AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology:
Human Perception and Performance

ISSN: 0096-1523

2021, Vol. 47, No. 12, 1673–1680
https://doi.org/10.1037/xhp0000963

Attention, Task Demands, and Multitalker Processing Costs in Speech Perception

David Saltzman, Sahil Luthra, Emily B. Myers, and James S. Magnuson
Department of Psychological Sciences, University of Connecticut

Determining how human listeners achieve phonetic constancy despite a variable mapping between the acoustics of speech and phonemic categories is the longest standing challenge in speech perception. A clue comes from studies where the talker changes randomly between stimuli, which slows processing compared with a single-talker baseline. These multitalker processing costs have been observed most often in speeded monitoring paradigms, where participants respond whenever a specific item occurs. Notably, the conventional paradigm imposes attentional demands via two forms of varied mapping in mixed-talker conditions. First, *target recycling* (i.e., allowing items to serve as targets on some trials but as distractors on others) potentially prevents the development of task automaticity. Second, in mixed trials, participants must respond to two unique stimuli (i.e., one target produced by each talker), whereas in blocked conditions, they need respond to only one token (i.e., multiple target tokens). We seek to understand how attentional demands influence talker normalization, as measured by multitalker processing costs. Across four experiments, multitalker processing costs persisted when target recycling was not allowed but diminished when only one stimulus served as the target on mixed trials. We discuss the logic of using varied mapping to elicit attentional effects and implications for theories of speech perception.

Public Significance Statement
This study highlights the importance of attention to the process of accommodating the unique way each individual speaks, which may not occur automatically unless the talker is relevant to the current situation. Understanding how listeners are able to adapt to each unique talker may later help in devising interventions for those with speech comprehension issues.

Keywords: talker normalization, word monitoring, automaticity, phonetic constancy

The mapping from the acoustic details of the speech signal to phonemes can vary tremendously depending on factors such as phonetic context, speaking rate, or ambient acoustic context; how listeners routinely perceive a talker's intended utterance despite this lack of invariance between the acoustic signal and perceptual categories is one of the oldest problems in speech perception (Liberman et al., 1957), and it remains unsolved today. Critically, the lack of invariance problem is exacerbated by the fact that individual talkers may produce their speech sounds in substantially different ways (with acoustic consequences), both for vowels (Peterson & Barney, 1952) and consonants (Dorman et al., 1977). Nonetheless, listeners typically perceive the content of the speech signal with ease, achieving phonetic constancy despite talker variability.

Researchers have proposed that in order to accommodate talker variability, listeners must adjust the mapping between acoustic details and phonetic categories on the basis of talker information (e.g., Joos, 1948; Ladefoged & Broadbent, 1957; Nearey, 1989; Nusbaum & Magnuson, 1997). In a classic monograph, Joos (1948) suggested a talker accommodation process by which listeners might make the necessary mapping adjustments. Joos proposed that listeners might use an initial sample of a talker's speech (e.g., a conventional greeting, such as "How do you do?") to map the talker's speech onto phonological (perceptual) categories, and then "shift" or "distort" either the incoming speech or their internal representations to bring the two into registration. This perspective is consistent with a large body of literature suggesting that listeners' interpretation of speech is modulated by acoustic information encountered in preceding auditory contexts (Bosker, 2018;

David Saltzman https://orcid.org/0000-0002-7244-5274
Sahil Luthra https://orcid.org/0000-0002-3517-2609
Emily B. Myers https://orcid.org/0000-0002-9475-764X
James S. Magnuson https://orcid.org/0000-0003-0158-2367

This research was supported by National Institutes of Health Grant R01 H14-001 (Principal Investigator [PI] Emily B. Myers) and National Science Foundation (NSF) Grants NRT 1747486 and PAC 1754284 (PI James S. Magnuson). Sahil Luthra was supported by an NSF graduate research fellowship. This work was presented at the 61st Annual Meeting of the Psychonomic Society, November 19-22 2020, Online. The preregistration plan for this study is available on the Open Science Framework (https://osf.io/wx4kd), and all stimuli and analysis code are available on GitHub (https://github.com/disaltzman/TalkerTeam-Mapping).

Correspondence concerning this article should be addressed to David Saltzman, Department of Psychological Sciences, University of Connecticut, 406 Babbidge Road, Unit 1020, Storrs, CT 06269, United States. Email: david.saltzman@uconn.edu

Ladefoged & Broadbent, 1957; Laing et al., 2012; Sjerps et al., 2019; Stilp, 2019; Zhang et al., 2013).[1]

A number of speech perception studies show that listeners are slower and/or less accurate in identifying words when the talker varies from word to word compared with when all the words are spoken by a single talker (Carter et al., 2019; Choi et al., 2018; Choi & Perrachione, 2019a, 2019b; Heald & Nusbaum, 2014; Kapadia & Perrachione, 2020; Magnuson & Nusbaum, 2007; Mullennix et al., 1989; Nusbaum & Morin, 1992; Verbrugge et al., 1976; Wong et al., 2004). Some have interpreted these multitalker processing costs as being a consequence of talker normalization or talker accommodation.[2] On such a view, each time a new talker is encountered, listeners must reengage the normalization/accommodation mechanism, and a processing cost is incurred as a result.

Much of our understanding of the processing costs associated with talker variability comes from studies that have used a speeded monitoring task (e.g., Antoniou et al., 2015; Heald & Nusbaum, 2014; Magnuson & Nusbaum, 2007; Magnuson et al., 2021; Nusbaum & Morin, 1992; Wong et al., 2004). In this paradigm, listeners hear a series of stimuli (e.g., *jolt*, *depth*, *ball*, *romp*) and must press a button whenever they hear a target item, indicated visually (e.g., BALL). In blocked-talker trials, one talker produces both the target and distractor items, whereas in mixed-talker trials, two different talkers produce both the target and distractor items and the talker alternates pseudo randomly from item to item (the total number of items in a mixed-talker trial is identical to a blocked-talker trial, as each talker produces half of the items). As expected by normalization/accommodation accounts, listeners are slower to identify the target word in mixed-talker trials compared with blocked-talker trials.

Nusbaum and his colleagues (Francis & Nusbaum, 1996; Heald & Nusbaum, 2014; Magnuson & Nusbaum, 2007; Nusbaum & Magnuson, 1997; Nusbaum & Morin, 1992; Nusbaum & Schwab, 1986) have proposed that achieving phonetic constancy despite the apparent lack of invariance between acoustics and percepts requires active, attention- and resource-demanding processes. Thus, when Nusbaum and Morin (1992, p. 122) described the features of the speeded monitoring task that they applied to the challenge of talker normalization, they pointed out that, by design, the blocked- and mixed-talker conditions differ in that blocked-talker conditions are amenable to automaticity (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977) but mixed-talker conditions are not. They noted that in a blocked-talker trial, participants must make a response to a single target item, while in mixed trials, the target items are produced by two talkers, and therefore participants must respond to two distinct stimuli (one produced by each talker): They noted that "from a cognitive perspective, recognition in the mixed-talker condition should require more effort and attention than recognition in the blocked-talker condition." (Nusbaum and Morin, 1992, p. 122).

On this logic, the mixed-talker condition is designed to reveal increased attentional demands induced by talker normalization/accommodation. If speech perception is normally a highly automatized, efficient process, detecting subtle differences in attentional demands induced by a talker change may require stressing the system. Crucially, Nusbaum and Morin (1992) proposed that the computations required to adjust acoustic-perceptual mappings after a talker change would require attention. If this were the case, a simple attentional manipulation like digit load should produce an interaction with talker condition, exacerbating the multitalker processing cost. This is precisely what they observed in their third experiment. With a one-digit preload, they observed larger-than-normal mixed-talker processing costs (∼30 ms vs. ∼20 ms in previous studies). With a 3-digit preload, there was virtually no change in response times in blocked-talker conditions (if anything, there was a slight numerical decline), but the multitalker cost increased to nearly 60 ms. This significant interaction is consistent with the logic that the added attentional demands of the mixed-talker condition would stress the (normally automatic, efficient) processes of speech perception detectably.

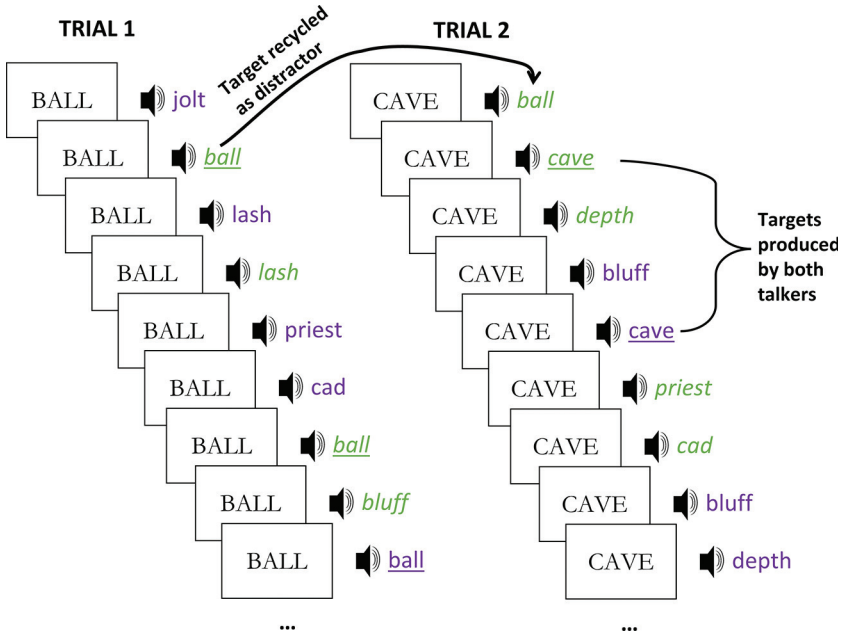
Previous evidence supports the conclusion that talker normalization is influenced by attentional demands, and the mixed-talker trials in the speeded monitoring paradigm were intentionally designed to allow this influence to be observed. However, the speeded monitoring task as conventionally implemented includes another deviation from the preconditions for automaticity: targets are *recycled*. That is, a word that appears as a target on one trial may appear on subsequent trials as a distractor. (In Figure 1, we schematize both deviations from consistent mapping.) In the classic visual search studies of Schneider and Shiffrin (1977), target recycling prevented the development of automaticity, as it violates the principle of consistent mapping. For example, Schneider and Shiffrin (1977) presented participants with displays with one or more target symbols. Participants then had to indicate whether any targets were present in a subsequent display with few or many distractors. Initially, reaction time (RT) increased with the number of distractors. However, if targets were never recycled as distractors, RT flattened out (with little increase with number of distractors), as though participants could search the display in a parallel fashion. This change did not occur if targets were recycled, identifying one of the key preconditions for the development of automaticity.

Unlike the "multiple talkers producing targets" deviation from consistent mapping we have already discussed, this design detail is not constrained to mixed-talker trials; target recycling also occurs for blocked-talker trials. However, it could be that target recycling interacts with talker mixing, as Nusbaum and Morin (1992) found for digit load. That is, it may generate difficulty for blocked- or mixed-talker trials but interact such that its impact is amplified by the attentional and/or resource demands imposed by talker mixing. This led us to ask whether either or both forms of attentional demand (target recycling or multiple talker tokens) disrupt the

[1] In the present work, we focus on normalization based on preceding speech, often termed *extrinsic normalization*. In contrast, most proposals for intrinsic normalization hold that each speech sample contains sufficient information to map acoustics to perceptual categories (Ainsworth, 1975; Lobanov, 1971; Syrdal & Gopal, 1986) and thus do not predict that talker changes should induce processing costs. Whereas extrinsic and intrinsic normalization could be complementary mechanisms that promote phonetic constancy (Nearey, 1989), we focus on contextual tuning theories of extrinsic normalization (Magnuson & Nusbaum, 2007; Magnuson et al., 2021), which explicitly predict processing costs due to talker changes (and subsequent re-computation of the acoustics-to-percepts mapping).

[2] Because normalization is often associated with the notion of *destructive abstraction*, whereby speech is stripped of surface details and mapped to abstract phonological and/or lexical categories, Magnuson and Nusbaum (2007) proposed that a better term might be *talker accommodation*. (They also discussed the fact that most proposals for talker normalization do not explicitly or implicitly propose destructive abstraction.)

Figure 1
Two Abbreviated Mixed-Talker Trials in the Standard Speeded Monitoring Paradigm



Note. Nine items of the full 16 are shown here to conserve space. The rectangles represent the visual display the participant will see (which always has the target listed on screen) with the auditory stimulus they hear to the right. Different talkers are indicated both by color and typeface (green and italicized represents one talker, while purple and normal typeface represents the other talker). In the standard design, items that serve as a target (underlined in this schematic) on one trial can serve as a distractor on subsequent trials; in this example, BALL is the target for the first trial but a distractor for the second. Furthermore, each talker produces the target item on every mixed-talker trial (both the "green" talker and the "purple" talker produce the target items), meaning that participants must respond to two unique productions; by contrast, they need only respond to one unique production on blocked trials. See the online article for the color version of this figure.

normalization process such that multitalker processing costs can be observed. We confirmed that Nusbaum predicted that removing either attentional demand (varied mapping or multiple target tokens) could dampen or wipe out mixed-talker effects, but that whether either or both are crucial for observing talker variability effects had not been explicitly tested (H. C. Nusbaum, personal communication, August 21, 2020).

If we were to remove the two sources of attentional demand in speeded monitoring—having two talkers produce the target items in the mixed-talker condition (doubling the number of unique tokens that need to be monitored for) and target recycling—at least four outcomes are possible. First, it is possible that talker changes have sufficient impact that we would still observe increased processing difficulty in mixed-talker trials relative to blocked-talker trials. Second, on the logic proposed by Nusbaum and Morin (1992), some degree of varied mapping may be required to induce sufficient demands on attention to induce detectable mixed-talker effects, and either form of varied mapping may suffice. Third, it may be that only one of the two aspects of varied mapping matters. Finally, it may be that both are required to induce sufficient attentional demands.

In the present study, we tested these possibilities. We first attempted to replicate previous studies that have shown a multitalker processing cost with the speeded monitoring paradigm, following the approach that has been used in previous work (Experiment 1); critically, this approach recycles targets as distractors and necessitates monitoring for multiple target tokens in mixed-talker trials (one per talker), but only one target token in blocked-talker trials. In subsequent experiments (Experiments 2 through 4), we modified the paradigm to eliminate target recycling and/or to control for the number of talkers producing target tokens for blocked-talker trials and mixed-talker trials. The 2 × 2 design for the experiments in this study is summarized in Figure 2.

Method

Stimuli

Stimuli were produced by four native speakers of American English (two men, two women), who were recorded in a sound-attenuated booth using a RØDE NT-1 condenser microphone (RØDE Microphones LLC, Sydney, Australia) with a Focusrite Scarlet 6i6 digital audio interface (Focusrite PLC, High Wycombe, United Kingdom). Each talker produced three repetitions of each of 19 phonetically distinct words from the word monitoring study

Figure 2
Overview of the Designs for the Experiments

	Number of talkers producing targets on mixed trials	
	Two	**One**
Target recycling	Experiment 1	Experiment 4
No target recycling	Experiment 3	Experiment 2

Note. In the standard design (Experiment 1), two talkers produce the target items on mixed-talker trials, and an item that serves as a target on one trial can serve as a distractor on a subsequent trial. The other experiments remove one or both design features (Experiment 2 removes both, Experiment 3 isolates the impact of multiple talkers producing mixed-trial targets, and Experiment 4 isolates the impact of recycling targets as distractors).

of Nusbaum and Morin (1992). Productions from two talkers (one man, one woman) were selected for the word monitoring experiments described in this study. We selected the best tokens from each talker's repetitions and edited them to remove leading and trailing silence. All stimuli were scaled to an RMS amplitude of 70 dB SPL in Praat (Boersma & Weenik, 2017). The stimuli were otherwise unmodified. We note that the durations of the female talker's stimuli (M = 606 ms) were significantly longer than those of the male talker (M = 568 ms), as indicated by a paired t test, t (18) = 2.20, p = .04; however, we do not believe that this difference has any theoretical or functional implications, and so we did not modify the original stimuli. Stimuli were delivered via OpenSesame v3.2.4 (Mathôt et al., 2012) through Sony MDR-7506 Stereo professional headphones (Sony Group Corporation, Minato City, Tokyo, Japan) or Sennheiser HD 595 headphones (Sennheiser electronic GmbH & Co. KG, Wedemark, Germany).

Participants

We analyzed data from 176 participants (47 men, 126 women, three not reported). Across all four experiments, 183 participants were recruited in total and seven were excluded based on poor accuracy, see Table 1). For all experiments, participants were recruited through the University of Connecticut Psychological Sciences participant pool. All participants indicated that they were monolingual English speakers with normal or corrected-to-normal vision and hearing and no history of speech, language, or neurological impairments. Written informed consent was obtained from every participant in accordance with the guidelines of the University of Connecticut Institutional Review Board. Participants received course credit for their participation.

Given that accuracy tends to be high in word monitoring experiments (e.g., Heald & Nusbaum, 2014; Magnuson & Nusbaum, 2007), we decided a priori to exclude participants with accuracy levels below 90% (collapsing across mixed-talker and blocked-talker trials). This criterion has been used in previous studies on talker normalization (e.g., Choi & Perrachione, 2019b). For each experiment, we recruited until we had 44 participants who met the 90% accuracy criterion. Our sample size was based on a different word monitoring study conducted in our lab where we considered how multitalker penalties (measured within subject) might be modulated by a between-subjects factor (Luthra et al., 2021). For that study, a power analysis of previous data (Magnuson & Nusbaum, 2007) demonstrated that 42 participants per level of the between-subjects factor were necessary for power of .90 at an α of .05 given an estimated mean effect size of approximately η^2_p = .114 (the effect size for the critical significant interaction in Magnuson & Nusbaum, 2007). In this study, there are no between-subjects factors, so 42 participants per experiment should be adequate for statistical power. We rounded this up to 44 so that our number is divisible by four (for counterbalancing whether subjects receive mixed/blocked-talker trials first and whether they receive male/female blocked trials first).

Procedure

Participants first went through the informed consent process and then were seated at a testing computer. They were instructed that in each trial they would hear a series of words and should press the spacebar on the keyboard as quickly as possible any time they heard the target word, which would be identified on-screen shortly before the trial began.

Each subject received 48 mixed-talker trials and 48 blocked-talker trials; we counterbalanced whether participants received all their mixed trials first or all their blocked trials first. In a given blocked trial, the stimuli were either all spoken by the male speaker or all by the female speaker. Within the blocked-talker trials, we counterbalanced whether participants received all the male or female blocked-talker trials first.

Each trial contained 16 auditory tokens, and the target appeared four times in each trial. The target did not appear in Positions 1 or 16, and there was always at least one distractor between two targets (i.e., targets did not appear consecutively). A unique randomization was generated for every subject. Following Heald and Nusbaum (2014), we set an intertrial interval (ITI) of 2,500 ms. This ITI consisted of a fixation cross for the first 1,000 ms; a blank screen for the next 250 ms; and then the visual presentation of the target word for the upcoming trial. Immediately following the ITI, the stimulus train for the trial began, with a stimulus onset

Table 1
Sample Size and Gender Composition of the Experiments

Sample size/gender composition	Experiment 1	Experiment 2	Experiment 3	Experiment 4
N	44	45	47	47
No. excluded for low accuracy	0	1	3	3
Gender				
Female	29	36	30	31
Male	13	8	13	13
Not reported	2		1	

Note. low accuracy is defined as accuracy levels below 90% on word monitoring task (collapsing across mixed-talker and blocked-talker trials).

asynchrony of 750 ms. The target word remained on screen for the duration of the trial. The outcome of interest was the RT to target items. Following Magnuson and Nusbaum (2007), RT was measured from stimulus onset, and RTs that occurred within 150 ms of stimulus onset were considered as a response to the previous item.

In Experiment 1, we sought to replicate the finding that that multitalker processing costs can be elicited in the standard word monitoring paradigm, as has been previously found (Heald & Nusbaum, 2014; Magnuson & Nusbaum, 2007; Magnuson et al., 2021; Nusbaum & Magnuson, 1997; Nusbaum & Morin, 1992). In keeping with the previous studies, items that served as targets could be used as distractors on subsequent trials. Furthermore, in every mixed-talker trial, the target was produced twice by the male talker and twice by the female talker. Thus, experiment 1 included both target recycling on blocked and mixed trials, and single-target tokens on blocked trials but targets produced by multiple talkers within mixed trials.

In Experiment 2, we modified the speeded monitoring paradigm to address two features of the standard paradigm that prevent conditions for consistent mapping. First, we ensured that target items would never be recycled as distractors in other trials. Second, we modified mixed-talker trials such that only one talker produced the target item, although both talkers produced distractor items. This maintains the same level of acoustic variability in the mixed-talker trials as in experiment 1 but reduces the potential working memory load, as subjects only need to monitor for one unique production.

In Experiment 3, we did not allow target recycling in the speeded monitoring paradigm to ensure that the mapping between targets and responses was fully consistent (i.e., items that served as a target on one trial could not serve as a distractor on another). However, as in the conventional design, targets in mixed-talker trials were produced by each talker. Thus, Experiment 3 tests whether multitalker processing costs in the monitoring paradigm can be driven solely by the need to respond to multiple target tokens.

In Experiment 4, we test the possibility that target recycling might be a sufficient condition for multitalker processing costs. That is, we asked whether multitalker processing costs persist even when target items are spoken only by one talker, but target recycling is allowed.

Analysis

RT data from trials with correct responses were submitted to a generalized linear mixed-effects model that was implemented in R (R Core Team, 2019) with the packages *lme4* (Bates et al., 2015) and *afex* (Singmann et al., 2020). No responses to target items were filtered or removed based upon their RT. Lo and Andrews (2015) have argued that RT transformations may obscure meaningful differences between conditions and therefore that raw RTs are a more theoretically justified dependent variable. We therefore used generalized linear mixed models for analyzing RTs; such an approach allows for the use of raw RTs as the dependent variable while allowing the user to specify a statistical distribution that reflects the actual distribution of RT. As suggested by Lo and Andrews, we specified a gamma distribution with an identity link. For all experiments, chi-square tests indicated that this approach yielded significantly better model fit than equivalent linear mixed-effects models with either the raw RT data or log-transformed RT data. W

As outlined in our preregistered analysis plan, we identified the most parsimonious random effects structure using a backward-stepping procedure (Matuschek et al., 2017). Likelihood ratio tests were implemented using the "mixed" function in the R *afex* package to test for effects of our fixed factors; we report chi-squared values and associated *p* values from these tests.

Results

Data from all four experiments were submitted to an omnibus analysis that used a generalized linear mixed model with fixed factors of condition (blocked vs. mixed, sum-coded), target recycling (present vs. absent, sum-coded), Number of talkers producing mixed-trials targets (one vs. two, sum-coded), and the accompanying two- and three-way interactions. The model with by-subject random slopes for condition was estimated to have the best fit (see Table 2). There was a significant main effect of condition (χ^2 = 19.63, $p < .001$), indicating that across the experiments, responses were slower to mixed talker trials than blocked talker trials and a significant main effect of target recycling ($\chi^2 = 4.55$, $p = .03$), indicating that responses were slower for experiments where target recycling was absent (Experiments 2 and 3) compared with those where it was present (Experiments 1 and 4). Only the two-way interaction between condition and number of talkers producing mixed-trials targets was significant ($\chi^2 = 6.94$, $p = .008$).

This interaction was explored using the R package *emmeans* (Lenth, 2020) to compare estimated marginal means (EMM) for the effect of condition at each level of number of talkers producing mixed-trials targets. There was a significant difference between blocked and mixed trials when one talker produced the target items in mixed-talker trials (EMM = –5.85, $p = .01$), though this difference was much larger when two talkers produced the target items in mixed talker trials (EMM = –21.56, $p < .0001$), indicating that multitalker processing cost was smaller (though not

Table 2
Generalized Linear Mixed-Effects Model

Fixed effects	$\chi^2(1)$	*p*
Condition (blocked/mixed)	19.63	<.001
Target recycling	4.55	.033
Number of talkers producing mixed-trials targets	0.46	.499
Condition × Target Recycling	0.00	>.999
Condition × Number of Talkers Producing Mixed-Trials Targets	6.94	.008
Target Recycling × Number of Talkers Producing Mixed-Trials Targets	2.59	.108
Condition × Target Recycling × Number of Talkers Producing Mixed-Trials Targets	0.00	.969

Note. For the omnibus analyses, we used the R package *afex* (Singmann et al., 2020).

nonexistent) when only one talker produced the target items in mixed talker trials.

General Discussion

Over the course of four experiments, we investigated the possibility that one, both, or neither of the attention-demanding features in conventional speeded monitoring paradigms might be crucial for observing multitalker processing costs. Specifically, we tested whether detecting this processing cost requires (1) the recycling of target items as distractor items on subsequent trials and/or (2) two talkers producing the target items in mixed-talker trials, thereby requiring the listener to monitor for twice as many unique items as the blocked-talker trials. We found evidence for the latter, as multitalker processing costs were elicited when the mixed-talker condition required responses to two unique tokens (Experiments 1 and 3) but substantially reduced when responses were made to a single target in both mixed and blocked talker (Experiments 2 and 4).

While previous work by Schneider and Shiffrin (1977) led us to hypothesize that the recycling of targets might also be a critical factor governing the emergence of multitalker processing costs[3], we did not find evidence to support this hypothesis. This may be because the visual search task used by Schneider and Shiffrin may differ too much from the word monitoring paradigm. The difference in modality (visual vs. auditory) notwithstanding, a key difference between the auditory monitoring task and their visual search task is the amount of practice participants had with the task; participants in Schneider and Shiffrin's studies had substantial exposure to repeated targets and distractors before the crucial test data were collected (on the order of thousands of trials), while participants in our study had fewer trials, and no prior training with items before data was collected. Schneider and Shiffrin posited that the two criteria for achieving automaticity in processing are consistent mapping and practice to reinforce that mapping; consistent mappings without substantial practice are not sufficient to develop automaticity. Thus, even when targets were not recycled (as in Experiments 2 and 3), participants may have been engaging in controlled processing as the mapping between certain words and their status as a "target" had little reinforcement, and the length of the paradigm used in the reported experiments was unlikely to be sufficient practice to reinforce that. To further investigate this possibility, future work might test whether multitalker processing costs dissipate if targets are not recycled, and participants receive considerable practice with the task. The present results fit into a broader literature suggesting that it is difficult (and perhaps impossible) to consider the problem of talker normalization without considering other aspects of cognitive processing, including the mapping between stimuli and responses, an individual's level of practice with the experimental task, and the degree of cognitive load.

Rather than finding evidence that target recycling was the key factor for eliciting multitalker costs, our results suggest that in the speeded monitoring paradigm, the presence of a multitalker processing cost depends on how many talkers produce the target stimuli in mixed-talker trials. As the speeded monitoring paradigm does not require participants to make responses to most items, and because participants in Experiments 2 and 4 only needed to respond to one talker's productions for a given mixed-talker trial, it is possible that listeners may have been able to effectively ignore the second talker who was only producing task-irrelevant distractors. As such, performance on mixed-talker trials may have been similar to performance on blocked-talker trials insofar as there was only a single target to monitor for on a given trial. This observation is consistent with the well-attested "cocktail party" effect (see Shinn-Cunningham, 2008; for a review)—the ability for listeners to attend to and segment one stream of speech from competing, irrelevant information (Cherry, 1953). When only a single talker produces the target items, the selective-attention required for mixed-talker trials changes—the target consists of only a single combination of talker and item, which reduces the cognitive demands in place, and perhaps allowing normalization to occur automatically and nearly undetectably. That said, it is important to note that the identity of the talker producing the target items on mixed-talker trials varied from trial to trial, so subjects could not have known in advance which talker they needed to attend to (at least prior to the first target on a given mixed-talker trial).

In other words, our results suggest that the key factor governing the emergence of multitalker processing costs in the speeded word monitoring paradigm is whether both talkers are behaviorally relevant as regards participants' responses. Our results suggest that when all the target items are produced by one talker (i.e., only one talker is behaviorally relevant), then the costs involved in talker normalization are dramatically reduced. The latter position—namely, that talker normalization is a highly-automatized process that is only observable when listeners must engage in highly controlled processing—is consistent with the stance taken by Nusbaum and Morin (1992).

Our findings suggest that to produce measurable multitalker penalties in speeded monitoring paradigms, researchers should ensure that both talkers are behaviorally relevant (i.e., that listeners must make behavioral responses to both talkers) in order to elicit multitalker processing costs. However, while both talkers are indeed behaviorally relevant in the standard speeded monitoring paradigm (Experiment 1), the standard design has inherent asymmetries between mixed-talker and blocked-talker trials with regard to the number of tokens (i.e., unique stimuli) to which listeners must respond. This makes it difficult to determine whether the observed multitalker processing costs are truly a result of talker normalization per se or a result of general acoustic variation. While previous work by Magnuson and Nusbaum (2007) suggests that not all acoustic variation (e.g., changes in amplitude) elicits a processing cost, we suggest that additional studies are needed to distinguish whether the processing costs in this paradigm are specifically due to talker variation.

It is important to acknowledge that multitalker processing costs have also been observed in other paradigms, and thus are unlikely to be an artifact of the monitoring paradigm. For example, Mullennix et al. (1989) assigned participants to either a blocked-talker group or multitalker group and asked them to identify what words were spoken. Across a range of signal-to-noise ratios, participants in the multitalker group were reliably slower to respond and less accurate than those in the blocked-talker group. Regardless of whether they were asked to type the word or speak it aloud,

[3] As we noted earlier, whereas target recycling occurs in both blocked- and mixed-talker conditions in the monitoring paradigm, it could have interacted with talker mixing by contributing additional attentional demand to allow multitalker processing costs to be observed.

participants in the multitalker group were reliably slower to respond and less accurate than those in the blocked-talker group. Multitalker processing costs have also been repeatedly observed in the speeded classification paradigm (Carter et al., 2019; Choi et al., 2018; Choi & Perrachione, 2019a, 2019b; Kapadia & Perrachione, 2020; Lim, Qu, et al., 2019), where listeners hear a single item (e.g., *boot*) on each trial and must indicate what they heard from a limited set of response options (e.g., *boot* or *boat*). Notably, in both tasks, listeners must make a behavioral response on every item, meaning that both talkers are behaviorally relevant. This again points to the fact that normalization may only occur (or that multitalker processing costs may only be measurable) when changes in talker are kept in the attentional focus.

More generally, in considering the utility of multitalker processing costs as a tool for studying talker normalization, it is worth noting that some researchers have suggested that multitalker processing costs may emerge simply because there is a break in low-level acoustic information that disrupts auditory streaming (Choi & Perrachione, 2019b; Lim, Shinn-Cunningham, et al., 2019), rather than reflecting talker normalization per se. Specifically, when listeners hear speech from one talker, they can attribute ongoing variation in the auditory signal to a single physical source with relative ease—that is, they can group the relevant auditory input into a single auditory object. By contrast, when the speech signal alternates between two talkers, the formation of one auditory object (for the first talker) may be disrupted by the need to form a second auditory object (for the second talker). In their view, this makes it harder to attend to—and thus harder to perceptually analyze—the speech signal, yielding multitalker processing costs. However, our results appear inconsistent with this notion. The streaming account should predict multitalker processing costs even when mixed-talker trial targets are produced by only one talker (since there is still talker variability within the trial, with equivalent numbers of talker changes), which was not the case in Experiments 2 and 4. Such a result suggests that multitalker processing costs indeed reflect a process of talker normalization/accommodation, rather than emerging simply because of disruptions in auditory object formation.

Conclusion

It is worth underscoring that the lack of invariance problem remains a critical issue for research on speech perception, and despite decades of concerted effort, as a field, we are still far from understanding how listeners accommodate sources of variance, including variation between talkers. For proponents of the view that phonetic constancy results from active, controlled processing, our results identify the potentially crucial attentional aspect of speeded monitoring for detecting the operation of talker normalization. These findings also call into question the automaticity of talker normalization, suggesting that processing penalties may only emerge (or may only be observable) when the talker change is in attentional focus. Future work will be required to further elucidate the nature of the attention-demanding processing mechanisms that appear to be associated with maintaining phonetic constancy, and to fully equate sources of variability between blocked- and mixed-talker trials.

References

Ainsworth, S. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. A. A. Tatham (Eds.), *Auditory analysis and the perception of speech* (pp. 103–113). Academic Press.

Antoniou, M., Wong, P. C. M., & Wang, S. (2015). The effect of intensified language exposure on accommodating talker variability. *Journal of Speech, Language, and Hearing Research*, *58*(3), 722–727. https://doi.org/10.1044/2015_JSLHR-S-14-0259

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Boersma, P., & Weenik, D. (2017). *Praat: Doing phonetics by computer*.

Bosker, H. R. (2018). Putting Laurel and Yanny in context. *The Journal of the Acoustical Society of America*, *144*(6), EL503–EL508. https://doi.org/10.1121/1.5070144

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

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with two ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229

Choi, J. Y., & Perrachione, T. K. (2019a). 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

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

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. https://doi.org/10.3758/s13414-017-1395-5

Dorman, M. F., Studdert-Kennedy, M., & Raphael, L. J. (1977). Stop-consonant recognition: Release bursts and formant transitions as functionally equivalent, context-dependent cues. *Perception & Psychophysics*, *22*(2), 109–122. https://doi.org/10.3758/BF03198744

Francis, A. L., & Nusbaum, H. C. (1996). Paying attention to speaking rate. In H. T. Bunnell & W. Idsardi (Eds.), *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96* (Vol. 3, pp. 1537–1540).

Heald, S. L. M., & Nusbaum, H. C. (2014). Talker variability in audio-visual speech perception. *Frontiers in Psychology*, *5*, 698. https://doi.org/10.3389/fpsyg.2014.00698

Joos, M. (1948). Acoustic phonetics. *Language*, *24*(2), 5–136. https://doi.org/10.2307/522229

Kapadia, A. M., & Perrachione, T. K. (2020). Selecting among competing models of talker adaptation: Attention, cognition, and memory in speech processing efficiency. *Cognition*, *204*, 104393. https://doi.org/10.1016/j.cognition.2020.104393

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *The Journal of the Acoustical Society of America*, *29*(1), 98–104. https://doi.org/10.1121/1.1908694

Laing, E. J. C., Liu, R., Lotto, A. J., & Holt, L. L. (2012). Tuned with a tune: Talker normalization via general auditory processes. *Frontiers in Psychology*, *3*, 203. https://doi.org/10.3389/fpsyg.2012.00203

Lenth, R. V. (2020). emmeans: Estimated Marginal Means, aka Least-Squares Means (R package version 1.5.3.). https://CRAN.R-project.org/package=emmeans

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368. https://doi.org/10.1037/h0044417

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

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. https://doi.org/10.3758/s13414-019-01684-w

Lo, S., & Andrews, S. (2015)., August To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171. https://doi.org/10.3389/fpsyg.2015.01171

Lobanov, B. M. (1971). Classification of Russian vowels spoken by different speakers. *The Journal of the Acoustical Society of America*, *49*(2B), 606–608. https://doi.org/10.1121/1.1912396

Luthra, S., Saltzman, D., Myers, E. B., & Magnuson, J. S. (2021). Listener expectations and the perceptual accommodation of talker variability: A pre-registered replication. *Attention, Perception & Psychophysics*, *83*(6), 2367–2376. https://doi.org/10.3758/s13414-021-02317-x

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology*, *33*(2), 391–409. https://doi.org/10.1037/0096-1523.33.2.391

Magnuson, J. S., Nusbaum, H. C., Akahane-Yamada, R., & Saltzman, D. (2021). Talker familiarity and the accommodation of talker variability. *Attention, Perception & Psychophysics*, *83*(4), 1842–1860. Advance online publication. https://doi.org/10.3758/s13414-020-02203-y

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods*, *44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315.

Mullennix, J. W., Pisoni, D. B., & Martin, C. S. (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–378. https://doi.org/10.1121/1.397688

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, *85*(5), 2088–2113. https://doi.org/10.1121/1.397861

Nusbaum, H., & Magnuson, J. (1997). Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J. Mullenix (Eds.), *Talker variability in speech processing* (p. 109132). Academic Press.

Nusbaum, H., & Morin, T. (1992). Paying attention to differences among talkers. In Y. Tohkura, E. Bateson, & Y. Sagisaka (Eds.), *Speech perception, production, and linguistic structure* (pp. 6694). IOS Press.

Nusbaum, H. C., & Schwab, E. C. (1986). The role of attention and active processing in speech perception. In E. C. Schwab & H. C. Nusbaum (Eds.), *Pattern Recognition by Humans and Machines*, Volume *1*: *Speech Perception* (pp. 113–157). Academic Press. https://doi.org/10.1016/B978-0-12-631403-8.50009-6

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the Acoustical Society of America*, *24*(2), 175–184. https://doi.org/10.1121/1.1906875

R Core Team. (2019). *R: A language and environment for statistical computing*. https://www.R-project.org/

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review*, *84*(1), 1–66. https://doi.org/10.1037/0033-295X.84.1.1

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological Review*, *84*(2), 127–190. https://doi.org/10.1037/0033-295X.84.2.127

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). afex: Analysis of factorial experiments (R package version 0.28-0). https://CRAN.R-project.org/package=afex

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

Stilp, C. E. (2019). Auditory enhancement and spectral contrast effects in speech perception. *The Journal of the Acoustical Society of America*, *146*(2), 1503–1517. https://doi.org/10.1121/1.5120181

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America*, *79*(4), 1086–1100. https://doi.org/10.1121/1.393381

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talker's vowel space? *The Journal of the Acoustical Society of America*, *60*(1), 198–121. https://doi.org/10.1121/1.381065

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience*, *16*(7), 1173–1184. https://doi.org/10.1162/0898929041920522

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. https://doi.org/10.1016/j.bandl.2013.05.010

Received January 12, 2021
Revision received June 30, 2021
Accepted August 24, 2021 ■