Attention, task demands, and multi-talker processing costs in speech perception

David Saltzman

Sahil Luthra

Emily B. Myers

James S. Magnuson

University of Connecticut

Author Note

This research was supported by NIH grant R01 H14-001 (PI: EBM) and NSF grants NRT 1747486 and PAC 1754284 (PI: JSM). SL was supported by an NSF Graduate Research Fellowship. This work was presented at the 61st Annual Meeting of the Psychonomic Society. As noted in the manuscript, the preregistration plan for this study is available on the Open Science Framework (https://osf.io/wx4kd), and all stimuli and analysis code can be found at https://github.com/disaltzman/TalkerTeam-Mapping.

CORRESPONDING AUTHOR:
David Saltzman
Psychological Sciences
University of Connecticut
Storrs, CT 06269-1020
david.saltzman@uconn.edu

Abstract

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 to a single-talker baseline. These *multi-talker 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* (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 (one target produced by each talker), whereas in blocked conditions, they need only respond to one token (*multiple target tokens*). We ask whether observing multi-talker processing costs depends upon either or both of these attentional demands. Across four experiments, multi-talker processing costs persisted when target recycling was not allowed but dissipated 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.

Keywords

Talker normalization, word monitoring, automaticity, phonetic constancy

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.

Word count: 6283

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, Harris, Hoffman, & Griffith, 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, Studdert-Kennedy, & Raphael, 1977). Nonetheless, listeners typically perceive the content of the speech signal with ease, achieving *phonetic constancy* in spite of 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; Ladefoged & Broadbent, 1957; Laing, Liu, Lotto, & Holt, 2012; Sjerps, Fox, Johnson, &

Chang, 2018; Stilp, 2019; Zhang, Peng, & Wang, 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 to when all the words are spoken by a single talker (Carter, Lim, & Perrachione, 2019; Choi, Hu, & Perrachione, 2018; Choi & Perrachione, 2019b, 2019a; Heald & Nusbaum, 2014; Kapadia & Perrachione, 2020; Magnuson & Nusbaum, 2007; Mullennix, Pisoni, & Martin, 1989; Nusbaum & Morin, 1992; Verbrugge, Strange, Shankweiler, & Edman, 1976; Wong, Nusbaum, & Small, 2004). Some have interpreted these *multi-talker 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 re-engage 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, Wong, & Wang, 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, the target and distractor items are produced by two different talkers

[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. While 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, Nusbaum, Akahane-Yamada, & Saltzman, 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.)

who are intermixed. As expected by normalization/accommodation accounts, listeners are slower to identify the target word in mixed-talker trials compared to 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 pointed to the fact that in a blocked-talker trial, participants must make a response to a single target item, while in mixed trials, participants must respond to two distinct stimuli (one produced by each talker): therefore, they noted, "…from a cognitive perspective, recognition in the mixed-talker condition should require more effort and attention than recognition in the blocked-talker condition."

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 multi-talker processing cost. This is precisely what they observed in their third experiment. With a 1-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 multi-talker 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.

The present study was motivated by our observation that 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*[3]. Unlike the "multiple target" 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 (varied mapping/target recycling or multiple talker tokens) in the paradigm are crucial for detecting mixed talker effects. We confirmed that Nusbaum (personal communication, August 21, 2020) predicted

[3] 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 increased with the number of distractors. However, if targets were never recycled, reaction time 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.

that removing either attentional demand (varied mapping or multiple target tokens) could damp or wipe out mixed-talker effects, but that whether either or both are crucial for observing talker variability effects had not been explicitly tested.

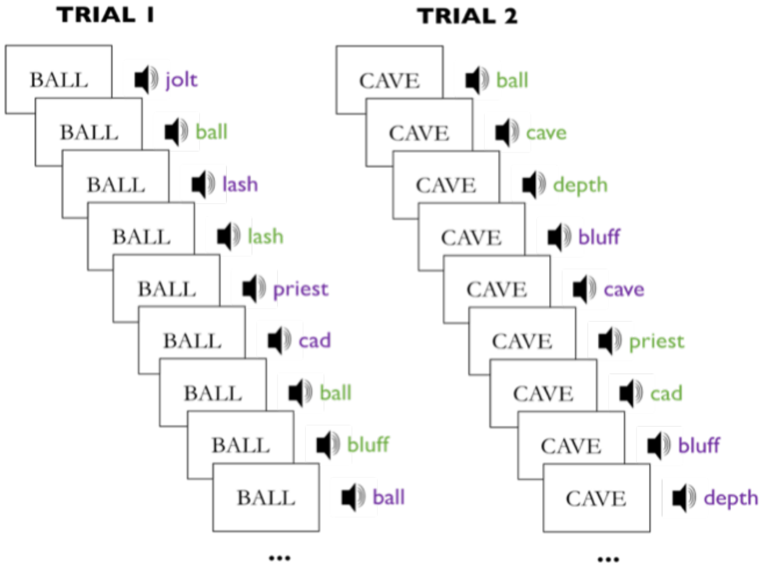

Figure 1. Example of two mixed-talker trials in the standard speeded monitoring paradigm. Different talkers are indicated with colored text. In the standard design, items that serve as a target on one trial can serve as a distractor on subsequent trials; in this example, *BALL* is the target for the first trial but a distractor for the second. Furthermore, each talker produces the target item on every mixed-talker trial, meaning that participants must respond to two unique productions; by contrast, they need only respond to one unique production on blocked trials.

If we were to remove the two sources of attentional demand in speeded monitoring – multiple target tokens rather than a single target stimulus in the mixed-talker condition, 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 multi-talker 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 multiple target tokens in mixed-talker trials (1 per talker), but only one target token in blocked-talker trials. In subsequent experiments (Experiments 2-4), 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 x 2 design for the experiments in this study is summarized in Figure 2.

Study Design		**Number of talkers producing targets on mixed trials**	
		Two	**One**
Can target items be recycled as distractors?	**Yes**	Experiment 1	Experiment 4
	No	Experiment 3	Experiment 2

Figure 2. Overview of the designs for the four experiments testing whether multi-talker processing costs observed in previous studies are attributable to specific design features of the speeded word monitoring paradigm. 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 of these design features.

General Methods

We pre-registered our experimental design and analysis plans on the Open Science Framework (https://osf.io/wx4kd) prior to data collection. For expository clarity, we have revised the order of the experiments in this paper. All stimuli and analysis scripts are available at https://github.com/disaltzman/TalkerTeam-Mapping.

Stimuli

Stimuli were produced by four native speakers of American English (two males, two females), who were recorded in a sound-attenuated booth using a RØDE NT-1 condenser microphone with a Focusrite Scarlet 6i6 digital audio interface. Each talker produced three repetitions of each of 19 phonetically distinct words from the word monitoring study of Nusbaum and Morin (1992). Productions from two talkers (one male, one female) 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, $t(18) = 2.20$, $p = 0.04$; 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 through Sony MDR-7506 or Sennheiser HD-595 headphones.

Participants

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 IRB. 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 and blocked 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 whose data could be used in analyses. Our sample size was based on a different word monitoring study being conducted in our lab where we consider how multi-talker penalties (measured within subject) might be modulated by a between-subjects factor. 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 0.90 at an α of 0.05 given an estimated mean effect size of approximately partial η2 = 0.114 (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 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 trials, we counterbalanced whether participants received all of the male or female blocked 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 inter-trial interval (ITI) of 2500 ms. This ITI consisted of a fixation cross for the first 1000 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 asynchrony of 750 ms. The target word remained on screen for the duration of the trial. The outcome of interest was the reaction time (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.

Analysis

In all four experiments, RT data were submitted to a generalized linear mixed-effects model that was implemented in R (R Core Team, 2019) with the packages *lme4* (Bates, Mächler, Bolker, &

Walker, 2015) and *afex* (Singmann, Bolker, Westfall, Aust, & Ben-Shacar, 2020). In our pre-registered analysis plan, we described our intent to use a linear mixed-effects model with either log-transformed RT data or raw RT data. However, Lo and Andrews (2015) have argued that neither of these approaches is optimal. Instead, they argue that RT transformations may obscure meaningful differences between conditions and therefore that raw RTs are a more theoretically justified dependent variable. However, linear mixed-effects models assume that the dependent variable has a Gaussian distribution, and raw RTs violate this assumption. Lo and Andrews therefore advocate for the use of a *generalized* linear mixed model 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 such, we implemented generalized linear models, specifying a gamma distribution with an identity link (as suggested by Lo and Andrews). 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. We therefore used generalized linear mixed models for all RT analyses, noting this as the only deviation from our pre-registered analysis plan.

As outlined in our pre-registered analysis plan, we identified the most parsimonious random effects structure using a backwards-stepping procedure (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 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.

Experiment 1

In Experiment 1, we first sought to replicate the finding that that multi-talker 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. Given previous research, we expected slower reaction times for the mixed-talker trials compared to the blocked-talker trials.

Participants

44 undergraduates (13 male, 29 female, 2 no report) were recruited from the University of Connecticut Psychological Sciences participant pool. No participants in Experiment 1 met our 90% accuracy exclusion criterion (mean accuracy = 98%, range = 92% to 100%), and thus data from all 44 participants were included in our analysis.

Results

Results are shown in Figure 3A and Figure 3B. On average, participants were faster to identify targets in the Blocked condition (*M* = 508 ms, *SD* = 119 ms) compared to the Mixed condition (*M* = 531 ms, *SD* = 111 ms).

To confirm the presence of a multi-talker processing cost, we first created a model with a fixed effect of Condition (Blocked vs. Mixed, sum-coded). Our backward-stepping procedure selected the maximal random effect structure (by-subject random intercepts and by-subject random slopes for Condition). The fixed effect of Condition was significant, $\chi 2 = 8.03$, $p = 0.005$), indicating that the presence of mixed talkers slowed participants' responses to the target item.

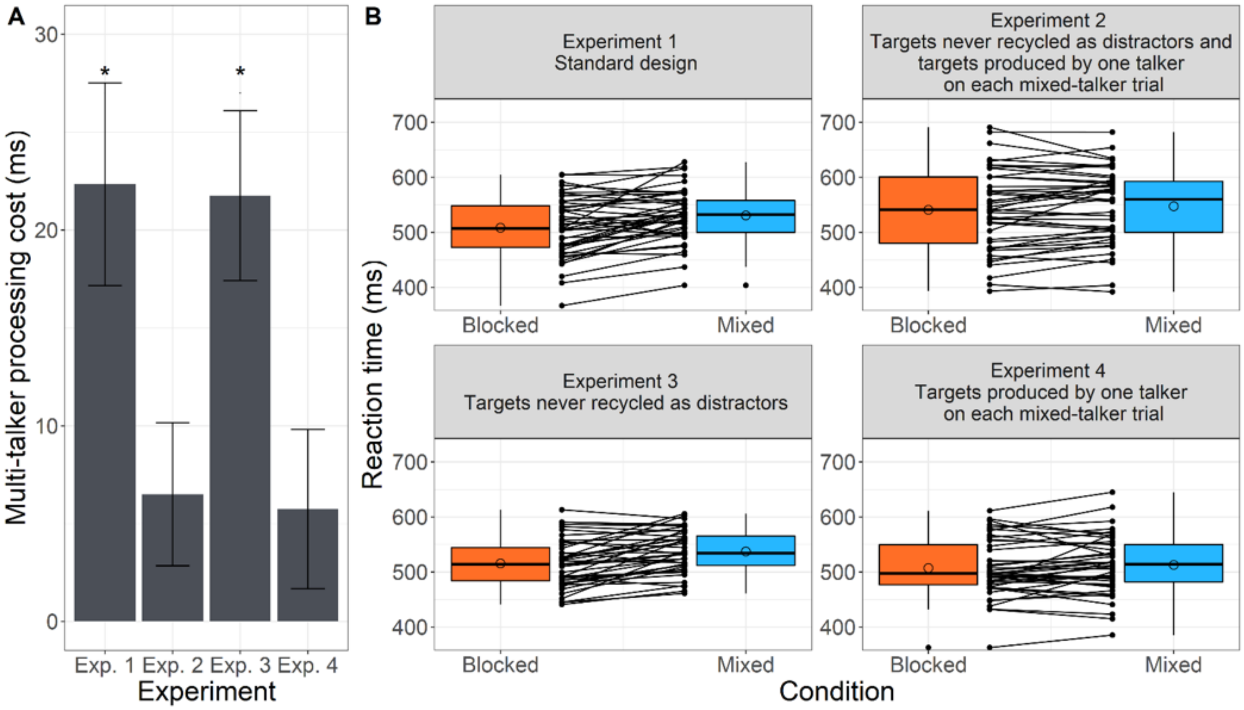

Figure 3. (A) Multi-talker processing cost in each experiment, calculated by subtracting the average reaction time for the Blocked condition from the Mixed condition. Stars indicate significance (*p* < 0.01). Error bars reflect standard error of the mean. (B) Reaction time data as a function of condition for each of the four experiments. In the box-and-whisker plots, the median is represented by a horizontal line, while the mean is represented by an open circle. Points with connecting lines represent data from each individual participant.

Discussion

Results of Experiment 1 replicated previous work showing that multi-talker processing costs are elicited using the standard speeded monitoring paradigm. However, as we discussed in the introduction, there are two aspects of the word-monitoring paradigm that may be essential for detecting this effect: (1) the recycling of target items as distractors in later trials, and (2) the fact that listeners have to monitor for two distinct productions in the mixed-talker trials compared to just one production in the blocked-talker trials. For Experiment 2, we eliminated both aspects from the paradigm.

Experiment 2

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. If detecting multi-talker processing costs in this paradigm requires some elevation of attentional demands, we should not observe multi-talker processing costs in this experiment.

Participants

47 participants were recruited for this experiment. Three participants failed to meet our accuracy-based inclusion criterion of 90% accuracy, so these participants' data were excluded. For the 44 participants (8 male, 36 female) included in analyses, accuracy in the word monitoring task was near ceiling (*M* = 97%, range = 92% to 100%).

Procedure

The procedure for Experiment 2 was identical to that of Experiment 1, with two exceptions. First, words that served as a target item in one trial (*ball*, *tile*, *cave*, and *done*) could not be recycled as a distractor in any subsequent trials. Second, in mixed-talker trials, only one of the two talkers produced the target items in a given trial, while both talkers produced distractors. Importantly, on these mixed trials, the identity of the talker who was producing the target items varied randomly from trial to trial, and participants heard an equal number of stimuli from each talker on every trial. To be precise, in Experiment 1, each talker contributed two targets and six distractors to each mixed trial; in Experiment 2, one talker contributed all four targets and four distractors, while the other talker contributed eight distractors (keeping the number of items per talker constant in each mixed trial). Both talkers were assigned to produce the target items in equal numbers of mixed trials (24 trials each) and the identity of the talker producing the targets in the mixed condition varied from trial to trial. Counterbalancing strategies were the same as in Experiment 1.

Results

Results are shown in Figure 3A and Figure 3B. Participants were on average faster to identify targets in the Blocked condition (*M* = 541 ms, *SD* = 129 ms) compared to the Mixed condition (*M* = 547 ms, *SD* = 124 ms).

The same analysis approach was used as in Experiment 1. As before, our backward-stepping procedure identified the maximal random effect structure (by-subject random intercepts and slopes and for Condition) as the most parsimonious. The model indicated that the fixed effect of Condition (Blocked vs. Mixed, sum-coded) was not significant, *p* = 0.23. Thus, we did not observe a robust multi-talker processing cost in this study.

Discussion

In Experiment 2, the multi-talker processing cost in the mixed talker trials was eliminated when the paradigm was modified to eliminate both features of the standard speeded monitoring paradigm that induce attentional demands. Crucially, this suggests that detecting multi-talker processing costs in the speeded monitoring paradigm requires some degree of attentional demands.

However, it is not clear from the results of Experiment 2 whether one or both forms of varied mapping in the conventional design are sufficient to elevate attentional demands to the degree needed to detect multi-talker processing costs such as those observed in Experiment 1. In Experiments 3 and 4, we tested the contribution of target recycling and multiple-target mapping individually. If both are required to induce attentional demands sufficient for detecting multi-talker processing costs, we should not observe those costs in either experiment. If either one is sufficient, we should observe mixed-talker costs in both experiments. If we only observe significant mixed-talker costs in one study, this will identify an attentional demand that may be essential for detecting mixed-talker costs in this paradigm.

Experiment 3

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 multi-talker processing costs in the monitoring paradigm can be driven solely by the need to respond to multiple target tokens in the mixed talker condition.

Participants

47 undergraduates were recruited for this experiment. Three participants met our exclusionary criterion, and their data were not included. Thus, analyses reflect data from 44 participants (13 male, 30 female, 1 no report), who had near-ceiling accuracy on the word monitoring task (*M* = 98%, range = 93% to 99%).

Procedure

The procedure for Experiment 3 was identical to that of Experiment 1, with the added constraint that a word that served as a target item in one trial could not be recycled as a distractor in any subsequent trials. This rule was implemented across both the blocked talker and mixed talker trials. The counterbalancing strategies were the same as in Experiment 1.

Results

Results are shown in Figure 3A and Figure 3B. Participants were on average faster to identify targets in the Blocked condition (*M* = 515 ms, *SD* = 117 ms) compared to the Mixed condition (*M* = 537 ms, *SD* = 114 ms).

The RT data from Experiment 3 were submitted to the same generalized linear mixed-effects model approach as in Experiments 1 and 2. As before, the selected model had a fixed factor of Condition (Blocked vs. Mixed, sum-coded) as well as by-subject random intercepts and random slopes for Condition. Results indicated a significant effect of Condition ($\chi 2$ = 10.67, p = 0.001), indicating that the presence of mixed talkers slowed participants' responses to the target item when multiple target tokens required responses in mixed talker trials, even when targets were not recycled.

Experiment 4

Experiment 3 indicated that having target items produced by both talkers was a sufficient condition for observing multi-talker processing costs. That is, multi-talker processing costs persisted even when target recycling was eliminated. In Experiment 4, we test the possibility that target recycling could also be a sufficient condition for multi-talker processing costs. That is, we asked whether multi-talker processing costs persist even when target items are spoken only by one talker but target recycling is allowed.

Participants

45 undergraduates were recruited for this experiment. One met our pre-specified criterion for exclusion, leaving 44 participants (13 male, 31 female) in our analysis. Among included participants, accuracy in the word monitoring task was near ceiling (*M* = 98%, range = 91% to 100%).

Procedure

The procedure for Experiment 4 was identical to that of Experiment 1, except for one change in the mixed-talker trials. Specifically, in mixed trials, target items were only produced by one of the talkers on a given trial, whereas distractors were produced by both talkers. Critically, participants heard an equal number of stimuli produced by each talker on every mixed trial; as in Experiment 2, one talker contributed all four targets and four distractors, while the other talker contributed eight distractors to keep the number of items per talker constant in each mixed trial. The identity of the talker producing the targets in the mixed condition varied from trial to trial. The blocked talker trials were the same as described in Experiment 1, as were the counterbalancing strategies.

Results

Results are shown in Figure 3A and Figure 3B. Participants were on average slightly faster to identify targets in the Blocked condition (*M* = 507 ms, *SD* = 119 ms) compared to the Mixed condition (*M* = 513 ms, *SD* = 115 ms). Trial-level data were submitted to a generalized linear mixed model with a fixed factor of Condition (Blocked vs. Mixed, sum-coded), as in previous experiments. As before, our backward stepping procedure identified the maximal random effect structure (by-subject random intercepts and by-subject random slopes for Condition) as the most parsimonious structure. Results indicated that the effect of Condition was not significant ($p = 0.30$), indicating that there was not a robust multi-talker processing cost in this experiment[4].

Discussion

In Experiment 4, we did not observe robust multi-talker processing costs when listeners only needed to respond to one token on mixed trials (though crucially, these mixed trials still included talker variability, as distractors were produced by both talkers). This is consistent with the proposal of Nusbaum and Morin (1992) that this aspect of the design plays an important role in creating attentional demands that stress the system sufficiently that mixed-talker effects can be observed, possibly as the result of an attention-demanding process that recomputes acoustic-percept mappings when there is a talker change.

General Discussion

Over the course of four experiments, we investigated the possibility that the either, both, or neither of the attention-demanding features in conventional speeded monitoring paradigms might be crucial for observing multi-talker 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) a difference in the number of unique target productions between blocked-talker trials and mixed-talker trials, which leads to different response demands. We found evidence for the

[4] Results of binomial tests supported the analyses presented in the main text, as the MTPC was significantly greater than 0 in Exp. 1 and Exp. 3 ($p < 0.01$ for both) but no different from 0 in Exp. 2 and Exp. 4.

latter, as multi-talker processing costs were elicited when the mixed-talker condition required responses to two unique tokens (Experiments 1 and 3) but not 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 would be a critical factor governing the emergence of multi-talker processing costs, 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 versus 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. Schneider and Shiffrin posited that the two criteria for achieving automaticity in processing are consistent mapping and practice; 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. To further investigate this possibility, future work might test whether multi-talker processing costs dissipate if targets are not recycled and participants receive considerable practice with the task.

Rather than finding evidence that target recycling was the key factor for eliciting multi-talker costs, our results suggest that in the speeded monitoring paradigm, the presence of a multi-talker 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. That said, it is important to note that the identity of the talker producing the target items on mixed 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 multi-talker processing costs in the speeded word monitoring paradigm is whether both talkers are *behaviorally relevant* with regard to 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 either talker normalization does not occur, or the task can no longer detect talker normalization. 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 multi-talker 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 multi-talker 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 multi-talker 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 multi-talker 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 multi-talker group and asked them to identify what words were spoken . Across a range of signal-to-noise ratios, participants in the multi-talker 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, participants in the multi-talker group were reliably slower to respond and less accurate than those in the blocked-talker group. Multi-talker 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, Tin, Qu, & Perrachione, 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 of these 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 multi-talker processing costs may only be measurable) when changes in talker are kept in the attentional focus.

More generally, in considering the utility of multi-talker processing costs as a tool for studying talker normalization, it is worth noting that some researchers have suggested that multi-talker 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, & Perrachione, 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 multi-talker processing costs. However, our results appear inconsistent with this notion. The streaming account should predict multi-talker 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 Experiment 2 and 4. Such a result suggests that multi-talker processing costs indeed reflect a process of talker normalization/accommodation, rather than emerging simply because of disruptions in auditory object formation.

Conclusions

In closing, 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, W. (1975). Intrinsic and extrinsic factors in vowel judgments. In *Auditory Analysis and Perception of Speech* (pp. 103-113.).

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.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. Journal of Statistical Software, 67(1), 1-48. doi: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.

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

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, and Psychophysics*, *80*(3), 784–797.

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, 1–7.

Choi, J. Y., & Perrachione, T. K. (2019b). Time and information in perceptual adaptation to speech. *Cognition*, *192*, 103982.

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.

Francis, A. L., & Nusbaum, H. C. (1996). Paying attention to speaking rate. In *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*, 1–9.

Joos, M. (1948). Acoustic phonetics. *Language*, *24*(2), 5–136.

Kapadia, A. M., & Perrachione, T. K. (2020). Selecting among competing models of talker adaptation: Attention, cognition, and memory in speech processing efficiency. *Cognition*, *204*.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, *29*(1), 98–104.

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*, 1–9.

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.

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.

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

Phonetic Sciences.

Lo, S., & Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*(August), 1–16.

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

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.

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

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.

Nearey, T. M. (1989). Static, dynamic, and relational properties in vowel perception. *Journal of the Acoustical Society of America*, *85*(5), 2088–2113.

Nusbaum, H. C., & Magnuson, J. S. (1997). Talker normalization: Phonetic constancy as a cognitive process. *Talker Variability and Speech Processing*, 109–132.

Nusbaum, H. C., & Morin, T. M. (1992). Paying attention to differences among talkers. In *Speech Perception, Production and Linguistic Structure* (pp. 133–134).

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* (1st ed., pp. 113–157). Academic Press.

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.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: 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.

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.

Singmann, H., Bolker, B., Westfall, J., Aust, F., & Ben-Shachar, M. S. (2020). afex: Analysis of Factorial Experiments. R package version 0.27-2. https://CRAN.R-project.org/package=afex

Sjerps, M. J., Fox, N. P., Johnson, K. A., & Chang, E. F. (2018). Speaker-normalized vowel representations in the human auditory cortex. *Nature Communications*, (2019), 1–38.

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.

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.

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), 198–212.

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). The neural basis of talker normalization. *Journal of Cognitive Neuroscience*, *16*, 1173–1184.

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.