Cognition 204 (2020) 104393

Contents lists available at ScienceDirect

Cognition

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

Brief article

Selecting among competing models of talker adaptation: Attention, cognition, and memory in speech processing efficiency

Alexandra M. Kapadia, Tyler K. Perrachione*

Department of Speech, Language, and Hearing Sciences, Boston University, Boston, MA, United States of America

ARTICLE INFO

Keywords:
Speech perception
Phonetic variability
Processing cost
Talker adaptation
Auditory streaming

ABSTRACT

Phonetic variability across talkers imposes additional processing costs during speech perception, often measured by performance decrements between single- and mixed-talker conditions. However, models differ in their predictions about whether accommodating greater phonetic variability (i.e., more talkers) imposes greater processing costs. We measured speech processing efficiency in a speeded word identification task, in which we manipulated the number of talkers (1, 2, 4, 8, or 16) listeners heard. Word identification was less efficient in every mixed-talker condition compared to the single-talker condition, but the magnitude of this performance decrement was not affected by the number of talkers. Furthermore, in a condition with uniform transition probabilities between two talkers, word identification was more efficient when the talker was the same as the prior trial compared to trials when the talker switched. These results support an auditory streaming model of talker adaptation, where processing costs associated with changing talkers result from attentional reorientation.

1. Introduction

Variation in the acoustic realization of speech across talkers is the principal source of phonetic variability in speech signals (Kleinschmidt, 2019). Listeners are nonetheless highly successful in extracting stable phonemic information from talkers' speech despite the lack of consistent acoustic-phonetic mapping (Pierrehumbert, 2003). However, it has been shown that listening to speech from a variety of talkers incurs additional processing costs to accommodate utterance-to-utterance variation and maintain phonetic constancy (Johnson, 2005; Nusbaum & Magnuson, 1997). These processing costs have been repeatedly demonstrated through *interference effects*—slower response times and reduced accuracy during mixed-talker speech processing tasks (Green et al., 1997; Mullennix & Pisoni, 1990; Nusbaum & Morin, 1992).

Current models of speech processing under uncertainty appeal to different mechanisms to explain the increased costs associated with processing mixed-talker speech, ranging from allocation of cognitive resources (Heald & Nusbaum, 2014; Nusbaum & Magnuson, 1997) to accessing memories (Kleinschmidt & Jaeger, 2015; Pierrehumbert, 2016) to feedforward auditory streaming (Bressler et al., 2014; Choi & Perrachione, 2019a). However, the different mechanisms implicated by these models make conflicting predictions about how speech processing efficiency should be affected by increasing phonetic uncertainty. We aimed to evaluate these predictions by investigating how speech processing costs scale when identifying words spoken by increasing numbers of possible talkers and thereby increasing acoustic-phonemic uncertainty.

Resolving talker variability has been proposed to employ an *active control process* (Heald & Nusbaum, 2014; Nusbaum & Magnuson, 1997; Nusbaum & Schwab, 1986). Here, cognitive resources are allocated by listeners to accommodate processing of new information as uncertainty arises, allowing the perceptual system to remain flexible in situations where acoustic-phonemic correspondences are variable or uncertain. Under this account, listeners' *expectation* of variability imposes a cost on speech processing, as the system pre-allocates some of its limited cognitive resources in anticipation of resolving variability.

Another preeminent model of processing variability in speech—the *ideal adapter framework* (Kleinschmidt & Jaeger, 2015)—posits that processing efficiency depends on costs associated with resolving the number of possible competing interpretations of a speech signal. Under this model, a listener uses available indexical information, such as relevant talker-specific representations, to pare down the number of potential interpretations. Here, memories ("models") of talker characteristics, established through prior experiences with individuals or classes of talkers, are accessed for comparison with incoming acoustic information. Reducing the number of possible internal models of an

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

https://doi.org/10.1016/j.cognition.2020.104393
Received 21 February 2020; Received in revised form 14 June 2020; Accepted 29 June 2020

acoustic signal reduces the decision space for choosing the correct interpretation, thereby making speech processing more efficient.

In contrast with the models above, recent work has begun to consider talker variability in the context of models of *auditory attention and streaming* (Shinn-Cunningham, 2008; Winkler et al., 2009). This framework suggests that the additional processing costs associated with mixed talkers primarily reflect the costs of attentional switching when reorienting from one auditory source (talker) to another. That is, speech processing is most efficient when listeners' attention to a single, continuous talker is undisrupted, whereas the auditory stream discontinuity associated with a talker change incurs additional processing costs as listeners switch their attention from one source to another (Bressler et al., 2014; Choi & Perrachione, 2019a, 2019b; Lim, Shinn-Cunningham, & Perrachione, 2019; Mehraei et al., 2018).

Although these models all account for the presence of increased processing costs, they make different predictions about how processing costs should scale with increasing numbers of talkers. The active control hypothesis predicts that there will be a decrement in speech processing efficiency whenever there is the potential for talker variability, because cognitive resources are pre-allocated based on either the expectation or experience of variability (Magnuson & Nusbaum, 2007). Under the ideal adapter framework, processing speech from a limited number of potential talkers (e.g., 2 or 4) should be more efficient than a larger number of talkers (e.g., 8 or 16) because more constraints on model selection reduce the number of possible interpretations of the acoustic signal. An auditory streaming interpretation, however, predicts that listening to any number of mixed talkers will be equally inefficient compared to a single talker, as the processing costs of mixed-talker speech are specifically associated with auditory stream disruption arising from talker discontinuity. Thus, decreases in speech processing efficiency should not depend on the number of different talkers, but on the occurrence of talker switches.

Moreover, these models make different predictions about the effect of listeners' expectations about the upcoming talker. The active control hypothesis predicts that processing costs will be incurred any time there is potential uncertainty (Magnuson & Nusbaum, 2007), whereas the ideal adapter framework predicts that specific expectations about the identity of an upcoming talker will facilitate speech processing by constraining the model selection process (Kleinschmidt & Jaeger, 2015). Auditory streaming predicts that continuity in an auditory source, whether expected or unexpected, will facilitate speech processing, whereas any discontinuity, such as a talker change, will incur processing costs, even if expected (Carter et al., 2019; Mehraei et al., 2018).

Existing empirical observations about talker variability-induced processing costs are insufficient to evaluate the predictions of the various models or favor one potential mechanism over another. Most prior studies have not parametrized the number of talkers, instead opting to operationalize talker variability via mixed-talker conditions with fixed numbers of talkers. However, numerous methodological differences preclude the ability to compare the effect sizes of talker variability across such studies, including differences regarding stimuli, task, response type, dependent variable, and, especially, whether talker variability was tested as a within- vs. between-subjects variable (Bradlow et al., 1999; Choi et al., 2018; Green et al., 1997; Mullennix & Howe, 1999; Mullennix & Pisoni, 1990; Morton et al., 2015; Perrachione et al., 2016; Sommers et al., 1997; Sommers et al., 1994; Wong et al., 2004, Zhang & Chen, 2016; inter alia).

Due to models' diverging predictions and the lack of empirical work addressing this question, we investigated how processing costs vary with respect to the amount of talker variability, operationalized as number of talkers. Specifically, we tested whether speech processing efficiency decreases monotonically as a function of the number of talkers, as predicted by the ideal adapter framework, or whether a fixed loss of efficiency occurs in the presence of any variability, as predicted by the active control hypothesis and auditory streaming framework. We also tested the models' diverging predictions about whether speech processing efficiency is affected by expectations of talker continuity vs. change, particularly whether an *expected talker change* can make speech processing more efficient (as predicted by the ideal adapter framework, but not the active control hypothesis or auditory streaming framework), or whether *unexpected talker continuity* can make speech processing more efficient (as predicted by the auditory streaming framework, but not the active control hypothesis or ideal adaptor framework).

2. Methods

2.1. Participants

Native speakers of American English (N = 72; 55 female, 17 male; mean age = 20.3 ± 2.5, 18–30 years) participated in this study. All participants reported a history free from speech, language, or hearing disorders. Participants provided informed written consent, approved and overseen by the Institutional Review Board at Boston University.

2.2. Stimuli

The naturally spoken English words “boot” and “boat” were recorded by 16 native speakers of American English (eight female, eight male). These words were chosen because their minimally contrastive

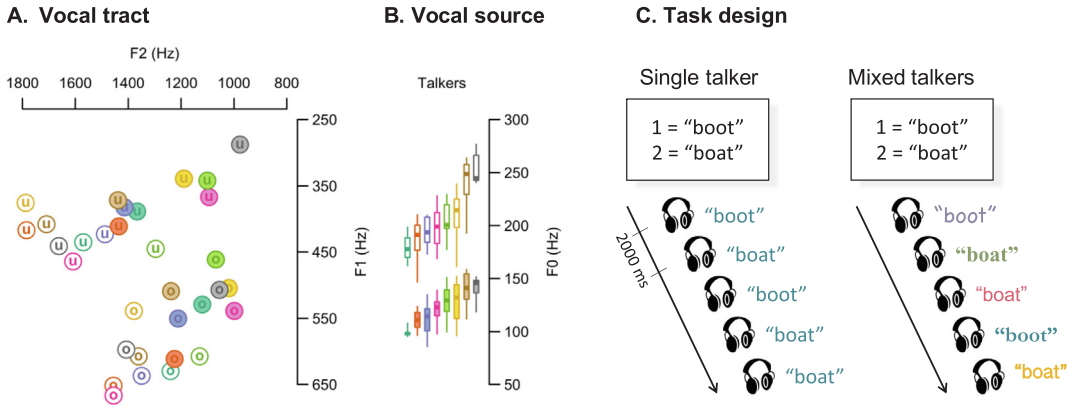

Fig. 1. A. Phonetic variation across all 16 talkers for the target words (“boot” and “boat”). Circles indicate the location of each talker's vowel (/u/ or /o/) in F1-F2 space. B. Box plots show the fundamental frequency (f_0) range for these stimuli across talkers. Filled and empty points/boxes for male and female talkers, respectively, with unique colors for each talker. C. Schematic representation of stimulus delivery. Participants performed a speeded word identification task while listening to speech produced by a single talker or mixed talkers. Mixed talker conditions included 2, 4, 8, or 16 talkers (depicted by different fonts and colors). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Measures of speech processing efficiency by number of talkers (mean ± s.d. across participants).

Number of talkers	Efficiency	Accuracy	Response time
1	1.49 ± 0.27	96.8 ± 4.5%	673 ± 126 ms
2	1.28 ± 0.26	95.5 ± 5.2%	772 ± 140 ms
4	1.29 ± 0.27	96.4 ± 4.1%	781 ± 163 ms
8	1.26 ± 0.25	96.1 ± 4.0%	791 ± 151 ms
16	1.29 ± 0.27	95.6 ± 7.0%	768 ± 144 ms

vowels (/u/ vs. /o/) have the greatest potential acoustic-phonemic ambiguity across talkers (Choi et al., 2018; Hillenbrand et al., 1995). Phonetic variability across all 16 talkers for both words is illustrated in Fig. 1A–B.

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. Recordings were normalized in Praat (Boersma, 2001) such that the RMS presentation level of each stimulus was 65 dB SPL. Stimulus durations ranged from 220 to 788 ms (mean ± s.d.: 390 ± 130 ms).

2.3. Procedure

Participants performed a speeded word-identification task in which they identified the target word ("boot" or "boat") as quickly and accurately as possible by pressing a corresponding number on a keypad. Participants received verbal instructions at the beginning of the experiment. Written directions assigning a number to each target word were displayed on the screen throughout the experiment (Fig. 1C). Words were presented with a 2000 ms stimulus onset asynchrony, and stimulus delivery was controlled using PsychoPy2 (v1.85.2) (Peirce, 2007) via Sennheiser HD-380 Pro headphones.

The task was divided into separate blocks, with the number of talkers (1, 2, 4, 8, or 16) varying parametrically across blocks. Written directions at the beginning of each block informed participants of the number of talkers in that block (Magnuson & Nusbaum, 2007). Each block was 64 trials long, with each target word presented 32 times per block. Stimuli were presented in a pseudorandom order such that the same word did not repeat more than three times in a row and the same talker did not repeat on adjacent trials during the 4-, 8-, or 16-talker conditions.

Natural limitations on how stimuli can be ordered in a 2-talker condition provided an additional opportunity to explore how listeners' expectations of talker repetition affect speech processing efficiency. If, like the other mixed-talker conditions, the same talker did not repeat on adjacent trials, then the two talkers would alternate predictably between trials (ABABABAB …); thus, listeners could anticipate the identity of the upcoming talker with perfect certainty, eliminating the potential ambiguity in interpreting speech associated with a talker switch. Alternatively, by presenting two talkers in a pseudorandom order (AABABBAB …), the same talker would occur on adjacent trials, potentially affecting processing efficiency via feedforward benefits of talker continuity. To evaluate the various models' predictions of speech processing efficiency under these different trial structures, each participant completed both variations of the 2-talker condition within the overall experiment. These two variations were always presented one after another. In the *alternating 2-talker condition*, the two talkers appeared consistently on every other trial. In the *random 2-talker condition*, the transition probability of a talker repeat or a talker switch was equal: there were 32 trials with the same talker as the previous trial and 32 trials with a different talker from the previous trial. All participants completed all six conditions.

Talkers were randomly selected for each participant, such that there were an equal number of female and male voices in each condition (the single talker block was split into a female talker half and a male talker half). The order of talker-number conditions was randomized across participants.

2.4. Data analysis

Accuracy and response time data were collected on each trial. Response time (RT) was measured as the time delay between the onset of the stimulus and the participant's keypad response. RT was log transformed to more closely approximate a normal distribution. Trials with incorrect responses or log-transformed RTs exceeding three standard deviations from the participant's mean for that condition were excluded from RT analysis (4.0% of all trials). *Efficiency* was calculated as the quotient of mean accuracy and mean RT per participant per condition (Lim, Shinn-Cunningham, & Perrachione, 2019; Townsend & Ashby, 1978). Accuracy and response time data from every included trial were analyzed in (generalized) linear mixed-effects models in R using the packages *lme4* and *lmerTest*. Because efficiency was calculated as a summary statistic over all trials, resulting in one value per participant per condition, these data were analyzed in repeated-measures analyses of variance (ANOVA) using the package *ez*.

3. Results

3.1. Effects of talker variability

3.1.1. Efficiency

We assessed how the number of talkers affected processing efficiency, and differentially affected both RT and accuracy (Table 1). We conducted a repeated-measures ANOVA on the efficiency measure using a within-subjects factor of *number of talkers* (1, 2, 4, 8, and 16). Significant effects were evaluated by *post-hoc* paired *t*-tests.

There was a significant effect of number of talkers ($F(4, 284) = 40.30$; $p \ll 0.0001$; $\eta^2_G = 0.092$; Fig. 2A). Post-hoc tests revealed that efficiency was significantly reduced in the 2-talker condition compared to a single talker ($t(71) = 11.28$; $p \ll 0.0001$), but that there was no further reduction in efficiency for increasing numbers of talkers (4-talker vs. 2-talker: $t(71) = -0.26$; $p = 0.80$; 8-talker vs. 4-talker: $t(71) = 1.35$; $p = 0.18$; 16-talker vs. 8-talker: $t(71) = -1.57$; $p = 0.12$).

3.1.2. Accuracy

We analyzed accuracy in a generalized linear mixed-effects model with fixed-effects terms including the categorical factor *number of talkers* (1, 2, 4, 8, or 16) and a continuous covariate for *condition order*; random-effects terms included by-participant intercepts and by-participant slopes for all fixed factors, and by-stimulus intercepts. The significance of fixed effects was tested in a type-III ANOVA using Wald chi-square tests followed by contrasts on the factor levels that coded for successive differences (i.e., 1-vs.-2, 2-vs.-4, etc.). We adopted the significance criterion $\alpha = 0.05$, with *p*-values based on the Satterthwaite approximation for degrees of freedom.

There was a significant effect of number of talkers ($\chi^2(4) = 12.96$, $p = 0.011$; Fig. 2B). Successive-differences contrasts on the levels of the *number of talkers* factor in the model revealed that accuracy fell significantly when talker variability was introduced (2-talker vs. 1-talker: $\beta = -0.46$, *s.e.* $= 0.13$, $z = -3.46$, $p = 0.0005$) but did not change further with increasing numbers of talkers (4-talker vs. 2-talker: $\beta = 0.12$, *s.e.* $= 0.12$, $z = 0.95$, $p = 0.34$; 8-talker vs. 4-talker: $\beta = -0.074$, *s.e.* $= 0.12$, $z = -0.61$, $p = 0.54$; 16-talker vs. 8-talker: $\beta = 0.16$, *s.e.* $= 0.14$, $z = 1.13$, $p = 0.26$).

3.1.3. Response time

We analyzed RT in a linear mixed-effects model with the same structure as that for accuracy. There was again a significant effect of number of talkers ($\chi^2(4) = 203.89$, $p \ll 0.0001$; Fig. 2C). Successive-differences contrasts on the levels of the *number of talkers* factor in the

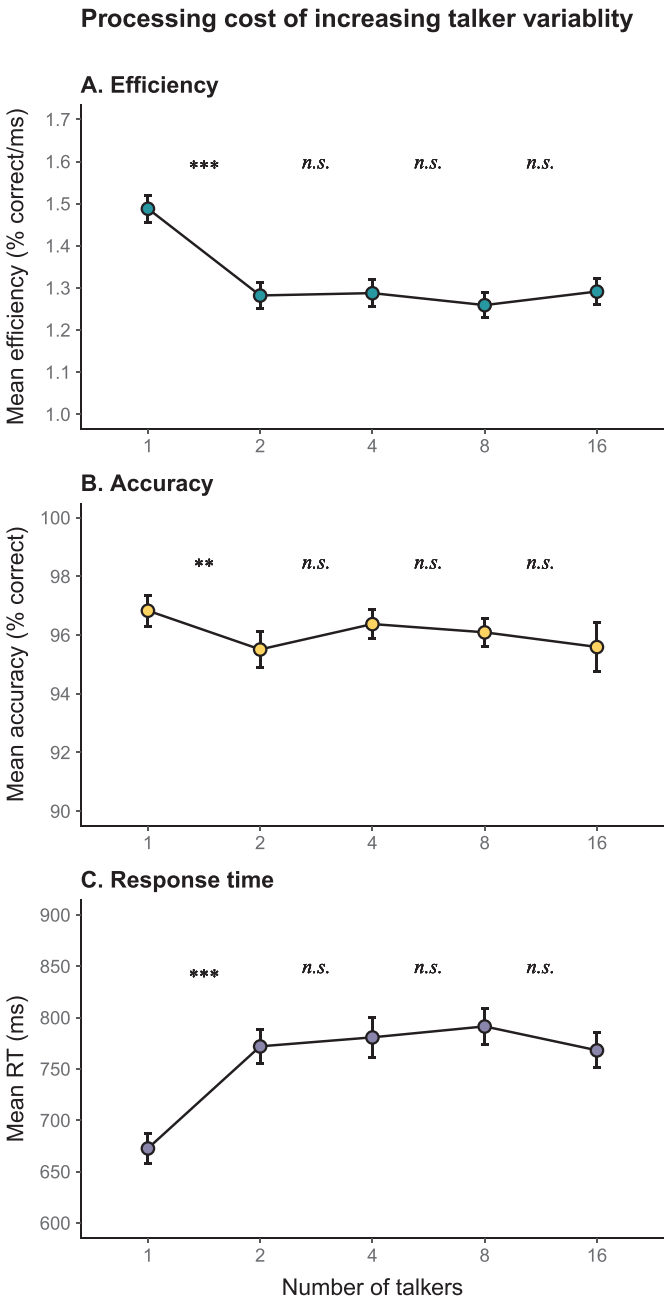

Fig. 2. Processing cost as a function of number of talkers. A. Mean *efficiency*, calculated as mean accuracy divided by mean response time for each participant for each condition. B. Mean accuracy, as percent of trials. C. Mean RT to correct trials. Greater processing costs are indicated by decreases in efficiency and accuracy and increases in RT. Significance of pairwise contrasts is indicated above each line. Error bars indicate ± 1 SEM across participants.

model revealed that RT was significantly slower in the 2-talker condition than in the 1-talker condition ($\beta = 0.057$, *s.e.* $= 0.005$, $t = 11.60$, $p \ll 0.0001$), with no significant increase in RT for additional talkers beyond two (4-talker vs. 2-talker: $\beta = 0.004$, *s.e.* $= 0.005$, $t = 0.83$, $p = 0.41$; 8-talker vs. 4-talker: $\beta = 0.005$, *s.e.* $= 0.006$, $t = 0.95$, $p = 0.35$; 16-talker vs. 8-talker: $\beta = -0.009$, *s.e.* $= 0.005$, $t = -1.86$, $p = 0.07$).

3.2. Effects of talker continuity

Next, we unpacked listeners' performance on the two different designs of the 2-talker condition to understand how talker (dis)continuity and listeners' expectations about the upcoming talker affected speech processing (efficiency, accuracy, and RT). Specifically, we compared those values obtained from the following trial types (examples underlined): (i) predictable talker-repeat trials from the *1-talker condition* (AAAAAA), (ii) unpredictable talker-repeat trials from the *random 2-talker condition* (ABAABB), (iii) unpredictable talker-change trials from the *random 2-talker condition* (ABAABB), and (iv) predictable talker-change trials from the *alternating 2-talker condition* (ABABAB).

Statistical analyses and model structures were the same as above. Here, our within-subjects categorical fixed factor was *trial structure* (1-talker, random 2-talker repeats, random 2-talker changes, and alternating 2-talker changes; Table 2).

Table 2
Measures of speech processing efficiency by talker (dis)continuity (mean ± s.d. across participants).

Trial type	Efficiency	Accuracy	Response time
1-talker (AAAA)	1.49 ± 0.27	96.8 ± 4.5%	673 ± 126 ms
Random 2-talker repeats (ABBA)	1.37 ± 0.30	96.5 ± 5.8%	735 ± 146 ms
Random 2-talker changes (ABBA)	1.24 ± 0.27	95.0 ± 6.5%	795 ± 149 ms
Alternating 2-talker (ABAB)	1.27 ± 0.27	95.2 ± 5.9%	780 ± 144 ms

3.2.1. Efficiency

There was a significant effect of *trial structure* ($F(3,213) = 65.33$; $p \ll 0.0001$ $\eta^2_G = 0.11$; Fig. 3A). Efficiency was significantly greater for anticipated talker continuity compared to unanticipated talker continuity (1-talker vs. random 2-talker repeats; $t(71) = 5.37$; $p \ll 0.0001$). Efficiency was also significantly greater on trials when the talker repeated than when it changed, notwithstanding listeners' inability to have anticipated any such repetition (random 2-talker repeats vs. random 2-talker changes; $t(71) = 9.37$; $p \ll 0.0001$). However, even when listeners could perfectly predict the upcoming talker, their word recognition efficiency did not improve compared to trials where the talker change could not be anticipated (alternating 2-talker vs. random 2-talker changes; $t(71) = 1.33$; $p = 0.19$).

3.2.2. Accuracy

There was no difference in accuracy when talker repetition could be reliably anticipated versus when the talker repeated unexpectedly (1-talker vs. random 2-talker repeats; $\beta = 0.14$, *s.e.* $= 0.16$, $z = 0.90$, $p = 0.37$; Fig. 3B). However, listeners were significantly more accurate when the talker repeated than when the talker changed, notwithstanding their inability to anticipate the upcoming talker (random 2-talker repeats vs. random 2-talker changes; $\beta = -0.64$, *s.e.* $= 0.16$, $z = -3.99$, $p \ll 0.0001$). Accuracy did not differ on trials where the talker change was unpredictable vs. predictable (random 2-talker changes vs. alternating 2-talker; $\beta = -0.05$, *s.e.* $= 0.16$, $z = -0.31$, $p = 0.75$).

3.2.3. Response time

RT was faster in the 1-talker condition than when the talker unpredictably repeated in the random 2-talker condition ($\beta = 0.037$, *s.e.* $= 0.006$, $t = 6.33$, $p \ll 0.0001$; Fig. 3C). RT was also significantly faster for trials when the talker repeated than when the talker changed, notwithstanding listeners' inability to anticipate the upcoming talker (random 2-talker repeats vs. random 2-talker changes; $\beta = 0.034$, *s.e.* $= 0.004$, $t = 9.65$, $p \ll 0.0001$). However, even when listeners could perfectly anticipate who the next talker would be, their RT did not differ compared to when the talker change was unpredictable (random 2-talker changes vs. alternating 2-talker; $\beta = -0.007$, *s.e.* $= 0.005$, $t = -1.46$, $p = 0.15$).

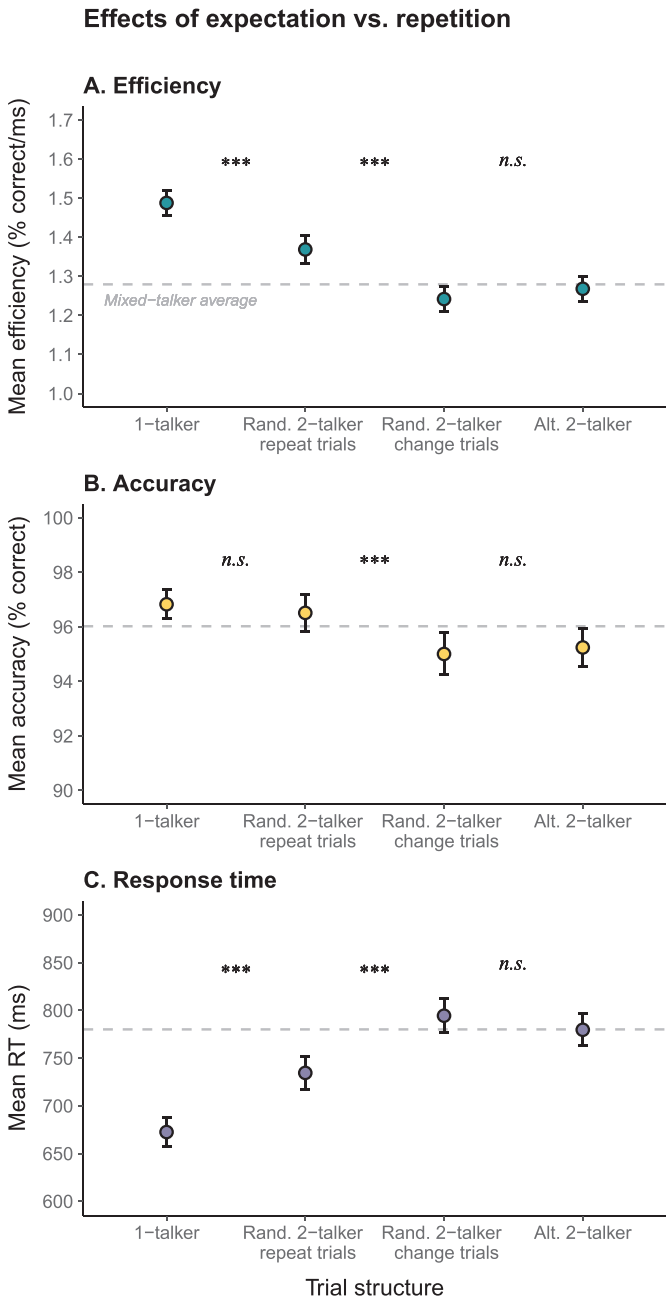

Fig. 3. Processing cost as a function of talker (dis)continuity. A. Mean *efficiency*, calculated as mean accuracy divided by mean response time for each participant for each condition. B. Mean accuracy, as percent of trials. C. Mean RT to correct trials. Greater processing costs are indicated by decreases in efficiency and accuracy and increases in RT. Significance of pairwise contrasts is indicated above each line. The horizontal dashed line in each panel represents the mixed-talker average (mean of the 4-, 8-, and 16- talker conditions). Error bars indicate ± 1 SEM across participants.

4. Discussion

In this study, we found that the costs associated with processing talker variability do not increase as a function of the amount of potential variability faced by the speech perception system. Listeners incur processing costs in both accuracy and response time for word identification when the number of possible talkers increases from one to two, but do not incur any additional processing costs as the number of talkers increases from two to 16. Furthermore, these processing costs appear to be primarily a result of feedforward disruption of talker continuity, rather than a result of selecting top-down interpretations to guide acoustic-phonemic mappings.

Neither of these observations appears consistent with a speech processing framework positing that acoustic-phonemic mappings become more efficient as the decision space of possible interpretations of an acoustic signal is reduced in a top-down, expectation-driven way (Kleinschmidt & Jaeger, 2015). Reducing the number of potential interpretations of an incoming speech signal does not offer any improvement in efficiency until that number is reduced to one. Furthermore, even when listeners could perfectly anticipate which new talker would speak on the next trial, word recognition was not more efficient than when the next talker was unpredictable. These two results suggest that, while the ideal adapter framework offers considerable explanatory power for the decision outcomes of speech perception when faced with signal uncertainty, the predictions of this model ultimately do not appear to account for the large literature on differences in processing efficiency for single- vs. mixed-talker speech (Choi & Perrachione, 2019a; Johnson, 1990; Green et al., 1997; Magnuson & Nusbaum, 2007; Mullennix & Pisoni, 1990; inter alia).

Instead, these observations appear consistent with an attentional model of speech processing, where processing costs are incurred when the auditory stream is disrupted, such as in talker discontinuity and, inversely, where efficiency gains are realized when there is continuity in the auditory source of speech (Bressler et al., 2014; Choi & Perrachione, 2019a; Lim, Shinn-Cunningham, & Perrachione, 2019; Shinn-Cunningham, 2008). Listeners incur a processing cost whenever auditory stream coherence is disrupted, and the magnitude of this cost does not vary with the number of potential new interpretations of the signal after disruption. Furthermore, word recognition was more efficient (both faster and more accurate) when the same talker spoke on two consecutive trials, even when such continuity was not predictable. This result suggests an automatic, feedforward facilitatory effect of talker continuity on speech processing efficiency (Choi & Perrachione, 2019a; Mullennix & Howe, 1999) such that continuity in an auditory source drives attentional capture and offers a processing advantage over attentional reorientation (Bressler et al., 2014; Mehraei et al., 2018; Uddin et al., 2020). Indeed, many effects previously described as "extrinsic" talker normalization may reflect processing gains associated with auditory streaming (Choi & Perrachione, 2019a; Sjerps et al., 2011a, 2011b).

Here it is important to acknowledge that, while the stimuli in the present experiment had substantial acoustic-phonemic variability across talkers, our participants were nonetheless likely to find these phonetic-phonemic correspondences to be highly familiar (as all talkers and listeners were native speakers of American English, with little or no salient regional or social accent variation present in the stimuli). It is interesting to consider whether additional variability in phonetic-phonemic mapping, such as that from different regional, social, or foreign-language accents, would incur additional processing costs, and whether these, too, would be independent of the number of possible interpretations of the acoustic signal. Indeed, knowing how the processes that support talker adaptation as listeners encounter each new talker (e.g., Magnuson & Nusbaum, 2007; Choi & Perrachione, 2019a, 2019b) are related to the processes that support talker adaptation when the phonetic-phonemic mappings are unfamiliar to listeners (e.g., Norris et al., 2003; Xie & Myers, 2017) remains a "missing link" in the literature on processing and representing variability in speech (Bent & Holt, 2017).

If efficiency gains attributed to talker adaptation were strictly about stream coherence afforded by talker continuity, we would have expected all talker continuity to be equally beneficial. However, the results of the different 2-talker manipulations revealed that talker continuity cannot be the only mechanism at play. Accuracy for an unexpected talker repetition was significantly higher than an expected talker change, and equally good as for a single continuous talker (see also Morton et al., 2015). However, while RT was faster when there was

unexpected talker continuity compared to both expected and unexpected talker changes (Carter et al., 2019), it was still not as fast as when listening to a single talker. While it may be the case that listeners need more than a single trial to fully tune in to a new auditory stream (cf. Lim, Qu, et al., 2019), this result is also consistent with the predictions of an active control mechanism that prioritizes speech perception accuracy by allocating cognitive resources to processing uncertainty whenever variability is expected (Heald & Nusbaum, 2014; Magnuson & Nusbaum, 2007; Nusbaum & Magnuson, 1997; Nusbaum & Schwab, 1986). Under this framework, the ongoing expectation of uncertainty requires listeners to pre-allocate cognitive resources to processing variability, meaning that these resources were limited even for talker-repeat trials, slowing processing compared to the single-talker condition, but maintaining high accuracy.

An integrated view of talker adaptation—incorporating feed-forward attentional facilitation by source continuity with top-down allocation of cognitive resources to resolve anticipated variability—parsimoniously accounts for all the results in the present study without appealing to memory mechanisms that depend on top-down predictions about the identity of a talker. These results suggest that speech processing efficiency gains associated with talker adaptation may be best understood as effects of feedforward attentional capture via auditory streaming.

CRediT authorship contribution statement

Alexandra M. Kapadia: Methodology, Software, Investigation, Data curation, Visualization, Writing - original draft.**Tyler K. Perrachione:**Conceptualization, Methodology, Validation, Writing - review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgments

We thank Yaminah Carter, Jessica Tin, Ja Young Choi, Sung-Joo Lim, Jayden Lee, Kamilah Harruna, Nicole Chen, Chinazo Otiono, Trista Lin, Grace Mecha, Maya Saupe, Amabel Antwi, and Michelle Njoroge. Portions of these results were presented at the 19th International Congress of Phonetic Sciences. This work was supported by the National Institute on Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health under awards R03 DC014045 (to TKP), R01 DC004545 (to Gerald Kidd), and T32 DC013017.

Appendix A. Supplementary data

Supplementary data to this article, including all of the stimuli, stimulus delivery code, participant data, and analysis code, can be found online at https://doi.org/10.1016/j.cognition.2020.104393.

References

Bent, T., & Holt, R. F. (2017). Representation of speech variability. *WIREs Cognitive Science*Article e01434. https://doi.org/10.1002/wcs.1434.

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

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perceptual Psychophysics, 61*, 206–219. https://doi.org/10.3758/BF03206883.

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. https://doi.org/10.1007/s00426-014-0555-7.

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, August 2019)*.

Choi, J. Y., Hu, E. R., & Perrachione, T. K. (2018). Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing. *Attention, Perception, & Psychophysics, 80*, 784–797. https://doi.org/10.3758/s13414-017-1395-5.

Choi, J. Y., & Perrachione, T. K. (2019a). Time and information in perceptual adaptation to speech. *Cognition, 192*, 103982. https://doi.org/10.1016/j.cognition.2019.05.019.

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

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. https://doi.org/10.3758/BF03206015.

Heald, S. L. M., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience, 8*, 35. https://doi.org/10.3389/fnsys.2014.00035.

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. https://doi.org/10.1121/1.411872.

Johnson, K. (1990). The role of perceived speaker identity in F0 normalization of vowels. *Journal of the Acoustical Society of America, 88*(2), 642–654. https://doi.org/10.1121/1.399767.

Johnson, K. (2005). Speaker normalization in speech perception. In D. B. Pisoni, & R. E. Remez (Eds.). *The handbook of speech perception* (pp. 363–389). Malden: Blackwell.

Kleinschmidt, D. F. (2019). Structure in talker variability: How much is there and how much can it help? *Language, Cognition and Neuroscience, 34*(1), 43–68. https://doi.org/10.1080/23273798.2018.1500698.

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*, 148–203. https://doi.org/10.1037/a0038695.

Lim, S.-J., Qu, A., Tin, J. A. A., & Perrachione, T. K. (2019). Attentional reorientation explains processing costs associated with talker variability. *Proceedings of the 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*, 1167–1177. https://doi.org/10.3758/s13414-019-01684-w.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology: Human Perception and Performance, 33*, 391–409. https://doi.org/10.1037/0096-1523.33.2.391.

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

Morton, J. R., Sommers, M. S., & Lulich, S. M. (2015). The effect of exposure to a single vowel on talker normalization for vowels. *Journal of the Acoustical Society of America, 137*, 1443–1451. https://doi.org/10.1121/1.4913456.

Mullennix, J. W., & Howe, J. N. (1999). Selective attention in perceptual adjustments to voice. *Perceptual and Motor Skills, 89*(2), 447–457 (doi:10.2466%2Fpms.1999.89.2.447).

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics, 47*, 379–390. https://doi.org/10.3758/BF03210878.

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*, 204–238. https://doi.org/10.1016/S0010-0285(03)00006-9.

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

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).

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

Peirce, J. W. (2007). PsychoPy: Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017.

Perrachione, T. K., Del Tufo, S. N., Winter, R., Murtagh, J., Cyr, A., Chang, P., ... Gabrieli, J. D. E. (2016). Dysfunction of rapid neural adaptation in dyslexia. *Neuron, 92*, 1383–1397. https://doi.org/10.1016/j.neuron.2016.11.020.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech, 46*(2–3), 115–154. https://doi.org/10.1177/00238309030460020501.

Pierrehumbert, J. B. (2016). Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics, 2*(1), 33–52. https://doi.org/10.1146/annurev-linguistics-030514-125050.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences, 12*, 182–186. https://doi.org/10.1016/j.tics.2008.02.003.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011a). Constraints on the processes responsible for the extrinsic normalization of vowels. *Attention, Perception, & Psychophysics, 73*(4), 1195–1215. https://doi.org/10.3758/s13414-011-0096-8.

Sjerps, M. J., Mitterer, H., & McQueen, J. M. (2011b). Listening to different speakers: On the time-course of perceptual compensation for vocal-tract characteristics. *Neuropsychologia, 49*(14), 3831–3846. https://doi.org/10.1016/j.neuropsychologia.2011.09.044.

Sommers, M. S., Kirk, K. I., & Pisoni, D. B. (1997). Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format. *Ear and Hearing, 18*, 89. https://doi.org/10.1097/00003446-199704000-00001.

Sommers, M. S., Nygaard, L. C., & Pisoni, D. B. (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *The Journal of the Acoustical Society of America, 96*(3), 1314–1324. https://doi.org/10.1121/1.411453.

Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan, & F. Restle (Vol. Eds.), *Cognitive theory. Vol. 3.*

Cognitive theory (pp. 200–239). Hillsdale, NJ: Erlbaum.
Uddin, S., Reis, K. S., Heald, S. L., Van Hedger, S. C., & Nusbaum, H. C. (2020). Cortical mechanisms of talker normalization in fluent sentences. *Brain and Language, 201*, 104722. https://doi.org/10.1016/j.bandl.2019.104722.
Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: Predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences, 13*, 532–540. https://doi.org/10.1016/j.tics.2009.09.003.
Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience, 16*, 1173–1184. https://doi.org/10.1162/0898929041920522.
Xie, X., & Myers, E. B. (2017). Learning a talker or learning an accent: Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language, 97*, 30–46. https://doi.org/10.1016/j.jml.2017.07.005.
Zhang, C., & Chen, S. (2016). Towards an integrative model of talker normalization. *Journal of Experimental Psychology–Human Perception and Performance, 42*, 1252–1268. https://doi.org/10.1037/xhp0000216.