of adaptation to talkers during speech processing still lacks an implementational description of (i) how the system operates in real time to arrive at a specific decision outcome among multiple possible interpretations of target speech acoustics, (ii) how much and what kinds of information the system uses to achieve such a decision, and (iii) the timescale in which the system integrates information about the talker-specific phonetic context of speech to facilitate its decision process. In this chapter, we report a preliminary empirical foundation that describes these three key constraints on the implementational level of talker adaptation, and we propose a potential theoretical framework through which talker adaptation can be explored as the integration between domain-general attentional allocation and linguistic representations. Neuroimaging studies have shown that adaptation to talker-specific speech is associated with reduced physiological cost (Perrachione et al., 2016; Wong, Nusbaum, & Small,

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

Listeners are faster and more accurate at processing speech from a single talker compared to mixed talkers presumably because they learn something about talker-specific idiosyncrasies from previous speech to adapt to each talker, making future speech processing more efficient. In this study, we aimed to further our understanding of how listeners take advantage of preceding speech context to facilitate perceptual decisions about speech. In particular, we wanted to

determine how speech processing efficiency is affected by (i) the amount of prior information that listeners have about a talker's speech and (ii) how much time they have to integrate that information prior to a perceptual decision. These questions are fundamental to establishing an implementational understanding of talker adaptation, as current models of processing talker variability in speech do not elaborate on how and when relevant information about the target talker's speech is ascertained during speech perception (Kleinschmidt & Jaeger, 2015; Pierrehumbert, 2016). To assess this question, we carried out a series of three experiments that explore the relationship between the amount of information listeners have about the phonetics of a talker's speech, the amount of time they have to process that information before making a perceptual decision, and the efficiency with which they can access speech content. In these experiments, listeners identified whether they heard the word “boot” or “boat”— a challenging speech distinction given the substantial overlap across talkers in the acoustic-phonetic-phonemic realization of the sounds /u/ and /o/ (Choi, Hu, & Perrachione, 2018; Hillenbrand et al., 1995). Because of the enormous potential confusability of these phonemes across talkers, we expected listeners to be much slower to make this decision in mixed-talker conditions, where the trial-by-trial correspondence between speech acoustics and phonemic targets is less stable, compared to single-talker conditions. In each of the three experiments, we manipulated the amount of information that listeners have about the current talker and the amount of time they have to integrate that information prior to identifying the word (“boot” / “boat”) by prepending the target words with carrier phrases of various lengths and contents. Specifically, we focused on how the response time to make the word identification changes as a consequence of listening to mixed talkers as opposed to single talker, which we refer to as the interference effect of talker variability. In Experiment 1, we established that speech processing efficiency is impacted by

preceding information about a talker and time to process it. By comparing the reduction in interference imparted by shorter vs. longer carrier phrases, we discovered that interference from mixed talkers is reduced as a function of the amount of preceding speech context. In Experiment 2, we examined how the quality of information in the carrier phrase serves to reduce interference. By comparing the reduction in interference made by a phonetically “complex” carrier phrase vs. a phonetically “simple” one, we discovered that the richness of phonetic information conveyed in the carrier phrase does not affect the magnitude of perceptual adaptation when the temporal duration of the carrier phrase is kept constant. In Experiment 3, we investigated how the speech perception system integrates phonetic information over time. By comparing the duration and temporal proximity of the carrier phrases to the target word, we discovered that a sustained stream of information is necessary over the duration of the context for the perceptual system to maximally facilitate adaptation to the talker. Overall, these experiments reveal (i) that the speech perception system appears to need surprisingly little information about a talker's phonetics in order to facilitate efficient speech processing, (ii) that the facilitation effect builds up with longer preceding exposure to a talker's speech, but (iii) that this gain depends on temporal continuity between adapting speech and word targets. Together, these experiments reveal how the psychological implementation of rapid perceptual adaptation to speech makes use of continuous integration of phonetic information over time to improve speech processing efficiency.

2.2 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- or

mixed-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.

2.2.1 Methods

2.2.1.1 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.

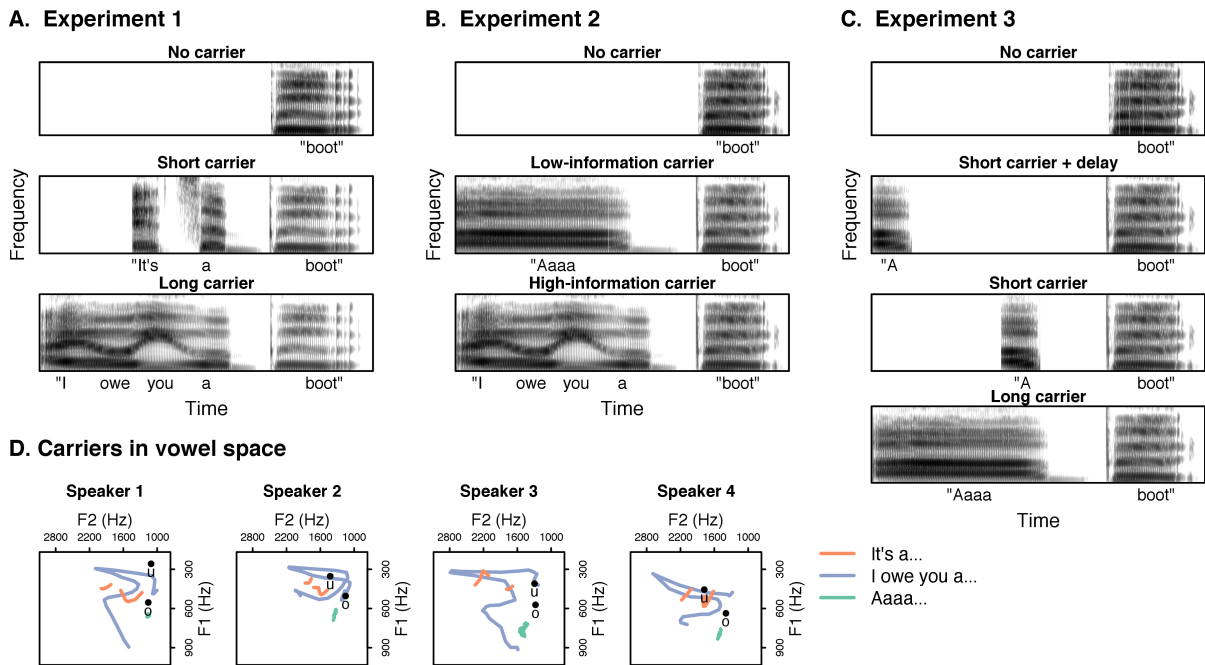


Figure 2.1. Stimuli for Experiments 1-3. (A,B,C) Spectrograms of example stimuli produced by Speaker 2 used in Experiments 1-3 in each condition. (D) Lines indicate the F1-F2 trajectory of all carriers produced by each talker. Black points indicate the F1-F2 position of the /o/ and the /u/ vowels in 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>

2.2.1.2 Stimuli

Stimuli included two target words “boat” and “boot.” These target words were chosen because the phonetic-phonemic correspondence of the /o/-/u/ contrast is highly variable across talkers (Hillenbrand et al., 1995) and therefore highly susceptible to interference in a mixed-

talker setting (Choi, 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. 2.1A,D**).

Words and carrier phrases were recorded by two male and two female native speakers of American English in a sound-attenuated room with a Shure MX153 earset microphone and Roland Quad Capture sound card sampling at 44.1kHz 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. To ensure the naturalness of concatenated sentences, we manipulated pitch, amplitude, and the voice-onset time of the carrier phrase and target words. All the recordings were normalized for RMS amplitude to 65 dB SPL in Praat (Boersma, 2001). Short carrier phrases were 298-382 ms; long-carrier phrases were 544-681 ms. Examples of these stimuli are shown in **Fig. 2.1A**.

2.2.1.3 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. 2.2**). In each block of 48 trials, each target word occurred in 24 trials. Stimulus presentation was pseudo-randomized such that the same target word was not presented for more than three consecutive trials. The order of conditions was counter-balanced across participants using Latin square permutations. Each participant listened to words spoken by the same talker across all three single-talker blocks, and all four talkers served as the single talker across participants.

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

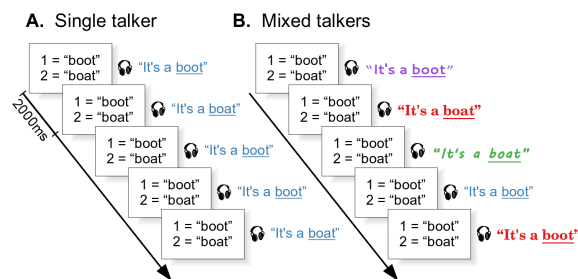


Figure 2.2. Task design for all experiments. Participants performed a speeded word identification task while listening to speech produced by either (A) a single talker or (B) mixed talkers. The short-carrier condition for Experiment 1 is shown.

2.2.1.4 Data analysis

Accuracy and response time data were analyzed for each participant in each condition. Accuracy was calculated as the proportion of trials where participants identified the word correctly out of the total number of trials. Response times were measured from the onset of the

target word. Response times were log-transformed to ensure normality. Only the response times from correct trials were included in the analysis. Outlier trials that deviated by more than three standard deviations from the mean log response time in each condition were also excluded from the analysis (< 1% of total correct trials). Data were analyzed in R using linear mixed-effects models implemented in the package *lme4*, with response times 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)¹.

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

2.2.2 Results

Participants' word identification accuracy was at ceiling (mean = 98% ± 2%). Consequently, the dependent measure for this experiment was response time (**Table 2.1**), as is usual for studies of perceptual adaptation 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 2.3**.

¹ Across experiments, these models took the form, in R notation:
 $\log_{10}(\text{response time}) \sim \text{indexical variability} * \text{context} + (1 + \text{indexical variability} + \text{context} | \text{subject})$

Table 2.1. Mean \pm s.d. response time (ms) in each condition in Experiment 1.

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

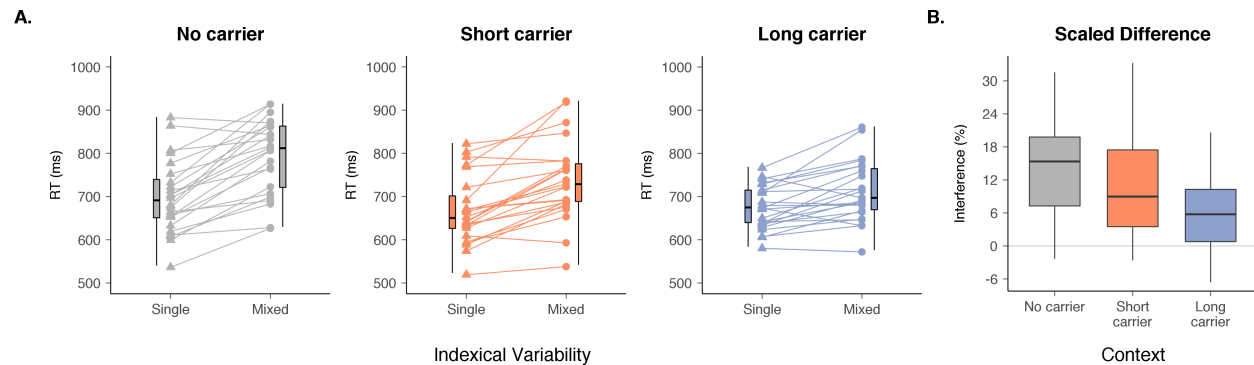


Figure 2.3. Results for Experiment 1. Effects of talker variability and context across talkers on response times. (A) Connected points show the change in response times for individual participants between the single- and mixed-talker conditions across three levels of context. Box plots in each panel 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: $((\text{mixed} - \text{single}) / \text{single}) \times 100$. 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.

Compared to the single-talker conditions, response times in the mixed-talker conditions were significantly slower overall (single 679 ms vs. mixed 747 ms; $F(1, 23.1) = 109.01$; $p \ll 0.001$). Post-hoc pairwise testing revealed that response times in the mixed-talker condition were significantly slower than in the single-talker condition for each of the three carrier conditions independently (**Table 2.1**): *no carrier* single-talker 698 ms vs. mixed-talker 792 ms ($\beta = 0.12$, $s.e. = 0.010$, $t = 11.96$, $p < 0.001$); *short-carrier* single-talker 666 ms vs. mixed-talker 736 ms ($\beta = 0.096$, $s.e. = 0.010$, $t = 9.21$, $p < 0.001$); *long-carrier* single-talker 672 ms vs. mixed-talker 711

ms ($\beta = 0.050$, $s.e. = 0.010$, $t = 4.84$, $p < 0.001$). 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.

In a model including all three carriers simultaneously, significant carrier \times variability interactions were observed ($F(2, 6658.6) = 27.52$; $p \ll 0.001$), 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.

2.2.3 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 an environment where the talker changes. 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 information about a talker's vocal tract and articulation: the short carrier phrase encompassed two vowels (/ɪ/, /ʌ/) that varied primarily in F2, while the long carrier phrase contained at least five distinct vowel targets (/a/, /i/, /o/, /u/, /ʌ/) including the target vowels (/o/, /u/) and effectively sampled the entire vowel space (**Fig. 2.1A,D**). That is, the long carrier not only contained richer and more relevant talker-specific detail about his/her speech production, but it also provided listeners with more time to adapt to the talker. In addition to how time and information is intertwined in the context manipulation in this experiment, the design of the experiment introduces another variable that might affect response times: because

subjects were given 2000ms between the onsets of trials regardless of the duration of carrier phrase, they were given less time to respond in long-carrier condition than in short carrier condition, which may have driven participants to respond more quickly for the long carrier condition.

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.

2.3 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. 2.1B,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 interference effect of mixed-talker speech should be lower in the high-information carrier phrase

than the low-information carrier (**Fig. 2.4A**). 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. 2.4B**).

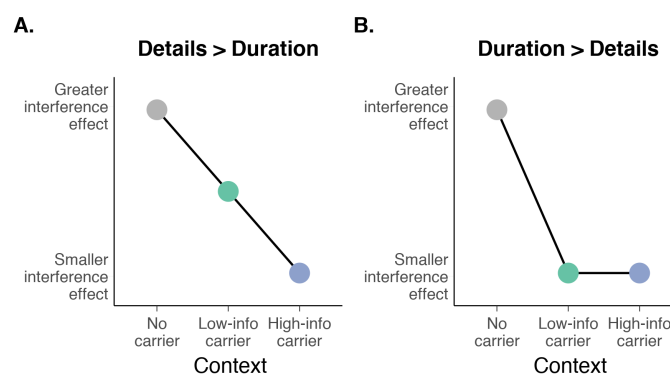


Figure 2.4. Hypothesized patterns of results for Experiment 2. Potential patterns for the interference effect of talker variability across the three experimental conditions, as predicted by the two different hypotheses about contextual effects on talker adaptation. **(A)** If the amount of talker-specific phonetic details in a carrier contributes more to talker adaptation than the duration of the carrier, the interference effect will be lower in the high-information carrier condition than in the low-information carrier condition. **(B)** If the duration of a carrier contributes more to talker adaptation than the richness of its phonetic details, the interference effect will not differ between the low- and the high-information carriers, as their durations are matched.

2.3.1 Methods

2.3.1.1 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.

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., “Aaaa [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. 2.1B**.

2.3.1.3 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. Each participant listened to words spoken by the same talker across all three single-talker blocks, and all four talkers served as the single talker across participants.

2.3.1.4 Data Analysis

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

2.3.2 Results

Participants' word identification accuracy was again at ceiling ($98\% \pm 2\%$), and so the dependent measure for this experiment was also response time (**Table 2**). Participants' response times in each condition are shown in **Figure 2.5**.

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; $F(2,6626.6) = 24.96, p \ll 0.001$). Pairwise analyses revealed that, for all three speech context conditions, response times were significantly slower in the mixed-talker condition than in the single-talker condition (**Table 2**): *no carrier* single-talker 705 ms vs. mixed-talker 784 ms ($\beta = 0.11, s.e. = 0.010, t = 10.67, p \ll 0.001$); *low-information carrier* single-talker 679 ms vs. mixed-talker 716 ms ($\beta = 0.053, s.e. = 0.010, t = 5.26, p \ll 0.001$); *high-information carrier* single-talker 662 ms vs. mixed-talker 697 ms ($\beta = 0.046, s.e. = 0.010, t = 4.63, p \ll 0.001$). As 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.

Table 2.2: Mean \pm s.d. response time (ms) in each condition in Experiment 2

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

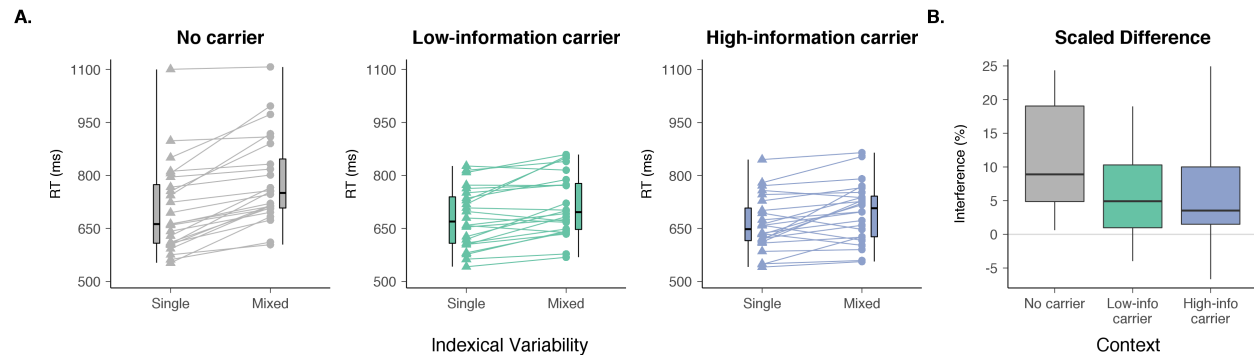


Figure 2.5. Results for Experiment 2. Effects of talker variability and context on response times. (A) Connected points show the response times in the single- and mixed-talker conditions across three levels of context for individual participants. Box plots in each panel show the distribution (median, interquartile range, extrema) for each variability-by-context condition. (B) The interference effect of indexical variability is shown for each level of context. The distribution of differences in response time between the mixed- and single-talker conditions is shown, scaled within participant to their response time in the single-talker condition: $((\text{mixed} - \text{single}) / \text{single}) \times 100$. 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. 2.4B).

We observed an interesting pattern of significant context \times variability interactions, indicating different effects on the magnitude of perceptual adaptation between the single- and mixed-talker conditions across speech contexts: 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. 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.

2.3.3 Discussion

The results from Experiment 2 refine our understanding of the temporal dimension of auditory adaptation to talkers and the source of information that facilitates this adaptation. As in Experiment 1, the interference effect of talker variability was greatest in the no-carrier condition where listeners were not given any preceding speech context, and the effect was reduced in both the low- and high-information carrier conditions where the brief preceding speech context allowed listeners to adapt to the talker on each trial. Surprisingly, Experiment 2 revealed that the increase in processing efficiency afforded by a carrier phrase in multi-talker speech contexts did not differ as a function of the amount 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 this representation and use it to understand speech (Kleinschmidt & Jaeger, 2015; Pierrehumbert, 2016). However, these models do not describe the implementational level of these computations; that is, they do not elaborate what kind of or how much talker-specific information is needed to affect perceptual outcomes, nor do they account for how or when the information must be integrated by listeners in order for them to utilize it for the perception of upcoming speech. The results from our experiment show that a carrier phrase that thoroughly samples the talker's vowel space is no more facilitatory than a much more impoverished form of carrier speech, suggesting that the amount of talker-specific information necessary to make speech processing more efficient is, in fact, minimal. It is possible that because inter-talker variability in the acoustic realization of speech is not completely random but rather structured regarding talkers' socio-indexical characteristics (Kleinschmidt, 2018), talker-specific cues with minimal phonetic information might have sufficiently facilitated talker adaptation in Experiment 2.

Coupled with the results of Experiment 1 where a longer carrier phrase afforded greater facilitation of speech processing efficiency than a shorter carrier, the results of Experiment 2 also suggest that the speech perception system requires a sufficient amount of time to integrate talker-specific information to facilitate the processing of future speech content. This raises the question of how the timecourse of such integration unfolds. Some authors have claimed that episodic models of speech processing – in which reactivation of listeners' memories of prior speech experiences guides future speech processing – can account for talker normalization / adaptation phenomena (Goldinger, 1998). Contemporary computational models have explicitly incorporated these mnemonic mechanisms into their perceptual decision processes (Kleinschmidt & Jaeger,

2015; Pierrehumbert, 2016): When a listener hears speech from a particular talker, the speech processing system will implicitly recognize that talker, re-activate related memories of their speech, and integrate them into perceptual processing in order to guide talker-specific interpretation of upcoming speech sounds. However, memory reactivation is a time-dependent process. Consequently, one implication of an episodic account of talker adaptation is that integration of talker specific information will be *ballistic*; that is, once a new talker is encountered, memories of that talker's speech automatically tune the speech perception system to facilitate processing that talker's speech, but a certain amount of time is required for the auditory system to reactivate the relevant memories underlying its perceptual recalibration. Alternatively, rather than the time-dependent reactivation of memories of similar speech as predicted by episodic/mnemonic models of speech processing, the integration of talker-specific information may depend on continuous integration of a talker's speech over time, akin to auditory streaming and auditory object formation (Shinn-Cunningham, 2008; Winkler et al., 2009). In this account, continuous exposure to a talker's speech facilitates attentional orientation to the relevant auditory features associated with that talker, such that there is a facilitatory effect of not only the length of an adapting speech context, but also its temporal proximity to a speech target. To adjudicate between a mnemonic/ballistic model of talker adaptation and an object continuity/streaming model, we therefore undertook a third experiment in which we varied both the *duration* of the adapting speech context and its *continuity* with respect to the target word.

2.4 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. 2.1C**).

The mnemonic/ballistic and the object-continuity/streaming models of talker adaptation predict different patterns of facilitation effected by these carrier-phrase conditions in the mixed-talker context. If talker adaptation is ballistic, then once speech is encountered and talker-specific memories are reactivated we should expect equal amounts of facilitation by the long-carrier and short-carrier-with-delay conditions. Because the onset of speech in these conditions occurs equidistant from the target lexical item, listeners will have had equal time to re-activate talker-specific memories. Correspondingly, both of those conditions should offer greater facilitation than the short-carrier-without delay, in which speech onset occurs closer to the target word and thus affords less time for activation and integration of talker-specific memories (**Fig. 2.6A**). Alternatively, if talker adaptation depends on attentional reorientation via auditory streaming across time, then the pattern of results should be markedly different (**Fig. 2.6B**): 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.

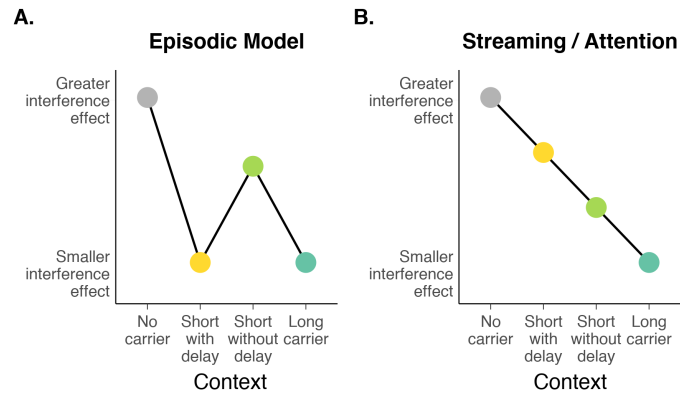


Figure 2.6. Hypothesized patterns of results for Experiment 3. Potential patterns for the interference effect of talker variability across the four experimental conditions, as predicted by the two different hypotheses of the contribution of temporal continuity of context. **(A)** The predicted pattern from an episodic account of speech perception. Due to having the greatest time available to reactivate talker-specific memories, the long-carrier and short-carrier-with-delay conditions should have the smallest (and equal) interference effects. The short-carrier-without-delay has less time to access memories, and so should have a larger interference effect than either of the other carriers. **(B)** The predicted pattern from an attention/streaming model of speech perception. In contrast to the episodic account, this model predicts a greater interference effect in the short-carrier-with-delay condition than either the short-carrier-without-delay condition or the long-carrier condition. In these latter two conditions, the temporal proximity between the adapting speech and the target word should facilitate the emergence of a talker-specific auditory object and improve processing efficiency.

2.4.1 Methods

2.4.1.1 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.

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. 2.1C). 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). We ensured that the total duration of each speaker's short-duration carriers with an intervening pause matched the duration of that speaker's long-duration carrier. Each speaker's three carrier phrases were then concatenated with the target words spoken by the same speaker to produce natural-sounding recordings.

2.4.1.3 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. Each participant listened to words spoken by the same talker across all three single-talker blocks, and all four talkers served as the single talker across participants.

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

2.4.1.4 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).

2.4.2 Results

Participants' word identification accuracy was again at ceiling ($99\% \pm 2\%$), 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 2.7**.

Table 2.3: Mean \pm 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 \pm 72	649 \pm 60	651 \pm 72	640 \pm 71
Mixed talkers	754 \pm 85	706 \pm 67	698 \pm 77	671 \pm 67
Differences	84 \pm 56	57 \pm 53	47 \pm 44	31 \pm 54

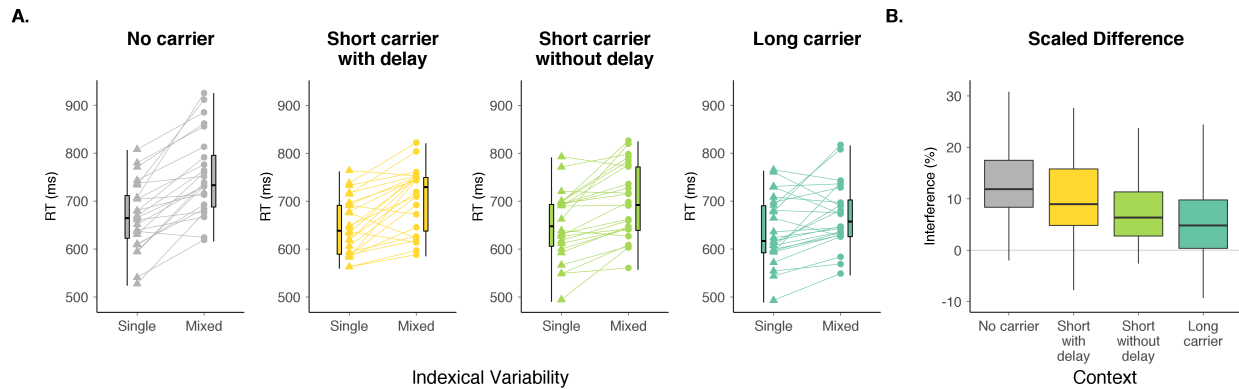


Figure 2.7. Results for Experiment 3. Effects of talker variability and context across talkers on response times. **(A)** Connected points show the change in response times for individual participants between the single- and mixed-talker conditions across four levels of context. Box plots in each panel 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: $((\text{mixed} - \text{single}) / \text{single}) \times 100$. 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. 2.6B**).

Compared to the single-talker conditions, response times in the mixed-talker conditions were significantly slower overall (single 652 ms vs. mixed 707 ms; $F(1, 23) = 71.89$; $p \ll 0.001$). For all four carrier conditions independently, we observed significantly slower response times in the mixed-talker condition than in the single-talker condition (**Table 3**): *no carrier* single-talker 670 ms vs. mixed-talker 754 ms ($\beta = 0.11$, $s.e. = 0.011$, $t = 10.35$, $p \ll 0.001$); *short-carrier with delay* single-talker 649 ms vs. mixed-talker 706 ms ($\beta = 0.084$, $s.e. = 0.011$, $t = 7.74$, $p < 0.001$);

short-carrier without delay single-talker 651 ms vs. mixed-talker 698 ms ($\beta = 0.065$, $s.e. = 0.011$, $t = 6.07$, $p < 0.001$); *long-carrier* single-talker 640 ms vs. mixed-talker 671 ms ($\beta = 0.047$, $s.e. = 0.011$, $t = 4.42$, $p < 0.001$). 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.

To understand the carrier \times variability interaction across four levels of carrier, we turned to the pairwise successive difference contrasts on the interaction terms of the linear model. The variability by carrier interaction was significant for no-carrier vs. short-carrier with delay (+57 ms / 9%; $\beta = 0.014$, $s.e. = 4.7 \times 10^{-3}$, $t = 2.94$, $p < 0.01$). 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 ($\beta = 9.0 \times 10^{-3}$, $s.e. = 4.7 \times 10^{-3}$, $t = 1.90$, $p = 0.057$). 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 without delay* condition ($\beta = 8.9 \times 10^{-3}$, $s.e. = 4.7 \times 10^{-3}$, $t = 1.88$, $p = 0.059$).

To understand the pattern of results in terms of our hypotheses (Fig. 6), we turned to the polynomial contrasts on the interaction terms of the linear model. The hypothesized pattern of results expected by episodic model (**Fig. 2.6A**) would be approximated best by a cubic model whereas the hypothesized pattern of results expected by linear model (**Fig. 2.6B**) would be best approximated by a linear model. Polynomial contrasts on the interaction terms of the model demonstrate a significant effect of linear pattern ($\beta = 0.023$, $s.e. = 3.3 \times 10^{-3}$, $t = 7.0$, $p \ll 0.001$) but not cubic pattern ($\beta = 1.08 \times 10^{-3}$, $s.e. = 3.3 \times 10^{-3}$, $t = 0.32$, $p = 0.75$).

2.4.3 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. 2.6B**), rather than the time required to re-activate memories of a talker once encountered (**Fig. 2.6A**). An episodic model of talker adaptation would predict equally large reduction in interference by the short carrier with delay and by the long carrier. It would also predict a greater reduction in interference by a short carrier with delay than one immediately adjacent to the target speech. However, 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.

2.5 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 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 a prior speech context consisting of only a single sustained vowel had just as much facilitatory effect as another context that fully sampled each talker's entire vowel space, provided the preceding speech samples had the same duration. Thus, the gradient effect of carrier length on perceptual adaptation observed in Experiment 1 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 theoretical frameworks that invoke either memory or attention as the mechanism underlying 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.

2.5.1 Extension and refinement of prior models of talker adaptation

2.5.1.1 Contextual tuning models

Previous studies exploring the impact of extrinsic cues on the perception of following target speech have primarily emphasized the role of context as a frame of reference against which the target speech can be compared to affect the outcomes of perceptual decisions. For example, variation in the F1 of an introductory sentence can bias perceptual decisions for

following, acoustically identical, speech sounds (Ladefoged & Broadbent, 1957). This biasing effect of context is consistent with *contextual tuning theory*, which proposes that preceding speech provides talker-specific context (i.e., the talker’s vocal characteristics) for interpreting the following speech target (Nusbaum & Morin, 1992). Contemporary models have formalized such propositions for determining perceptual outcomes for speech, as in the *ideal adapter framework* (Kleinschmidt & Jaeger, 2015). However, context does more than just provide a reference for weighting perceptual decisions about speech categories; preceding speech also allows listeners to process target speech contrasts more efficiently, 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.

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).

2.5.1.2 Active control process models

The facilitatory effect of context on perceptual adaptation has been explained with models that treat speech perception as an active process of building possible hypotheses and testing them against the incoming signal. Such models often propose an active control mechanism (e.g., Magnuson & Nusbaum, 2007; Wong & Diehl, 2003), by which some cognitive process monitors incoming speech and initiates the computations underlying perceptual adaptation (e.g., Nearey, 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, listeners can engage this active control mechanism when they encounter each new talker's carrier phrase in mixed talker conditions. Consequently, the perceptual system will not need to expend as much cognitive effort to map the incoming acoustics of the subsequent target word to its intended linguistic representation. 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, while the long-carrier provides no additional phonetic information compared to the short-carrier-with-delay, it still affords greater adaptation to the target talker. This demonstrates that an active control process 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).

2.5.1.3 Episodic memory models

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 will be slower and less accurate. However, when the number of models is smaller – such as when a listener can limit model selection to a single talker – speech recognition will be faster. The computation underlying this internal model selection is described as an inference that draws not only on bottom-up evidence from the speech signal but also top-down expectation from signal-extrinsic cues such as visual or phonetic cues (Kleinschmidt & Jaeger, 2015, pp. 180-182). 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 of short-term perceptual adaptation to a talker’s speech: Namely, by showing how talker adaptation unfolds over time, our results suggest that the efficient allocation of auditory attention involved in streaming / object formation is likely the active cognitive process underlying talker adaptation.

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

Explaining the findings from Experiment 3, in which the duration of speech context and its temporal continuity with the target speech afforded maximal talker adaptation, requires us to 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, 2019), 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 disrupts the integration of the carrier and the target word into a coherent auditory object, resulting in less talker adaptation and a greater interference effect in mixed-talker environments than the other carrier phrases which were temporally continuous with the target speech.

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

2.5.3 Limitations and directions for future work

Across three experiments, we parametrically varied the length, content, and contiguity of speech context preceding target words to investigate how context facilitates speech processing. The pattern of results across these three experiments both sheds light on the temporal and informational factors underlying talker adaptation and emphasizes the potential contributions of domain-general attention and auditory streaming in talker adaptation. However, considerable future work remains to both replicate and extend the predictions made by this framework. In particular, our observations are based on a limited set of carefully-controlled laboratory stimuli – two words spoken by four talkers in the absence of any auditory distractor. While we chose these particular stimuli to optimize the processing interference from multiple talkers (Choi et al., 2018), it will be important to show that these results generalize to contrasts that are less confusable across talkers. Furthermore, the repetitious identification of either of two target words is a *de minimis* case of speech perception, whereas real world utterances are highly heterogeneous and depend on a larger variety of contextual cues. Likewise, auditory streaming has traditionally been explored in contexts where multiple sound sources compete for attention and perceptual organization, whereas the present experiments involved multiple sequential, rather than simultaneous, sound sources. While the suggestion that talker adaptation involves feedforward auditory streaming also offers an opportunity to bridge previously disparate work on talker variability and source selection in speech processing, a model approaching speech adaptation as a process of building auditory objects will need to be further studied in more complex auditory scenes. Future work involving open-set stimuli, more ecological utterances, and real- world listening environments – such as conversations – will be needed to better understand how talker adaptation entails auditory attention. Finally, a major requirement of

future work will be to reconcile the predictions and results of processes-based measures of talker adaptation (e.g., differences in response time to single vs. mixed talkers with vs. without carrier phrases) with outcome-based measurements (e.g., differences in perceptual biases for ambiguous vowels or consonants based on contextual information, as in Johnson, 1990).

2.6 Conclusions

The results from this study show that speech processing is made more efficient via the perceptual adaptation to a talker arising from preceding speech context. The pattern of results suggests that talker adaptation is facilitated by exposure to preceding speech from a talker that is brief (but not too brief), that is temporally continuous with the target speech, and that needs contain only minimal phonetic content. Together, these patterns of temporal and informational effects on talker adaptation raise the possibility that the efficiency gains in speech perception associated with talker adaptation may reflect the successful allocation of auditory attention to a target auditory object (i.e., a talker).

Chapter 3. Two distinct mechanisms of processing talker variability

Reproduced from

Choi, J.Y., Kou, R.S.N., & Perrachione, T.K. (2022). Distinct mechanisms for talker adaptation operate in parallel on different timescales. *Psychonomic Bulletin & Review*, 29, 627-634.

3.1 Introduction

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

attention (Bressler et al., 2014; Choi & Perrachione, 2019a; Kapadia & Perrachione, 2020; Lim et al., 2021).

An important contribution to understanding how listeners resolve talker variability comes from the role played by preceding speech context during word identification. For instance, early studies showed that talker-specific variation in the fundamental frequency of a preceding sentence could bias the interpretation of an ambiguous vowel sound (Johnson, 1990) – reflecting a phenomenon known as ‘extrinsic normalization’ of speech acoustics. Recent work has expanded on this to show how ongoing speech context improves speech processing efficiency by allowing listeners to form a coherent auditory stream (Bregman, 1980) with the current talker as its source. In mixed-talker contexts, the duration and temporal proximity of preceding speech – but not the richness of its phonetic content – affect word recognition efficiency (Choi & Perrachione, 2019a). Similarly, the cognitive demands of processing words spoken by multiple talkers are reduced when stimuli are blocked by talker (Perrachione et al., 2011; Stilp & Theodore, 2020). The improvements to accuracy and response time imparted by talker continuity appear to be automatic, immediate, and independent of listeners’ perceptual expectations (Bressler et al., 2014; Carter et al., 2019; Kapadia & Perrachione, 2020; Lim et al., 2019a; Morton et al., 2015), suggesting that talker continuity improves speech processing efficiency by feedforward capture of selective auditory attention (Shinn-Cunningham, 2008). Correspondingly, talker discontinuity (the abrupt change from one talker to another) appears to incur processing costs by disrupting listeners’ attention to one auditory object and requiring them to refocus their attention on a new source (Mehrai et al., 2018; Lim et al., 2019b; 2021; Wong et al., 2004). Thus, under an auditory streaming framework, the accuracy and response time differential between mixed- and single-talker speech contexts can be understood as speech processing

efficiency gains via successful allocation of feedforward auditory selective attention vs. efficiency losses from ongoing attentional disruption and reorientation.

An untested prediction of the auditory-streaming framework of talker adaptation is that there should be some duration of preceding speech from a continuous talker that is sufficient for fully capturing a listener's auditory attention, thereby rendering their speech processing maximally efficient. That is, in a context where a listener is hearing multiple different talkers in turn, there should be some duration of continuous speech from one talker that would allow speech processing to be as efficient as if the listener were in a single-talker context. Prior work has shown that target words are recognized more efficiently when they are preceded by a brief carrier phrase from the same talker, but the durations of carrier phrases tested (300 and 600 ms) did not fully ameliorate the additional processing costs from the mixed-talker context (Choi & Perrachione, 2019a). Extrapolating linearly from the trend in this prior report, we hypothesized that a continuous speech context of approximately 1100 ms should allow a listener to become fully adapted to a talker, even in a mixed-talker context. In contrast, the active control model of processing variability in speech (Magnuson & Nusbaum, 2007) postulates a different mechanism behind talker-related inefficiencies in speech perception, and thus makes a different prediction about how much benefit listeners can extract from talker continuity in a mixed-talker situation. In this model, when faced with potential uncertainty about the acoustic composition of speech sounds – such as in listening contexts involving multiple talkers – listeners pre-allocate cognitive resources in anticipation of the need to resolve that acoustic-phonetic uncertainty (Nusbaum & Magnuson, 1997; Nusbaum & Schwab, 1986). Under this account, top-down expectation of talker variability – triggered either by detection of a novel talker or by listeners' situational knowledge – engages additional cognitive resources for determining the phonetic content of

speech (Magnuson & Nusbaum, 2007; Heald & Nusbaum, 2014). This allows listeners' perceptual system to be flexible in accommodating variability in the mapping between incoming acoustic signals and internal phonetic representations, but at the cost of increased processing time compared to listening contexts where variability is not expected. Thus, while short-term talker continuity may refocus the listener's auditory attention and make speech processing more efficient (Choi & Perrachione, 2019a), this feedforward process may not ultimately be sufficient to completely ameliorate the cognitive costs of talker variability when listeners are expecting to hear multiple talkers and are thus subjecting the speech signal to additional top-down analysis in anticipation of phonetic ambiguity from the ongoing variability (Heald & Nusbaum, 2014). The active control model thus predicts that, faced with uncertainty from potential talker changes, speech processing will always be less efficient in a mixed-talker context compared to a single-talker context.

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

3.2 Methods

3.2.1 Participants

Native speakers of American English (N = 24; 20 female, 4 male; mean age 19.8 years, range 18-22 years) successfully completed this study. Additional participants were recruited for this study but were excluded from analysis because they had accuracy below 90% in any of the conditions (n = 5). All participants reported a history free from speech, language, or hearing disorders. No participant had previously participated in a similar experiment in our laboratory or had prior experience with the talkers. Participants provided written informed consent, approved and overseen by the Institutional Review Board at Boston University.

3.2.2 Stimuli

The naturally spoken English words “boot” and “boat” were recorded by 8 native speakers of American English (4 female, 4 male). These words were chosen because their minimally contrastive vowels (/u/ vs. /o/) have the greatest potential acoustic-phonemic ambiguity across talkers (Hillenbrand et al., 1995; Choi et al., 2018). In addition to the target words, speakers were also recorded producing a brief, sustained “uh” before the words “boot” and “boat” ([ʌ:but], [ʌ:bot]). These recordings were spliced at the end of the silent portion between the closure and the release burst of /b/ so that the sustained /ʌ:/ could be used as a carrier to elicit talker adaptation / attentional reorientation (Johnson, 1990; Choi & Perrachione, 2019a). Among numerous recordings of the carrier, the token with the most stable formant frequencies, amplitude, and fundamental frequency was selected for each talker. Then, using the pitch synchronous overlap-and-add algorithm (PSOLA; Moulines & Charpentier, 1990) implemented in the software Praat (Boersma, 2001), the duration of the voiced part of the carrier was adjusted so that the total duration of the carrier equaled 300, 600, 900, 1200, and 1500 ms. These carriers

were then prepended to the target words. This carrier was chosen because an isolated vowel carrier has been shown to induce as much adaptation as phonetically rich carrier phrases of equivalent duration (Choi & Perrachione, 2019a; Morton, Sommers & Lulich, 2015). Recordings were made in a sound-attenuated booth using a Shure MX153 microphone and Roland Quad Capture sound card, sampled at 44.1 kHz and 16-bit resolution. Stimuli were RMS amplitude normalized to 65 dB SPL in Praat.

3.2.3 Procedure

Participants performed a speeded word recognition task in which they identified the target word as quickly and accurately as possible by pressing a corresponding number on a keypad. Participants received verbal instructions at the beginning of the experiment, and written directions assigning a number to each target word were displayed on the screen throughout the experiment. Stimuli were presented with a 1500-ms interval between the onset of the target word and the onset of the following stimulus (**Figure 3.1**), and stimulus delivery was controlled using PsychoPy2 (v1.83.03) (Peirce, 2007) via Sennheiser HD-380 Pro headphones.

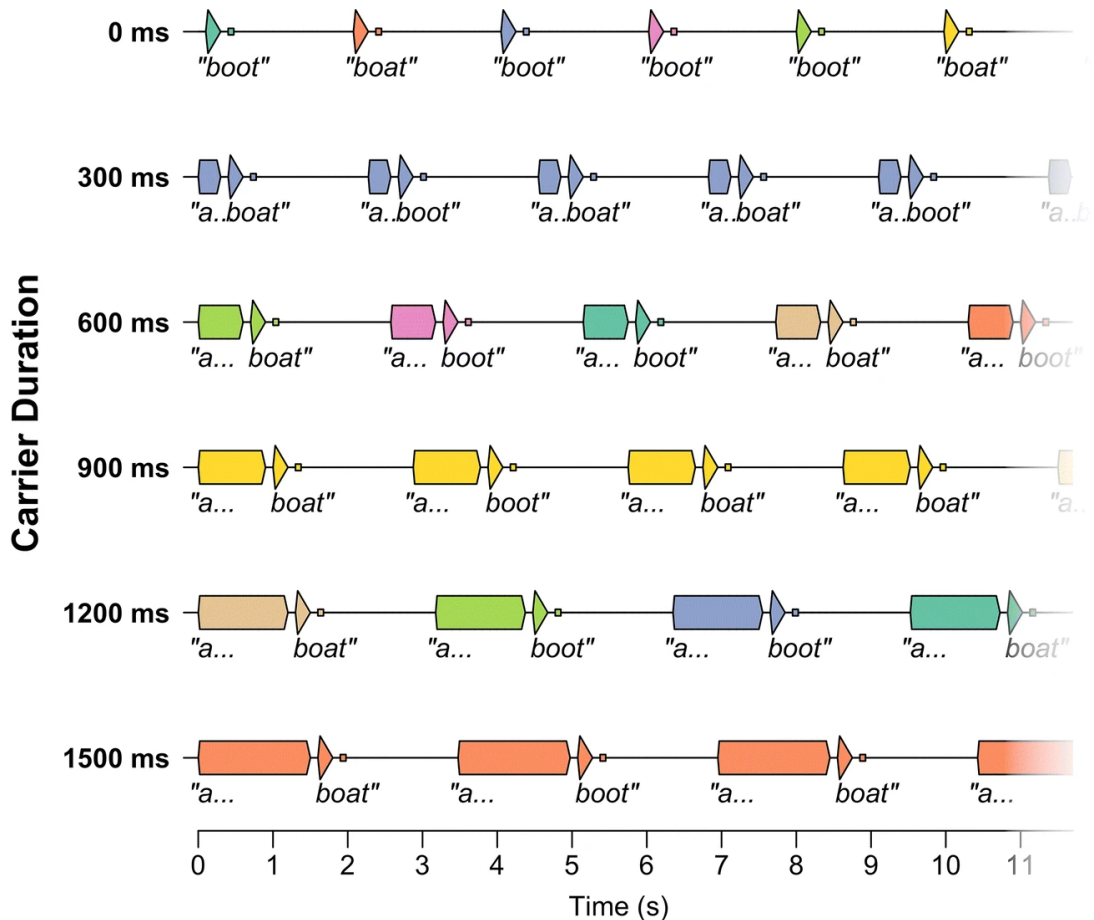


Figure 3.1. Schematic of the task design. A stylized version of the task acoustics is shown, depicting the speech waveforms from different talkers in different colors. Participants identified the target word (“boot” or “boat”) on each trial. Target words were presented in isolation (0ms carrier) or preceded by a sustained vowel /ʌ:/ from the same talker. Trials were blocked by carrier length (varying parametrically from 0 to 1500 ms) and by the single- or mixed-talker condition. Mixed-talker conditions are shown for 0-, 600-, and 1200-ms carriers; and single-talker conditions are shown for 300-, 900-, and 1500-ms carriers.

The task was divided into separate blocks, parametrically varying in talker variability (single vs. mixed) and carrier duration (0, 300, 600, 900, 1200, 1500 ms). Each block was 48 trials long, with each target word presented 24 times per block. Stimuli were presented in a pseudorandom order such that the same target word did not repeat in more than three consecutive trials and the same talker did not repeat in adjacent trials in mixed-talker conditions. Participants heard all talkers during the mixed talker conditions. The talker heard during each single talker condition

was counterbalanced across participants and carrier lengths. The order of conditions was counterbalanced across participants.

3.3 Results

Accuracy and response time (RT) data were collected on each trial. Accuracy in each condition was calculated as the proportion of trials where the participant responded correctly out of all trials. Because our a priori inclusion criteria required participants to have accuracy above 90% in every condition, our planned analyses focused on differences in RT alone, as in Choi & Perrachione (2019a). Analogous tasks have shown limited dynamic range for analyses of accuracy (Kapadia & Perrachione, 2020), and participants' word identification in the present study was at ceiling (mean accuracy $99\% \pm 1\%$ across participants). Correspondingly, the dependent measure of interest in the present study was RT, which serves as a metric of speech processing efficiency. RT was measured as the delay between the onset of the target word and the participant's keypad response. RT was log-transformed to more closely approximate a normal distribution. Only RTs from correct trials were included in the analysis. Outlier trials with log-transformed RTs exceeding three standard deviations from the participant's mean for that condition were also excluded from RT analysis (0.8% of all trials). Data were analyzed in R using linear mixed effects models implemented in the package lme4, with log-transformed RTs as the dependent variable. Categorical fixed factors included talker variability (single vs. mixed talker) and carrier duration (0, 300, ..., 1500 ms). The model also contained random effect terms for within-participant slopes for talker variability and carrier duration and random intercepts for participants (Barr et al., 2013). In the model design matrix, deviation-coded contrasts were

applied to the talker variability factor, and contrasts for successive differences (i.e., 0 vs. 300; 300 vs. 600; etc.) were applied to the carrier duration factor.

Table 3.1. Response time differences between talker variability conditions (by carrier duration)

Carrier duration (ms)	RT (mean \pm SD ms)			Difference of least square means			
	Single talker	Mixed talker	Difference	β	<i>SE</i>	<i>t</i>	<i>p</i>
0	674 \pm 131	752 \pm 103	78 \pm 70	0.115	0.011	10.40	\ll .0001
300	668 \pm 108	734 \pm 118	66 \pm 67	0.089	0.011	8.00	\ll .0001
600	685 \pm 108	720 \pm 121	35 \pm 43	0.047	0.011	4.22	\ll .0001
900	691 \pm 143	724 \pm 122	33 \pm 58	0.051	0.011	4.62	\ll .0001
1200	699 \pm 118	742 \pm 126	44 \pm 59	0.058	0.011	5.27	\ll .0001
1500	699 \pm 117	742 \pm 114	42 \pm 57	0.061	0.011	5.54	\ll .0001

Significance of fixed factors was determined in a Type III analysis of variance (ANOVA). Significant effects from the ANOVA were followed by post-hoc pairwise analyses using difference of least squares means implemented in the package *lsmeans* in R and testing contrasts on the terms of linear mixed effects model using the package *lmerTest* in R.

Significance of main effects and interactions was determined by adopting the significance criterion of $\alpha = 0.05$, with *p*-values based on the Satterthwaite approximation of degrees of freedom. The ANOVA of the linear mixed effects model of RT revealed a main effect of talker variability such that RTs in the mixed-talker conditions were significantly slower than those in the single-talker conditions overall ($F(1, 23) = 76.41, p \ll 0.0001$). Post-hoc pairwise analysis showed that, within every level of carrier duration, RTs in the mixed-talker condition were significantly slower than the corresponding single-talker condition (**Table 3.1; Figure 3.2A**).

Carrier duration had no significant effect on overall RT ($F(5,23) = 1.10, p = 0.39$).

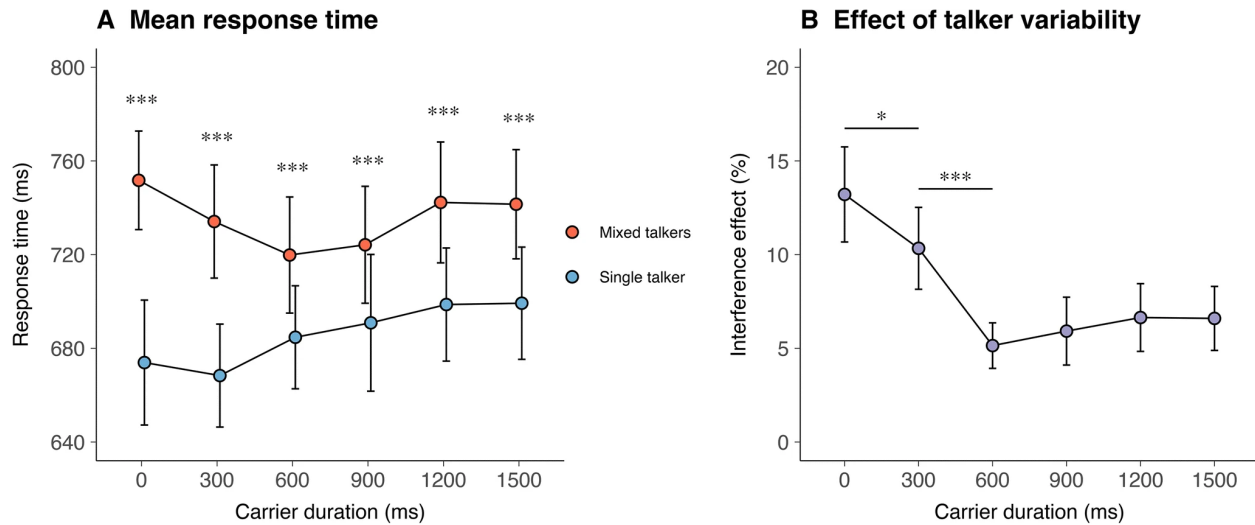


Figure 3.2. Effects of talker variability and carrier duration across talkers on response times. (A) Mean response times in single- and mixed-talker conditions in each carrier condition. (B) The effect of talker variability is shown for each level of carrier duration. The mean response time difference between the mixed- and single-talker conditions is shown, scaled within participant: $((\text{mixed} - \text{single})/\text{single}) \times 100$. Significance of pairwise contrasts is indicated above each line. Error bars indicate ± 1 SEM across participants.

Table 3.2. Interactions between talker variability and carrier duration on log response time.

Carrier duration (ms)	Interaction with talker variability			
	β	<i>SE</i>	<i>t</i>	<i>p</i>
0 vs. 300ms	-0.013	0.006	-2.251	<0.025
300 vs. 600 ms	-0.021	0.006	-3.559	<0.0004
600 vs. 900 ms	0.002	0.006	0.373	0.709
900 vs. 1,200 ms	0.004	0.006	0.616	0.538
1,200 vs. 1,500 ms	0.001	0.006	0.249	0.803

There was a significant interaction between talker variability and carrier duration ($F(5, 13084) = 10.08, p < 0.0001$). Contrast terms on the linear mixed effect model showed that this interaction was significant between the 0- and 300-ms conditions ($\beta = -0.013, s.e. = 0.0059, t = -2.25, p < 0.025$) and between the 300- and 600-ms conditions ($\beta = -0.021, s.e. = 0.0059, t = -3.56, p <$

0.0004), but not for any of the conditions with longer carriers (all $|\beta| < 0.004$, all $|t| < 0.62$, all $p > 0.53$) (Table 3.2, Figure 3.2B).

3.4 Discussion

In this study, we explored to what extent immediately preceding speech from a talker can ameliorate the processing costs that talker variability imposes on word identification. We found that the RT difference between single- and mixed-talker conditions steadily decreased as the duration of the preceding speech context increased from 0 to 600 ms. Beyond 600 ms, however, additional exposure to continuous speech from each talker in the mixed-talker condition did not further facilitate speech processing efficiency compared to the single-talker condition. Processing speed in the mixed-talker condition was always slower than in the single-talker condition, even when the target word was preceded by 1500 ms of continuous speech from the same talker. This piecewise pattern of results – a continuous reduction in mixed-talker interference for connected speech contexts up to 600 ms followed by a constant difference thereafter – is inconsistent with an account of talker adaptation based on a single mechanism. Instead, these data suggest that at least two independent mechanisms are in play: One for rapid adaptation, which continuously improves speech processing efficiency up to ~600 ms of exposure, and a second for sustained expectation of talker variability that operates over longer timescales, such as at least the length of one of the experimental blocks. The first mechanism supporting talker adaptation appears to be a stimulus-driven reorientation of auditory attention that unfolds up to ~600 ms. Within this time frame, listeners experience continuous improvements in speech processing efficiency after encountering a new talker. This is in line with other recent observations that lexical decisions are facilitated immediately after hearing speech from the same talker (Morton et al., 2015; Choi & Perrachione, 2019a; Carter et al., 2019;

Lim et al., 2019a; Kapadia & Perrachione, 2020). There are several reasons to think that these efficiency gains are the result of stimulus-driven reorientation of auditory attention: First, temporal discontinuity between the adapting speech and the target word disrupts this effect (Choi & Perrachione, 2019a), consistent with other evidence that temporal discontinuity interrupts attention to speech (Best et al., 2008; Bressler et al., 2014). Second, this process appears to depend on continuity in the auditory modality, as non-matching or non-auditory primes do not facilitate auditory word recognition (Morton et al., 2015). Third, this process appears to be engaged automatically and independent of listeners' top-down expectations about who the talker will be (Carter et al., 2019). Neurophysiological correlates of speech processing under talker variability also lend support to the idea that a feedforward, attention-based mechanism partially underlies talker adaptation: Abrupt talker discontinuity alters evoked neural responses to auditory onsets and desynchronizes attention-related neural alpha oscillatory power (Mehrai et al., 2018), and talker discontinuities evoke greater pupil dilation responses and larger late cortical potentials associated with distractor suppression (Lim et al., 2021). Similarly, noninvasive electrical stimulation of left temporal lobe disrupts the behavioral facilitation associated specifically with local talker continuity in global mixed-talker contexts (Choi et al., 2019b).

A second mechanism supporting talker adaptation appears to involve changes to the mental computations that support speech processing, which are realized over longer timescales than those involved in feedforward attentional reorientation. It is only during sustained periods of listening to one talker, free from the possibility of having to hear another talker, when listeners seem able to maximize their speech processing efficiency. One proposed difference in speech processing that fits this timeframe is in the extent to which top-down cognitive resources are deployed in anticipation of the processing demands associated with talker variability – a

mechanism described in the active control model of speech processing (Heald, Klos, & Nusbaum, 2015). According to this framework, speech perception is a cognitively active process, in which incoming speech signals are compared against their various possible interpretations, the cognitive-computational demands of which increase in contexts where there is greater acoustic-phonemic uncertainty (Nusbaum & Schwab, 1986; see also Kleinschmidt & Jaeger, 2015).

Several converging lines of evidence suggest that this active control mechanism is best characterized as a difference in mental states, in which listeners either expect to hear speech from a single talker (minimizing the computational demands of speech processing) or from more than one talker (triggering a more computationally demanding, and therefore less efficient, mode of speech processing; Heald & Nusbaum, 2014; Magnuson & Nusbaum, 2007). First, this mechanism operates over a relatively long timeframe and seems to be insensitive to short-term expectations about the talker: Knowing that speech will alternate predictably between two talkers does not improve word recognition efficiency compared to speech from random talkers (Kapadia & Perrachione, 2020). Priming the identity of the upcoming talker in mixed-talker contexts, whether via short auditory (Choi & Perrachione, 2019a) or visual cues (Morton et al., 2015), also does not allow speech to be processed as efficiently as in a single-talker context. Second, this mechanism appears to operate in a categorical fashion: The additional cognitive demands of speech processing are the same regardless of how many different talkers there are beyond one (Kapadia & Perrachione, 2020; Mullennix & Pisoni, 1990). Furthermore, when listeners are performing a task that involves the possibility of talker variability, they remain slower to process speech even during brief periods of talker continuity: Within a mixed-talker context, word recognition during 10-s blocks of speech from a continuous talker remained less efficient than word recognition in longer, single-talker contexts (Stilp & Theodore, 2020), and in a context

where the talker could change randomly on any trial, even 12-s spans of speech from a single talker did not offer additional improvement in speech processing efficiency versus a 4-s span (Lim et al., 2019). Although trial- by-trial talker continuity in an otherwise mixed-talker condition facilitates word recognition, it does not facilitate it as much as listening in a single-talker context (Morton et al., 2015; Choi & Perrachione, 2019a; Kapadia & Perrachione, 2020; Stilp & Theodore, 2020). Generally, the idea that two dissociable cognitive mechanisms might underlie talker adaptation over different timescales parallels the idea that multiple distinct sensory/perceptual mechanisms underlie different aspects of short-term auditory normalization (reviewed in Sjerps et al., 2011), together underscoring the computational complexity of ecological speech processing.

Of these two mechanisms, the shorter-timescale one appears clearly related to stimulus-driven reorientation of auditory attention (reviewed above and in Lim et al., 2021). However, it remains an open question what process or circuit accomplishes the cognitive resource allocation postulated by the active control model. There is some evidence that talker variability poses additional demands on working memory resources (Nusbaum & Morin, 1992; Antoniou & Wong, 2015; Lim et al., 2019b), but brain imaging studies that compare neural activation during speech recognition in single vs. mixed talker contexts find differences almost exclusively in bilateral temporal areas associated with speech processing (Belin & Zatorre, 2003; Wong et al., 2004; Chandrasekaran et al., 2011; Perrachione et al., 2016), not lateral prefrontal areas associated with domain-general cognitive operations (e.g., Fedorenko et al., 2013). This observation may not actually exclude working memory as a mechanism, as there is a growing body of research to suggest that working memory for speech itself may actually rely on the same superior temporal circuits that carry out speech recognition (Perrachione et al., 2017; Scott &

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

Chapter 4. Neurostimulation of left temporal lobe disrupts rapid talker adaptation

Reproduced from

Choi, J. Y., & Perrachione, T. K. (2019). Noninvasive neurostimulation of left temporal lobe disrupts rapid talker adaptation in speech processing. *Brain and Language*, 196(July), 104655. <https://doi.org/10.1016/j.bandl.2019.104655>

4.1 Introduction

Mapping acoustic speech signals onto abstract phonemic representations is a key challenge in speech perception, as the acoustic realization of speech varies substantially across talkers. Thus, when listeners encounter a new talker, they need to quickly ascertain the acoustic-phonemic mappings that correspond to that talker, resulting in an additional processing cost relative to when the talker does not change (Johnson, 2005). The additional processing costs incurred by talker variability have been extensively shown in previous behavioral studies, in which listeners' performance in speech perception tasks gets slower or less accurate when they listen to mixed talkers rather than a single talker (Assmann, Nearey, & Hogan, 1982; Choi, Hu, & Perrachione, 2018; Green, Tomiak, & Kuhl, 1997; Magnuson & Nusbaum, 2007; Mullennix & Pisoni, 1990; Strange, Verbrugge, Shankweiler, & Edman, 1976). Correspondingly, neuroimaging studies have routinely shown that listening to speech from mixed talkers leads to greater activation of superior temporal cortices compared to listening to speech from a single talker (Belin & Zatorre, 2003; Chandrasekaran, Chan, & Wong, 2011; Perrachione et al., 2016; Wong, Nusbaum, & Small, 2004; Zhang et al., 2016). One mechanism by which listeners adapt to a talker is by using the immediately preceding speech context (Johnson, 1990; Nearey, 1989).

Speech in real life almost always occurs in a continuous stream, rather than a word or a speech sound in isolation, and previous speech sounds produced by a talker provide listeners with contextual information about the phonetic space of that talker. Previous studies have shown that preceding speech context biases the decision outcome of speech perception (Johnson, 1990; Ladefoged & Broadbent, 1957) and reduces the processing costs associated with talker variability (Choi & Perrachione, 2019). These empirical results lend support to several related models of speech processing that account for how contextual information is integrated by the perceptual system. Contextual tuning theory treats preceding context as a frame of reference against which the following speech is compared (Nusbaum & Morin, 1992). Under this model, listeners use information embedded in the first speech sounds produced by a new voice to build an internal representation of the vocal tract (i.e., formant space) specific to the talker, which is then used to interpret following speech sounds produced by the same voice. Building upon this theory, Magnuson and Nusbaum (2007) proposed that speech perception is an active control process, in which listeners build hypotheses regarding the interpretation of incoming signals and check them against the speech sounds that they encounter. This process is proposed to be triggered when listeners detect a change of talker and to operate until a stable mapping between the speech sounds produced by the new talker and the listeners' internal phonetic categories is established. In an alternative framework, episodic models of speech perception (e.g., Goldinger, 1998) also highlight the role of previously encountered speech in processing subsequent speech signals. Recently formalized as the ideal adapter framework, this model posits that listeners use cues prior to a speech target to narrow down the range of possible interpretations of incoming speech based on prior experiences with an individual or class of speakers (Kleinschmidt & Jaeger, 2015). Despite the theoretical and empirical work on rapid talker adaptation using

context, the neural mechanisms of talker adaptation still remain elusive. Talker variability is consistently found to increase neural activation in superior temporal lobe (Belin & Zatorre, 2003; Chandrasekaran et al., 2011; Perrachione et al., 2016; Wong et al., 2004), but the causal contribution of this region to processing talker variability is still unknown. Animal models of auditory cortical dynamics and plasticity have elaborated the processes by which neural representations of behaviorally-relevant sounds can be tuned by context over short timescales on the order of seconds (Fritz, Shamma, Elhilali, & Klein, 2003; Froemke, Merzenich, & Schreiner, 2007; Herrmann, Henry, Fromboluti, McAuley, & Obleser, 2015; Jääskeläinen, Ahveninen, Belliveau, Raji, & Sams, 2007). Similar mechanisms may constitute the neurobiological basis for talker adaptation during speech perception by human listeners, but a synapse- or circuit-level understanding of adaptation in speech processing remains beyond the abilities of current human systems neuroscience research. However, a means for studying the causal contribution of larger brain structures in processing talker variability is possible through noninvasive brain stimulation. Transcranial direct current stimulation (tDCS) is a safe, noninvasive technique that modulates cortical excitability and plasticity by employing weak electric currents over the scalp, with anodal stimulation increasing cortical excitability and cathodal stimulation decreasing it (Nitsche & Paulus, 2000). Thus, causal evidence for the involvement of a particular brain area in processing talker variability can be inferred if targeted stimulation of that region results in behavioral changes in speech processing, and the direction and degree of change associated with each polarity of stimulation can better inform us of circuit-level understanding of talker adaptation.

In this study, we aimed to investigate whether the left superior temporal lobe causally underlies the brain's ability to adapt to talkers and, if so, the timescale of its involvement in

talker adaptation. While previous neuroimaging studies have shown that processing speech from multiple talkers vs. a single talker elicits greater response in bilateral superior temporal regions, the source of this increased activation may differ between the two hemispheres: Compared to a single-talker condition, a mixed-talker condition increases not only phonetic variability but also variability in the source of speech (i.e., talker identity). Several studies have specifically contrasted processing talker identity vs. speech content, and have consistently found left-lateralized processing of the verbal content in speech and right-lateralized processing of voice content (e.g., Stevens, 2004; von Kriegstein, Eger, Kleinschmidt, & Giraud, 2003). These results are consistent with the classic finding that phonological processing of speech is mediated by the left hemisphere (Liebenthal, Binder, Spitzer, Possing, & Medler, 2005; Obleser, Zimmermann, Van Meter, & Rauschecker, 2007; Scott, Blank, Rosen, & Wise, 2000; Wernicke, 1874). In a mixed between/within-subjects design, participants were assigned to groups receiving either anodal, cathodal, or sham high-definition (HD) tDCS to left superior temporal lobe while performing a word identification task. All listeners identified which of two phonetically-confusable target words they heard (“boot” or “boat”) while we factorially varied talker variability (single vs. mixed talkers) and speech context (isolated words vs. connected speech). Using participants’ response time to the target word as our dependent variable, we focused on how the speed of word identification changes as a consequence of listening to mixed talkers as opposed to a single talker, and how that difference varies as a function of speech context. This allowed us to explore talker adaptation at two different timescales – within each block (on the order of seconds) and within each trial (on the order of hundreds of milliseconds). Comparing the response time differences between different stimulation groups, we investigated how noninvasive stimulation of left superior temporal lobe influenced talker adaptation. We expected

to replicate the interference effect of talker variability, that response times are slower for mixed- vs. single-talker speech (Choi et al., 2018; Mullennix & Pisoni, 1990), and to replicate the finding that extrinsic talker adaptation leads to a smaller interference effect in connected speech vs. isolated words (Choi & Perrachione, 2019). We expected that anodal stimulation would facilitate talker adaptation, whereas cathodal stimulation would interfere with the process, as anodal stimulation of the left temporal region in healthy individuals has often been shown to improve performance in speech and language domain (reviewed in Zoefel & Davis, 2017). However, it is important to note that the heuristic hypothesis that anodal stimulation enhances, while cathodal stimulation impairs, a target behavior does not necessarily reflect the complex neurobiological mechanisms that electrical stimulation of the cortex affects (Bestmann, de Berker, & Bonaiuto, 2015; Dayan, Censor, Buch, Sandrini, & Cohen, 2013). Finally, we hypothesized that stimulation would affect talker adaptation for connected speech vs. isolated words differently, given the unique role of the left hemisphere in processing connected speech (Peelle, 2012).

4.2 Methods

4.2.1 Participants

Native English-speaking adults (N=60; 46 female, 14 male; age 18–31, M=20.4 years) participated in this study. Participants had no metallic implants and no history of speech, language, hearing, or neurological disorder or significant head trauma. All participants were right-handed as indicated by the Edinburgh Handedness Inventory (Oldfield, 1971). Participants gave informed, written consent approved and overseen by the Institutional Review Board at Boston University.

4.2.2 Stimuli

Stimuli included two target words, “boot” and “boat.” We chose these words because the acoustic-phonemic correspondence of the /u/-/o/ contrast is highly talker-dependent; the acoustic realization of the vowels /u/ and /o/ exhibits extensive overlap across talkers that listeners must resolve on a talker-specific basis to correctly identify the target phoneme (Hillenbrand et al., 1995) and therefore imposes greater processing interference in a mixed-talker environment (Choi et al., 2018). Target words were presented either in isolation or in connected speech, where they were preceded by the carrier phrase “I owe you a [boot/boat].” This carrier phrase was chosen because it provides an extensive sample of each talker’s vowel space (**Figure 4.1A**), offering listeners talker-specific phonetic details that they can use to calibrate their perception of the vowel in the following target word (Johnson, 1990; Joos, 1948; Nusbaum & Morin, 1992). Words and carrier phrases were recorded by two male and two female native speakers of American English (**Figure 4.1A**). The recordings were made in a sound-attenuated room with a Shure MX153 earset microphone and Roland Quad Capture sound card sampling at 44.1 kHz

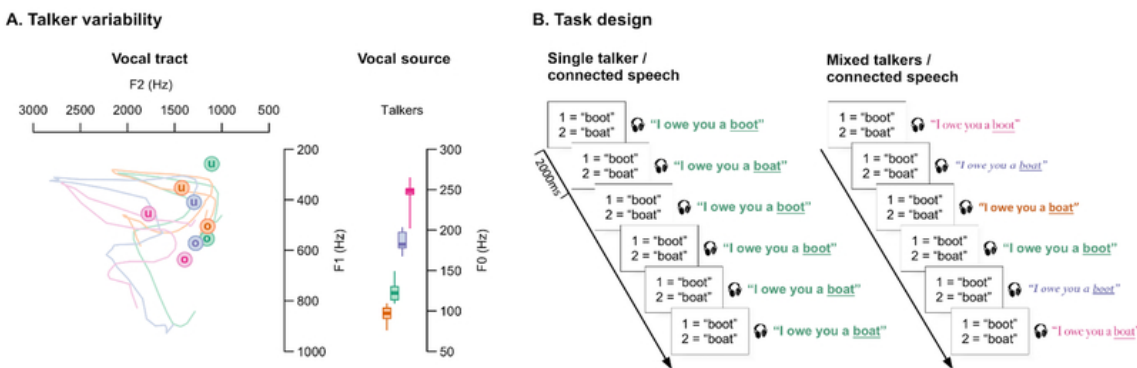


Figure 4.1. Stimulus variability across talkers and task design. (A) Phonetic variability of stimuli across talkers. Left: F1 and F2 of the target words (circles; /u/ boot, /o/ boat) and the F1-F2 trajectory of the carrier phrase (lines; “I owe you a”). Right: F0 (vocal pitch) distribution for all talkers’ recordings. Colors denote different talkers. (B) Behavioral task design. Participants identified words while listening to speech produced by either a single talker (left) or mixed talkers (right). The connected speech conditions are shown. Font/color combinations denote different talkers.

and 16bits. Among numerous tokens from these speakers, the recordings in which the boot / boat distinction was most evident based on their formant frequencies – and which were least dissimilar in noncontrastive features such as voice pitch, amplitude envelope, and duration – were chosen as the final stimulus set. The mean duration of the target words was 228 ms (range: 203–256 ms), and the mean duration of the prepended carrier phrases was 609 ms (range: 543–656 ms). Connected speech sentences were synthesized by concatenating the naturally-recorded carrier phrase to the target word, so that the same target word stimuli from each talker were used in all conditions. Carrier phrases and target words were normalized to 65 dB SPL RMS amplitude in Praat (Boersma, 2001).

4.2.3 Behavioral task

Participants’ task on each trial was to listen to the stimulus and indicate whether they heard “boot” or “boat” as quickly and accurately as possible by pressing the corresponding number on the keypad. Trials were organized into four blocks that factorially manipulated talker variability (single-talker vs. mixed-talker) and speech context (isolated words vs. connected speech), with each block corresponding to one of the four conditions. Each block consisted of 96 trials, with each target word occurring in 48 trials per block. Stimulus presentation was pseudo-randomized such that the same target word was not presented for more than three consecutive trials (**Figure 4.1B**). The order of conditions was counterbalanced across participants using Latin-square permutations. For each participant, the same talker served as the single talker in both single-talker blocks, and which of the four talkers was used in the single-talker conditions was counterbalanced across participants. The duration of each trial, including the duration of the stimulus and the time for participants to respond with the keypad, was kept at 2000 ms across all

conditions. Stimulus delivery was controlled using PsychoPy v.1.8.1 (Peirce, 2007). The total experiment duration was approximately 13 minutes.

4.2.4 High-definition transcranial direct current stimulation (HD-tDCS)

In a between-subjects design, participants were randomly assigned to receive either sham ($n=20$), anodal ($n=20$), or cathodal ($n=20$) HD-tDCS during the task. Stimulation was applied using a Soterix M \times N HD-tDCS system. Stimulating electrodes (cathodes for the cathodal condition, anodes for the anodal condition) were placed at electrode locations T7 and TP7 in the 10–10 system (Klem, Lüders, Jasper, & Elger, 1999); return electrodes (anodes for the cathodal condition, cathodes for the anodal condition) were placed at C3, CP3, PO7 and F7 (**Figure 4.2A**). This configuration, which approximates the center-surround stimulation design that has been shown to be optimal for achieving maximally focal stimulation intensity and current flow (Datta et al., 2009; Kuo et al., 2013), was chosen to focally target left superior temporal cortex.

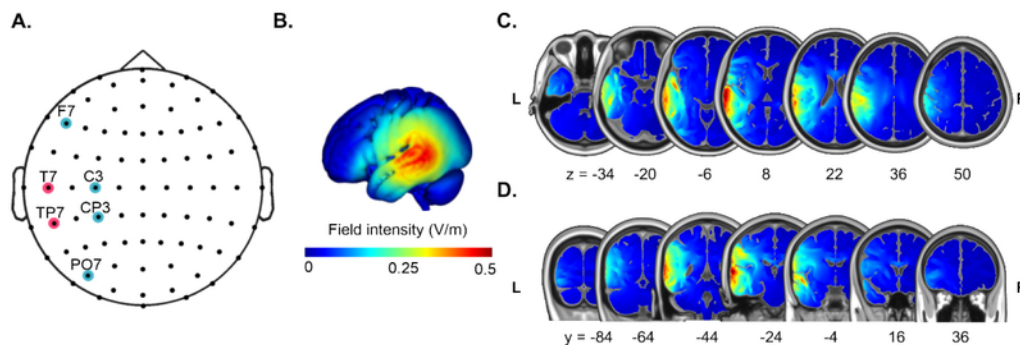


Figure 4.2. tDCS paradigm. (A) Electrode configuration. Stimulating electrodes are shown in red; reference electrodes are shown in blue. Simulated current flow estimated by HD-Explore in (B) 3D view, (C) axial view, and (D) coronal view. The y- and z-coordinates refer to the slice location in MNI stereotaxic space. Slices are shown in neurological convention.

Electrode locations were selected based on biophysical simulation of current flow in the human brain (Soterix HD- Explore, Soterix Medical, NY, USA). Peak estimated field intensity at the target location was 0.507 V/m (Fig. 2B–D). For anodal and cathodal HD-tDCS sessions, current

was increased to the maximum stimulation intensity of 2 mA using a 30-s linear ramp after initiation. Stimulation magnitude remained at 2 mA for the entire duration of the task (~13 min), followed by a 30-s linear ramp-down at termination. For sham HD-tDCS sessions, current was linearly ramped up to 2 mA over 30 s and then immediately ramped back down to 0 mA over 30 s, where it remained for the entire duration of the task. Sham HD-tDCS induces the initial mild dermal tingling sensation associated with HD-tDCS without stimulating the brain areas below the electrodes during the task, thus keeping participants unaware as to whether they were assigned to an active stimulation or sham control condition. Participants filled out a questionnaire after completing the experiment to ensure that HD-tDCS did not cause excessive discomfort. Electrode resistance was kept below 10 k Ω for all electrodes for all sessions.

4.2.5 Data analysis

Accuracy and response time data were analyzed for each participant in each condition. Accuracy was calculated as the proportion of trials in which the participant correctly identified the target words out of the total number of trials. Response times were log-transformed to more closely approximate a normal distribution expected by the model. Only response times from correct trials were analyzed. Outlier trials deviating from the mean log response time in each condition by more than three standard deviations were excluded from analysis (<1% of trials). Participants' response times were analyzed using a linear mixed-effects model with fixed factors including speech context (isolated words vs. connected speech), talker variability (single- vs. mixed-talker), and stimulation (anodal vs. cathodal vs. sham), and with random effects including by-participant intercepts and by-participant slopes for the effects of context and variability. Significance of factors was determined in a Type III analysis of variance (ANOVA). Significant effects from the ANOVA were followed by post-hoc pairwise analyses by testing contrasts on

the terms in the linear mixed-effects model using the package *lmerTest* in R. Contrasts were treatment-coded, with baseline levels of isolated words (*speech context*), single-talker (*talker variability*), and sham (*stimulation*). Significance of main effects and interactions was determined by adopting the significance criterion of $\alpha = 0.05$, with p -values based on the Satterthwaite approximation of the degrees of freedom.

4.3 Results

Participants' word identification accuracy was at ceiling ($98\% \pm 2\%$), with no effect of stimulation condition on participants' accuracy. As this study was primarily designed to investigate speech processing efficiency, the principal dependent measure was response time (**Table 4.1**). In the post-experiment questionnaire, the number of participants who reported scalp sensations related to HD-tDCS did not differ between sham and active (combined anodal and cathodal) stimulation groups ($\chi^2(1) = 1.68, p = 0.19$). Participants reported mild to moderate tingling (84% of all participants); mild pain (36%), and mild burning sensations (29%). The number of participants reporting each type of sensation did not differ between sham and active stimulation groups (tingling $\chi^2(1) = 0.19, p = 0.67$; pain $\chi^2(1) = 1.49, p = 0.22$; burning $\chi^2(1) = 0.22, p = 0.64$). The lack of group difference in these responses suggests that participants were effectively blinded as to whether they received active or sham stimulation.

Table 4.1 Mean \pm s.d. response time (ms) in each condition

	Sham		Anodal		Cathodal	
	Isolated Words	Connected Speech	Isolated Words	Connected Speech	Isolated Words	Connected Speech
Single-talker	745 \pm 104	679 \pm 81	700 \pm 76	654 \pm 75	717 \pm 85	645 \pm 59
Mixed-talker	836 \pm 122	708 \pm 78	780 \pm 87	702 \pm 79	805 \pm 100	697 \pm 58
Difference	91 \pm 66	29 \pm 49	79 \pm 48	48 \pm 51	88 \pm 82	52 \pm 49

4.3.1 Interference effects of talker variability

The ANOVA of the linear mixed-effects model revealed a robust main effect of *talker variability* ($F(1, 57) = 156.19; p \ll 0.0001$), showing that response times in the mixed-talker conditions were significantly slower than the single-talker conditions overall. Response times in the connected-speech conditions were also significantly faster overall compared to the isolated-word conditions (main effect of speech context; $F(1, 57) = 98.15; p \ll 0.0001$). We observed a significant speech *context* \times *talker variability* interaction effect ($F(1, 22275) = 89.74; p \ll 0.0001$), indicating that the magnitude of processing interference from the mixed-talker condition differed depending on whether the target words were embedded in continuous speech or presented in isolation. Listeners exhibited significantly more interference from talker variability when recognizing words in isolation than in connected speech.

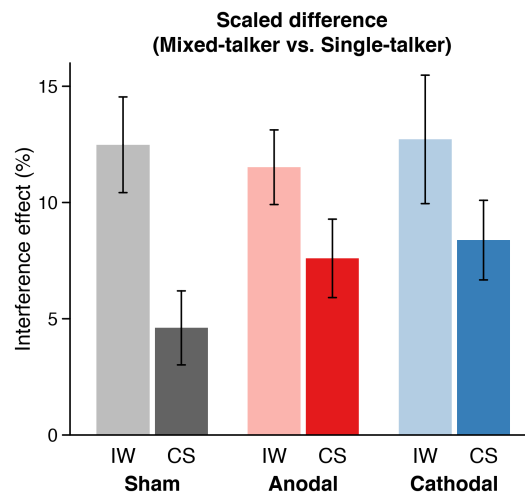


Figure 4.3. Processing cost of talker variability by speech context and stimulation condition. Mean interference effects of talker variability for isolated words (IW) and connected speech (CS) in each stimulation condition. Taller bars reflect greater differences in word identification response time between the mixed- vs. single-talker conditions. The interference effect of talker variability is calculated as the scaled difference between the average response time (RT) in mixed-talker condition and the single-talker condition: $((\text{mixed} - \text{single}) / \text{single}) \times 100$. Error bars indicate standard error of mean across participants.

4.3.2 Effects of neurostimulation on talker adaptation

The HD-tDCS manipulation did not have a significant effect on overall response time (no main effect of stimulation; $F(2, 57)=1.03$; $p=0.36$). There was also no significant stimulation \times talker variability interaction ($F(2, 57)=0.40$; $p=0.67$), nor *stimulation \times speech context* interaction ($F(2, 57)=1.14$; $p=0.33$). Critically, there was a significant *stimulation \times speech context \times talker variability* interaction ($F(2, 22275)=5.33$; $p < 0.01$), indicating that the amount of benefit obtained from connected speech under talker variability differed among the three stimulation conditions (Fig. 3). To understand the three-way interaction across three levels of the stimulation factor, we turned to the pairwise contrasts on the three-way interaction terms of the linear model: The *talker variability \times speech context \times stimulation* interaction was significant for anodal vs. sham ($\beta = 0.0038$, $s.e. = 0.0013$, $t = 2.97$, $p < 0.01$) and cathodal vs. sham ($\beta = 0.0034$, $s.e. = 0.0013$, $t = 2.65$, $p < 0.01$) stimulation. This indicates that the effect of connected speech on mitigating the interference effect of mixed talkers was smaller under anodal and cathodal stimulation conditions than under sham stimulation. Furthermore, in models on subsets of the data examining only the single- and mixed-talker conditions separately, the *stimulation \times speech context* interaction effect for mixed talkers was nearly three times larger than the respective effect for a single talker (Anodal: $\beta_{\text{interact.}}=0.06$ (50 ms, mixed) vs. 0.02 (20 ms, single); Cathodal: $\beta_{\text{interact.}}=0.03$ (20 ms, mixed) vs. 0.01 (-6 ms, single)). That is, compared to sham, HD-tDCS disrupted the brain's ability to use the immediately preceding speech context to rapidly adapt to each new talker in a mixed-talker context. In the isolated words condition alone, however, the magnitude of the talker variability effect (mixed vs. single talkers) was not affected by either of the active stimulation conditions compared to sham (the contrast on the *stimulation \times*

talker variability interaction term for isolated words only; sham vs. anodal $\beta = 0.014$, *s.e.* = 0.018, $t = 0.76$, $p = 0.45$; sham vs. cathodal $\beta = 0.0023$, *s.e.* = 0.018, $t = 0.13$, $p = 0.90$).

4.4 Discussion

In this study, we used noninvasive neurostimulation to investigate the causal role of left superior temporal lobe in talker adaptation. We observed a significant interaction with stimulation such that, compared to sham, both anodal and cathodal stimulation disrupted rapid talker adaptation in connected speech. When processing isolated words, however, the three different types of stimulation did not differentially affect processing efficiency between single- and mixed-talker speech. These results raise the possibility that there is a dissociation between two timescales of—or mechanisms for—adaptation to a talker using preceding speech context, in which disruption of neurocomputational processes in left superior temporal lobe impairs the brain's ability to rapidly adapt to a talker on a timescale as short as within a sentence (< 1 s), but not its ability to adapt over longer timescales.

4.4.1 Causal involvement of left superior temporal region in rapid talker adaptation

Our observations extend previous fMRI studies that have reported reduced activation in superior temporal areas in single-talker blocks relative to mixed-talker blocks (i.e., neural adaptation effects) when subjects performed tasks similar to our isolated-word condition (Belin & Zatorre, 2003; Chandrasekaran et al., 2011; Perrachione et al., 2016; Wong et al., 2004; Zhang et al., 2016). In addition to the correlation between speech processing behavior and neural activity established by those previous neuroimaging studies, we found that the extent to which connected speech can offset the interference effect of mixed talkers was disrupted by electrical stimulation of left superior temporal lobe. This result appears to be specific to rapid integration

of context information during talker adaptation from connected speech rather than a more general effect on speech processing efficiency. Increasing (or decreasing) cortical excitability of left superior temporal lobe via non-invasive neurostimulation did not generally speed up (or slow down) speech processing, neither overall nor in either speech context separately. This pattern of results suggests that left superior temporal lobe is causally involved in rapid integration of context information during connected speech. Thus, the early integration of talker and speech information likely occurs in this structure, where neural response differences between single- and mixed-talker speech likely reflect the additional computational demands in processing talker variability (Kaganovich, Francis, & Melara, 2006). In two conditions of this study, we preceded the target words with a carrier phrase to provide listeners with talker-specific vocal and phonetic details, giving them an extrinsic context from which they could develop expectations about the correspondence between speech acoustics and phonemic categories (Johnson, 1990; Magnuson & Nusbaum, 2007; Nusbaum & Morin, 1992). Auditory expectations sharpen neural responses to relevant stimulus features (Fritz et al., 2003; Todorovic, van Ede, Maris, & de Lange, 2011), which may underlie our behavioral outcomes showing overall faster response times when the target words were preceded by an adapting carrier phrase. Although we specifically operationalized acoustic-phonemic ambiguity as differences in speech phonetics across talkers, it is possible that the computations carried out by superior temporal lobe may contribute to resolving phonetic ambiguity more generally. For instance, in a phonetic category judgment task, recruitment of superior temporal lobe bilaterally is greater when listeners are less certain of phonetic category membership (Myers, 2007), suggesting that this region may be the locus of resolving variable acoustic-phonemic mappings even when the source of variability is not related to differences across talkers. However, neuroimaging studies of processing variability in speech

perception have also almost exclusively operationalized speech variability as phonetic variability between talkers, and future work must ascertain whether analogous normalizing processes also underlie within-talker variation arising from, for example, speech rate or coarticulation. These results also broach the question of whether talker adaptation comprises neurocomputational processes specific to speech processing or reflects a more domain-general phenomenon underlying auditory adaptation. Even non-speech extrinsic contexts have been shown to affect speech processing in a manner similar to talker adaptation (Laing, Liu, Lotto, & Holt, 2012; Sjerps, Mitterer, & McQueen, 2011), demonstrating that auditory perceptual adaptation to speech during talker adaptation may actually be occurring via more fundamental auditory processes underlying stimulus adaptation (Herrmann et al., 2015). Correspondingly, as we discuss below, stimulation of left superior temporal lobe may ultimately be affecting feedforward adaptation of auditory circuits, rather than computations specific to speech processing.

4.4.2 Effects independent of stimulation polarity

Behaviorally, there was no difference in the effect of stimulation between anodal and cathodal polarities, which are thought to increase and decrease cortical excitability, respectively (Nitsche & Paulus, 2000). This may be due to the fact that rapid re-tuning of auditory perception relies on the precise (re-)balancing between excitatory and inhibitory activity, rather than a unidirectional process. Although the behavioral effects of the two polarities were similar, the mechanism by which HD-tDCS disrupts talker adaptation may nonetheless differ: anodal stimulation may reduce the balanced precision between excitation and inhibition that underlies neocortical adaptation (Wehr & Zador, 2003), resulting in less precise re-tuning and thereby reducing perceptual efficiency. Cathodal stimulation, meanwhile, may reduce the magnitude of short-term changes to synaptic weights (Froemke et al., 2007), making them less specific.

Application of a small electric current over the scalp, anodal or cathodal, may have interrupted this balance between excitation and inhibition in different ways, thus degrading the facilitatory effect of feedforward stimulus continuity on perception. Moreover, as the effect of HD-tDCS varies depending on various factors such as simultaneity between stimulation and the task, stimulation magnitude and duration, electrode placements, and cognitive load (Ohn et al., 2008; Roe et al., 2016; Thair, Holloway, Newport, & Smith, 2017), HD-tDCS polarity effects in cognitive domains cannot simply be reduced to an “anodal-excitation and cathodal-inhibition” heuristic (Jacobson, Koslowsky, & Lavidor, 2012). Indeed, both anodal and cathodal stimulation of auditory cortex have been shown to increase the magnitude of various auditory evoked potentials (Zaehle, Beretta, Jäncke, Hermann, & Sandmann, 2011).

4.4.3 No effect of stimulation on talker adaptation to isolated words

In single-talker blocks, listeners can benefit from using the same talker-specific acoustic-phonemic mappings on every trial, even when they are listening to isolated words. When there is context that immediately precedes the target words, the processing costs associated with mixed-talker speech are reduced, because listeners can rapidly ascertain some talker-specific cues from the local context, even when the talker differs from the previous trial (Choi & Perrachione, 2019). By using both the isolated-word and connected-speech conditions in this experiment, we were able to investigate how the left superior temporal region is involved in talker adaptation on varying timescales. Our study showed that anodal and cathodal stimulation of left superior temporal lobe reduced the benefit of adaptation on short timescales (i.e., for connected speech) but did not reduce the adaptation effect on longer timescales (i.e., for isolated words). Since neurostimulation revealed no causal role of left superior temporal region in talker adaptation on the scale of seconds, such adaptation may be mediated by other brain regions. In addition to the

superior temporal lobe, Wong et al. (2004) found talker adaptation-related activation in superior parietal lobe. They suggested activation in this region may reflect the additional cognitive effort demanded by constant attentional reorientation to new talkers in mixed-talker blocks. Future work will need to assess whether applying noninvasive neurostimulation to superior parietal lobe will affect talker adaptation to isolated words, as predicted by the attentional-reorientation hypothesis. That left hemisphere stimulation did not affect talker adaptation from isolated words may also be due to hemispheric differences in temporal integration of connected vs. unconnected speech information. For instance, Peelle (2012) advances the idea that differences in left-lateralized vs. bilateral responses to speech depend primarily on whether speech is encountered in a connected (i.e., phrasal or sentential) vs. unconnected (i.e., individual words or syllables) context. Such a framework is consistent with our results, where left hemisphere HD-tDCS disrupted talker adaptation in a connected speech context, but not in an isolated word context, where the right hemisphere's putative role in processing unconnected speech was undisrupted. This pattern of results is also consistent with a longstanding supposition that the two cerebral hemispheres may be involved in integrating auditory information on different timescales (e.g., Abrams, Nicol, Zecker, & Kraus, 2008; Boemio, Fromm, Braun, & Poeppel, 2005; Zatorre & Belin, 2001), notwithstanding what those particular timescales may be. Future work is thus clearly needed to explore how HD-tDCS of right superior temporal lobe also affects talker adaptation, and whether it does so for connected vs. unconnected speech contexts.

4.4.4 Limitations and future directions

In our application of HD-tDCS to left superior temporal cortex, we observed a three-way interaction between stimulation, talker variability, and speech context, revealing a causal involvement of left superior temporal cortex in talker adaptation during connected speech.

However, it is important to note the large number of degrees of freedom that are available in the design and implementation of brain stimulation studies, including details of the behavioral paradigm, as well as the location, magnitude, and polarity of electrical stimulation.

Consequently, future work remains to both replicate and extend the observations from this study.

Our behavioral paradigm involved manipulations that affect speech processing efficiency (Choi & Perrachione, 2019; Choi et al., 2018). However, talker adaptation affects not only speech processing efficiency, but also the phonological and lexical decision outcomes of speech perception (Francis, Ciocca, Wong, Leung, & Chu, 2006; Johnson, 1990; Kleinschmidt & Jaeger, 2015; Laing et al., 2012). Similarly, talker variability during encoding has differential effects on short-term vs. long-term memories for speech (Lim, Shinn-Cunningham, & Perrachione, 2019; Palmeri, Goldinger, & Pisoni, 1993). Future work is therefore needed to understand how left superior temporal lobe is causally involved in recalculating acoustic-phonemic correspondences associated with talker adaptation, and how its role in talker-adaptation processes affects short- and long-term memories for speech. We found similar behavioral effects of anodal and cathodal stimulation on speech processing efficiency, but hypothesized that the mechanistic bases for these disruptions were nonetheless differentiable.

While the present study used a between-subjects design to parsimoniously establish the efficacy of HD-tDCS in studying talker adaptation, this design choice nonetheless precluded the ability to compare the relative effects of anodal vs. cathodal stimulation within individual participants.

Future studies may be able to gain better mechanistic insight into how and why these polarities differentially disrupt talker adaptation by comparing effect sizes under a within-subjects design.

Finally, although we found that HD-tDCS of left superior temporal cortex induced a significant and context-specific disruption of talker adaptation, this does not preclude the possibility that

other areas of the brain are also causally involved in speech adaptation, or that this region also participates in speech adaptation on other timescales. Stimulation of other sites implicated in talker adaptation (especially the right superior temporal lobe (Zhang et al., 2016; Belin & Zatorre, 2003; Perrachione et al., 2016) and superior parietal lobe (Wong et al., 2004)) must be undertaken in future studies. Similarly, these results should be validated by stimulation at other intensities and using other stimulation paradigms (e.g., transcranial alternating current stimulation) or technologies (e.g., transcranial magnetic stimulation) to replicate and extend our observation of a causal, context-specific role for left superior temporal cortex in talker adaptation.

4.5 Conclusions

The results from this study show that noninvasive neurostimulation of left superior temporal lobe interferes with the usage of local phonetic context to adapt to a talker and enhance speech processing efficiency, demonstrating that this region is causally involved in rapid talker adaptation.

Chapter 5. Conclusions

5.1 Summary

This dissertation examined the cognitive and neural mechanisms underlying listeners' resolution of talker variability in speech perception. As observed repeatedly in previous studies, talker variability imposes additional processing demands in listeners' speech perception. The magnitude of this cognitive processing cost, manifested as increased response times in mixed-talker speech relative to single-talker speech, was observed to be influenced by various factors of the preceding speech context. Specifically, this dissertation explored how the process of talker adaptation is implemented over time as a set of speech signals unfolds, as well as the type of information that the process requires.

Chapter 2 presented a study that analyzed the information of the immediately preceding carrier speech that gets incorporated into listeners' perceptual system in order to adapt to different talkers, and how the process operates as the carrier speech unfolds over time. Across all three experiments that factorially manipulated the duration, richness of phonetic detail, and temporal continuity of the carrier phrases along with talker variability, listeners were (i) significantly slower at word identification in the mixed-talker conditions than in the single-talker conditions, and (ii) more efficient with processing mixed-talker speech when the target word was presented with carrier speech than in isolation. The first experiment showed that a carrier phrase of 300 to 600 ms preceding the target speech facilitates talker adaptation relative to when the target speech was presented in isolation, and that the longer carrier speech had a more facilitatory effect on talker adaptation than the shorter carrier. The results suggest that the preceding speech phrase influences the efficiency of processing talker variability and that it does so as a function of the amount of speech context available. Delving further into precisely why the longer carrier

afforded more facilitation of talker adaptation than the short carrier did, the second experiment used two carriers that had the same duration but different richness of phonetic detail – one containing at least five distinct vowels, and the other being a static snapshot of one vowel. We observed that the amount of facilitation afforded by these two different carriers in the second experiment did not differ, suggesting that the richness of phonetic detail in the carrier speech did not have a significant effect on resolving talker variability while the duration of the carrier speech did. The third experiment attempted to adjudicate between different models of speech perception that purport to account for talker variability by manipulating temporal continuity between the carrier speech and the target word. The results showed that the carrier that is temporally continuous to the target speech had more facilitatory effect on talker adaptation than a carrier that was presented with a delay afterwards. These findings suggest that the effect of context on talker adaptation is a manifestation of auditory attentional re-orientation rather than a real-time computation of each talker's speech articulation that is facilitated by rich phonetic detail about the talker, or a ballistic activation of episodic memory associated with the talker.

The findings presented in Chapter 2 – that only a minimal amount of detail about the talker's speech production immediately prior to the target word can facilitate talker adaptation and that it does so as a function of its duration – motivated the following question: Would a simple carrier speech context with a sufficiently long duration make the processing of mixed-talker speech as efficient as single-talker speech, and, if so, how long does the carrier speech need to be? Chapter 3 presented an experiment where participants identified target words that were presented in isolation or preceded by a simple carrier vowel that varied in its duration. The results showed that the facilitatory effect of carrier speech increases as the duration of the carrier increases to 600 ms, but lengthening the duration beyond 600 ms did not further facilitate talker

adaptation. This set of results suggest that there are two different mechanisms of talker adaptation occurring in parallel: one feedforward mechanism that occurs on the scale of approximately 600 ms where the detection of talker variability and the continued stream of the talker allows for an attentional reorientation, and another top-down cognitive mechanism that operates in a longer timescale when talker variability is detected.

Chapter 4 examined the neural mechanism underlying talker adaptation, specifically the causal role of left STG in the utilization of carrier speech to resolve talker variability. Neurostimulation of left STG did not change the general pattern of listeners being slower at word identification in mixed-talker conditions relative to single-talker conditions. However, it did significantly reduce the facilitatory effect afforded by the carrier speech in resolving talker variability. Both anodal and cathodal stimulation had the same effect compared to sham stimulation, suggesting that electric stimulation, regardless of its polarity, may have disrupted a precise balance of neuronal activity in left STG that supports rapid talker adaptation. The effect of tDCS on talker adaptation was observed selectively in the effect of carrier speech, lending support to the proposition that the two hemispheres of the brain may operate on different timescales in integrating auditory information.

Together, these studies suggest that resolving talker variability in speech perception involves a feedforward auditory attentional reorientation that operates in a shorter timescale, and a feedback control mechanism that operates in a longer timescale when talker variability is detected or expected.

5.2 Future directions

In this dissertation, various factors of carrier speech have been investigated as potential elements that can affect the efficiency with which listeners identify the target word in mixed-talker settings. While our choice of target words and carrier phrases was intentional such that they optimize our observation of processing costs of talker variability, empirical studies are necessary to understand how generalizable our observations are to other types of stimuli, as real-life speech perception involves listening to a much less predictable set of words and sentences and thus making perceptual decisions in a considerably bigger decision space than the stimuli used in the studies in Chapters 2, 3, and 4. Also, auditory attention has been proposed as a potential explanation of talker adaptation, but the set of studies presented in this thesis did not use some of the classical methods of studying auditory attention, such as eliciting selective attention to one out of multiple simultaneous sources of auditory stimuli. To further develop a more sophisticated attentional framework of talker adaptation, it is necessary to study how listeners process talker variability in an auditory scene more complex than one talker producing speech sequentially after another talker without any auditory distractor. Furthermore, it would also be an important task to understand what the decision bias that results from extrinsic context (e.g., Johnson, 1990) means in an attentional framework of talker adaptation.

As laid out in Chapter 4, there are many directions in which neurostimulation can further enlighten the role of brain regions in resolving talker variability. Studies have found increased neural activity associated with talker variability in the right superior temporal lobe (Zhang et al., 2016; Belin & Zatorre, 2003; Perrachione et al., 2016). Case studies have presented patients with right hemisphere damage who show deficits in vocal identity processing but not in speech perception or voice discrimination (Luzzi et al., 2018; Van Lancker & Canter, 1982), suggesting

that vocal identification is a process that is at least partially dissociable from speech perception and resolution of talker variability. Using neurostimulation on the right hemisphere may be able to further enlighten how the brain processes varying speech signals to resolve talker variability. Based on the findings from the study presented in Chapter 3, the top-down cognitive control component of talker adaptation in longer timescale can also be further investigated via studying the role of frontal regions as well.

As speech perception involves more than the left STG that was studied in this dissertation, exploring how the structural and functional connectivity of the brain supports talker adaptation may also further our understanding of how the brain processes phonetic information and talker information and how the two integral sides of speech signals are integrated or separated depending on the task. Talker identification and speech comprehension are two different tasks that listeners can accomplish from the same speech signal; these tasks also interact with each other. An fMRI study by von Kriegstein and colleagues (2010), for example, revealed that bilateral STG/STS are sensitive to changing talkers, and that the functional connectivity between the left and right STG/STS is stronger when listeners are processing mixed talkers than when processing single talker. Further exploring how a network of regions operate may provide a more accurate picture of the neural mechanism underlying the solution of lack of invariance problem in speech perception.

Although the scope of this dissertation specifically is restricted to how listeners' resolution of the lack of invariance problem affects speech processing efficiency, humans face similar challenge in other perceptual domains in general. In vision, for example, recognizing an object requires resolving variation that results from different colors and sizes of the same type of object; the visual context that surrounds the object such as the direction of illumination,

brightness of the environment, viewing orientations and occlusion from other objects; and how prototypical each instance of an object is with regards to its category. While there are characteristics that make speech perception different from perceiving other types of stimuli that we encounter, it would be enlightening to understand the interaction between what underlies the general challenge of perceptual system versus what specifically supports resolution of variability in speech perception.

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.
- Baddeley, A. D. (1986). *Working memory*. Clarendon Press.
- Baddeley, A. D. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, 4(10), 829–839.svr
- 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." *Glott International*, 5, 341-345.

- Bregman, A. S. (1990). *Auditory scene analysis*. MIT Press.
- Bressler, S., Masud, S., Bharadwaj, H., & Shinn-Cunningham, B. (2014). Bottom-up influences of voice continuity in focusing selective auditory attention. *Psychological Research*, 78(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.
- Carter, Y. D., Lim, S.-J., & Perrachione, T. K. (2019). Talker continuity facilitates speech processing independent of listeners' expectations. Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia.
- Chandrasekaran, B., Chan, A.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.
- Choi, J. Y., & Perrachione, T. K. (2019). Time and information in perceptual adaptation to speech. *Cognition*, 192.
- 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.
- Datta, A., Bansal, V., Diaz, J., Patel, J., Reato, D., & Bikson, M. (2009). Gyri-precise head model of transcranial direct current stimulation: improved spatial focality using a ring electrode versus conventional rectangular pad. *Brain Stimulation*, 2(4), 201–207.
- 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.
- Fedorenko, E., Duncan, J., & Kanwisher, N. (2013). Broad domain generality in focal regions offrontal and parietal cortex. *Proceedings of the National Academy of Sciences*, 110(41), 16616–16621.
- Flege, J. E. (1981). The phonological basis of foreign accent: A hypothesis. *TESOL quarterly*, 15(4), 443-455.
- Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance*, 32(5), 1276–1293.
- 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.
- Gerstman, L. J. (1968). Classification of self-normalized vowels. *IEEE Transactions on Audio Electroacoustics*, AU-16,78–80.
- 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.
- Heald, S. L. M., Klos, S., & Nusbaum, H. C. (2015). Understanding speech in the context of variability. In G. Hickok & S. Small (Eds.), *Neurobiology of language* (pp. 195–206). Academic Press.

- Heald, S. L. M., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, 8, 35.
- Herrmann, B., Henry, M. J., Fromboluti, E. K., McAuley, J. D., & Obleser, J. (2015). Statistical context shapes stimulus-specific adaptation in human auditory cortex. *Journal of Neurophysiology*, 113(7), 2582–2591.
- 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.
- Holmes, E., & Johnsrude, I. S. (2020). Speech spoken by familiar people is more resistant to interference by linguistically similar speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(8), 1465.
- 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.
- Jacobson, L., Koslowsky, M., & Lavidor, M. (2012). tDCS polarity effects in motor and cognitive domains: a meta-analytical review. *Experimental Brain Research*, 216(1), 1-10.
- Jacquemot, C., & Scott, S.K. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Science*, 10, 480–486.
- 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.
- Kapadia, A. M., & Perrachione, T. K. (2020). Selecting among competing models of talker adaptation: Attention, cognition, and memory in speech processing efficiency. *Cognition*, 204, Article 104393.
- 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.

- Kiesel, A., Steinhauer, M., Wendt, M., Falkenstein, M., Jost, K., Philipp, A. M., & Koch, I. (2010). Control and interference in task switching—A review. *Psychological Bulletin*, *136*, 849–874.
- Kleinschmidt, D. F. (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.
- Klem, G. H., Lüders, H. O., Jasper, H. H., & Elger, C., (1999). The ten-twenty electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology*, *52*(3), 3–6.
- Koenig, L. L. (2000). Laryngeal factors in voiceless consonant production in men, women, and 5-year-olds. *Journal of Speech, Language, and Hearing Research*, *43*, 1211–1228.
- Koenigs, M., Acheson, D. J., Barbey, A. K., Solomon, J., Postle, B. R., & Grafman, J. (2011). Areas of left perisylvian cortex mediate auditory-verbal short-term memory. *Neuropsychologia*, *49*(13), 3612–3619.
- Kuo, H. I., Bikson, M., Datta, A., Minhas, P., Paulus, W., Kuo, M. F., & Nitsche, M. A. (2013). Comparing cortical plasticity induced by conventional and high-definition 4 × 1 ring tDCS: A neurophysiological study. *Brain Stimulation*, *6*(4), 644–648.
- 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*.
- Leff, A.P., Schofield, T. M., Crinion, J. T., Seghier, M. L., Grogan, A., Green, D. W., & Price, C. J. (2009). The left superior temporal gyrus is a shared substrate for auditory short-term memory and speech comprehension: Evidence from 210 patients with stroke. *Brain*, *132*, 3401–3410.
- Lieberman, 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., Carter, Y. D., Njoroge, J. M., Shinn-Cunningham, B. G., & Perrachione, T. K. (2021). Talker discontinuity disrupts attention to speech: Evidence from EEG and pupillometry. *Brain & Language*, *221*, Article 104996.

- Lim, S.-J., Qu, A., Tin, J.A.A., & Perrachione, T.K. (2019). Attentional reorientation explains processing costs associated with talker variability. *19th International Congress of Phonetic Sciences* (Melbourne, August 2019).
- Lim, S.-J., Shinn-Cunningham, B.G., & Perrachione, T.K. (2019). Effects of talker continuity and speech rate on auditory working memory. *Attention, Perception, & Psychophysics*, *81*(4), 1167-1177.
- Luthra, S. (2021). The role of the right hemisphere in processing phonetic variability between talkers. *Neurobiology of Language*, *2*(1), 138– 151.
- Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human perception and performance*, *33*(2), 391-409.
- Magnuson, J.S., Nusbaum, H.C., Akahane-Yamada, R., & Saltzman, D. (2021). Talker familiarity and the accommodation of talker variability. *Attention, Perception, & Psychophysics*, *83*, 1842–1860.
- Majerus, S. (2013). Language repetition and short-term memory: an integrative framework. *Frontiers in Human Neuroscience*, *7*, 357.
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. 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.
- Mehrai, G., Shinn-Cunningham, B., & Dau, T. (2018). Influence of talker discontinuity on cortical dynamics of auditory spatial attention. *NeuroImage*, *179*, 548–556.
- 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.

- Myers, E. B., & Theodore, R. M. (2017). Voice-sensitive brain networks encode talker-specific phonetic detail. *Brain and Language*, *165*, 33–44.
- Nearey, T.M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*, 2088-2113.
- Nitsche, M.A., & Paulus, W. (2000). Excitability changes induced in the human motor cortex by weak transcranial direct current stimulation. *Journal of Physiology*, *527*(3), 633–639.
- 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.
- Obleser, J., Zimmermann, J., Van Meter, J., & Rauschecker, J. P. (2007). Multiple stages of auditory speech perception reflected in event-related fMRI. *Cerebral Cortex*, *17*(10), 2251-2257.
- Ohn, S. H., Park, C. I., Yoo, W. K., Ko, M. H., Choi, K. P., Kim, G. M., ... & Kim, Y. H. (2008). Time-dependent effect of transcranial direct current stimulation on the enhancement of working memory. *Neuroreport*, *19*(1), 43-47.
- Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97-113.
- Peelle, J. E. (2012). The hemispheric lateralization of speech processing depends on what “speech” is: A hierarchical perspective. *Frontiers in Human Neuroscience*, *6*, 309.
- Peirce, J.W. (2007). PsychoPy: Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*, 8-13.
- Perrachione, T.K., Del Tufo, S.N., Winter, R., Murtagh, J., Cyr, A., Chang, P., Halverson, K., Ghosh, S.S., Christodoulou, J.A. & Gabrieli, J.D.E. (2016). Dysfunction of rapid neural adaptation in dyslexia. *Neuron*, *92*, 1383-1397.
- Perrachione, T. K., Ghosh, S. S., Ostrovskaya, I., Gabrieli, J. D. E., & Kovelman, I. (2017). Phonological working memory for words and nonwords in cerebral cortex. *Journal of Speech, Language, and Hearing Research*, *60*, 1959–1979.
- Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, *24*, 175–184.

- Pierrehumbert, J. (2002). Word-specific phonetics. In C. Gussenhoven and N. Warner (Eds.), *Laboratory Phonology 7*, pp. 101–139. Berlin: Mouton de Gruyter.
- 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.
- Potter, R. K., & Steinberg, J. C. (1950). Toward the specification of speech. *Journal of the Acoustical Society of America*, 22, 807–820.
- Roe, J. M., Nesheim, M., Mathiesen, N. C., Moberget, T., Alnæs, D., & Sneve, M. H. (2016). The effects of tDCS upon sustained visual attention are dependent on cognitive load. *Neuropsychologia*, 80, 1-8.
- Scott, T. L., & Perrachione, T. K. (2019). Common cortical architectures for phonological working memory identified in individual brains. *NeuroImage*, 202, Article 116096.
- Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in cognitive sciences*, 12(5), 182-186.
- Sjerps, M. J., 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.
- Stevens, A. A. (2004). Dissociating the cortical basis of memory for voices, words and tones. *Cognitive Brain Research*, 18(2), 162-171.
- Stilp, C. E., & Theodore, R. M. (2020). Talker normalization is mediated by structured indexical information. *Attention, Perception, & Psychophysics*, 82, 2237–22431.
- Strange, W., Verbrugge, R. R., Shankweiler, D. P., & Edman, T. R. (1976). Consonant environment specifies vowel identity. *Journal of the Acoustical Society of America*, 60, 213–224.
- 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.
- Sussman, H. M. (1986). A neuronal model of vowel normalization and representation. *Brain & Language*, 28, 12–23.

- Thair, H., Holloway, A. L., Newport, R., & Smith, A. D. (2017). Transcranial direct current stimulation (tDCS): a beginner's guide for design and implementation. *Frontiers in Neuroscience, 11*, 641.
- 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.
- von Kriegstein, K., Eger, E., Kleinschmidt, A., & Giraud, A. L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Cognitive Brain Research, 17*(1), 48–55.
- von Kriegstein, K., Smith, D. R., Patterson, R. D., Kiebel, S. J., & Griffiths, T. D. (2010). How the human brain recognizes speech in the context of changing speakers. *Journal of Neuroscience, 30*(2), 629–638.
- 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.
- Zatorre R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex, 11*(10), 946–953.
- Zhang, C., Peng, G., & Wang, W. S.-Y. (2013). Achieving constancy in spoken word identification: Time course of talker normalization. *Brain and Language, 126*(2), 193–202.
- 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.

