Listener expectations and the perceptual accommodation of talker variability: A pre-registered replication

Sahil Luthra[1], David Saltzman[1], Emily B. Myers[1,2], & James S. Magnuson[1,3,4]

[1] University of Connecticut, Department of Psychological Sciences

[2] University of Connecticut, Department of Speech, Language and Hearing Sciences

[3] BCBL. Basque Center on Cognition Brain and Language, Donostia-San Sebastián, Spain

[4] Ikerbasque. Basque Foundation for Science, Bilbao, Spain

Author Note

This research was supported by NSF 1754284, NSF IGERT 1144399 & NSF NRT 1747486 (PI: JSM) and NSF BCS 1554810 & NIH R01 DC013064 (PI: EBM). This research was also supported in part by the Basque Government through the BERC 2018-2021 program and by the Agencia Estatal de Investigación through BCBL Severo Ochoa excellence accreditation SEV-2015-0490. SL was supported by an NSF Graduate Research Fellowship. The authors have no known conflict of interest to disclose. All stimuli, data and analysis code are publicly available at https://github.com/disaltzman/TalkerTeam-Expectations. The experimental approach was approved by this journal prior to data collection.

CORRESPONDING AUTHOR:

Sahil Luthra

Psychological Sciences

University of Connecticut

Storrs, CT 06269-1020

sahil.luthra@uconn.edu

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. https://doi.org/10.3758/s13414-021-02317-x

Abstract

Researchers have hypothesized that in order to accommodate variability in how talkers produce their speech sounds, listeners must perform a process of *talker normalization*. Consistent with this proposal, several studies have shown that spoken word recognition is slowed when speech is produced by multiple talkers compared to when all speech is produced by one talker (a *multi-talker processing cost*). Nusbaum and colleagues have argued that talker normalization is modulated by attention (e.g., Nusbaum & Morin, 1992). Some of the strongest evidence for this claim is from a speeded monitoring study where a group of participants who expected to hear two talkers showed a multi-talker processing cost but a separate group who expected one talker did not (Magnuson & Nusbaum, 2007). In that study, however, the sample size was small and the crucial interaction was not significant. In this registered report, we present the results of a well-powered attempt to replicate those findings. In contrast to the previous study, we did not observe multi-talker processing costs in either of our groups. To rule out the possibility that the null result was due to task constraints, we conducted a second experiment using a speeded classification task. As in Experiment 1, we found no influence of expectations on talker normalization, with no multi-talker processing cost observed in either group. Our data suggest that the previous findings of Magnuson and Nusbaum (2007) be regarded with skepticism and that talker normalization may not be permeable to high-level expectations.

Introduction

Listeners typically recognize a talker's intended message with ease, a noteworthy feat given the amount of variability in how individual talkers produce speech sounds (Peterson & Barney, 1952). In general, studies of talker variability suggest that phonetic information is not processed independently from voice information; listeners show better identification of words spoken by familiar talkers (Nygaard, Sommers, & Pisoni, 1994) and listeners are slower to classify word-initial phonemes when there is variation in talker identity compared to when talker identity is held constant (Mullennix & Pisoni, 1990). One proposed interpretation of these findings is that speech perception must involve an extrinsic normalization process by which listeners adjust how they map from the acoustic signal to abstract perceptual categories, tuning to the characteristics of the talker as they hear them speak (Joos, 1948; Ladefoged & Broadbent, 1957; Nusbaum & Morin, 1992).

Extrinsic normalization accounts predict that when talkers are intermixed, listeners incur a processing cost each time they encounter a new talker, due to the need to compute the mapping between the talker's productions and perceptual categories. Indeed, performance advantages (faster and/or more accurate responses) are seen in speech processing tasks when items are blocked by talker as compared to when items from two or more talkers are intermixed (e.g., Mullennix, Pisoni, & Martin, 1989; Nusbaum & Morin, 1992). Such performance advantages have been observed across several paradigms, including in perceptual identification and word naming tasks (Mullennix et al., 1989) and in speeded classification (e.g., *is this a* /b/ *or* /p/?; Choi, Hu, & Perrachione, 2018).

In several word monitoring experiments by Nusbaum and Morin (1992), participants heard a series of syllables or words on each trial and were instructed to press a button whenever they heard a visually cued target stimulus (e.g., *press the key whenever you hear BALL*). In Blocked trials, all targets and distractors were produced by a single talker. In Mixed trials, targets and

distractors from both talkers were randomly interleaved. In general, participants were slower for Mixed than Blocked trials, which we refer to as a *multi-talker processing cost*. Nusbaum and Morin hypothesized that talker changes trigger an extrinsic normalization process that requires cognitive resources. In their Experiment 4, the multi-talker processing cost was not found with two female talkers with similar vowel spaces. This suggests the possibility that either listeners did not detect talker changes – since the talkers were similar, the same mapping from acoustics to perceptual categories worked for both – or that there could be a normalization mechanism that is triggered only when the errors are detected in the listener's current mapping from acoustics to phonetic categories. Finally, in their Experiment 5, Nusbaum and Morin (1992) found an interaction with cognitive load: The multi-talker processing cost was larger when participants had to maintain three numbers in working memory than when they only had to remember one, providing further evidence that normalization requires cognitive resources.

Nusbaum and colleagues have argued that these findings suggest that mapping speech to phonetic categories is an attentionally-demanding process carried out via an active control system: The speech perception system may need to re-compute a mapping from speech to phonetic categories when a talker change is perceived *or* if the current mapping leads to errors, such as failures to find lexical matches (Magnuson & Nusbaum, 2007; Nusbaum & Magnuson, 1997; Nusbaum & Morin, 1992). More recently, Choi et al. (2018) found that multi-talker processing costs were most pronounced when there was high ambiguity between potential target sounds, supporting the notion that talker normalization is an active process in which cognitive resources are used to resolve ambiguities in the speech signal. In that same study, Choi et al. also found that multi-talker processing costs were observed even when there was no ambiguity between potential alternatives (that is, when the same mapping between acoustics and percepts could be used for

both talkers), in line with the suggestion that talker normalization may occur whenever a change in talker is detected, not only in cases where normalization is necessary to avoid phonetic ambiguity.

Some of the strongest evidence that the accommodation of talker variability relies on an attention-modulated talker normalization process comes from a word monitoring experiment by Magnuson and Nusbaum (2007; Experiment 4). In this experiment, listeners heard synthetic speech, with some words produced at an average fundamental frequency (F0) of 150 Hz and some words produced with an average F0 of 160 Hz. Critically, one group of subjects was told that the synthetic speech was intended to simulate one talker with variable pitch, while another group was told that the speech simulated two similar-sounding talkers. (A third group was given no instructions about the number of talkers). The authors found a multi-talker processing cost when listeners thought they were hearing two talkers, but not when they believed they were hearing a variable single talker (or had no expectations about number of talkers), suggesting that phonetic processing can be modulated in a top-down fashion by the *expectation* that the listener would hear two voices rather than one. In other words, this finding suggests that when a listener encounters a small (but noticeable) acoustic difference, their expectations govern whether they treat it as within-talker variation or whether they re-compute the mapping between acoustics and percepts.

While this is an intriguing finding with potentially important implications for theories of speech perception, there are reasons to be cautious about the robustness of expectation effects in talker normalization. First, the critical interaction between Expectations (1-voice vs. 2-voice) and Blocking (Mixed vs. Blocked) – was not significant in the original data, $F(1,14) = 1.797$, partial

$\eta^2 = 0.114$, Cohen's $f = 0.358$, $p = 0.201$[1]. Furthermore, the study was underpowered; with only 8 participants per group, a post-hoc power analysis of the original data revealed the power to detect the interaction to be 0.27. Conversely, there are reasons to suspect that the original finding is not spurious, as other studies have shown that a listener's expectations about talker identity can influence phonetic processing. In a study by Fenn et al. (2011), for instance, listeners were more likely to notice a change in voice on a telephone call when they were actively monitoring for one, consistent with the notion that phonetic processing can be guided by expectations.

If the results presented by Magnuson and Nusbaum (2007) are indeed robust, they would provide strong support for theoretical perspectives in which talker normalization occurs via an active control process that requires cognitive resources and cognitive control. Such a perspective contrasts with the radically different view that talker variability can be accommodated non-analytically, as in episodic accounts (Goldinger, 1998). Episodic theories posit that word recognition relies on episodic traces that include all aspects of a spoken utterance, including both phonetic and non-linguistic information. Under this view, listeners do not need to actively re-compute the mapping between acoustics and phonemes as talkers vary; rather, they simply encode holistic memories of each speech token, including both linguistic content and non-linguistic information about talker identity. Tokens are then recognized by how they cluster with prior episodic memories. Such a perspective does not account for the evidence that mixed-talker effects are most pronounced when listeners have a working memory load (Nusbaum & Morin, 1992) but can readily account for talker specificity effects in recognition memory, whereby listeners have stronger recognition memory for words that are produced by the same talker between encoding

[1] Magnuson and Nusbaum (2007) reported the simple effect of Blocking (Mixed / Blocked) for each level of Expectations (1-voice / 2-voice), but the interaction was not reported. The interaction statistics we report here were calculated from the original data.

and test phases than for words that are produced by different talkers between encoding and test (e.g., Palmeri, Goldinger, & Pisoni, 1993).

In Experiment 1 of the current study, we conducted a well-powered, pre-registered experiment in an attempt to directly replicate the Magnuson and Nusbaum (2007) finding that expectations can modulate the emergence of multi-talker processing costs. Notably, the influence of expectations on multi-talker processing costs has only been shown in a word monitoring paradigm, but there are inherent asymmetries in the response demands of Mixed and Blocked trials in the standard monitoring paradigm (described below in the Discussion) that make it difficult to assess whether speeded monitoring studies are well-suited to investigating talker normalization (Saltzman, Luthra, Myers, & Magnuson, under review). Therefore, in Experiment 2, we examined whether the effect would emerge in a speeded classification task, which has also been used for studying talker normalization (Choi et al., 2018).

Experiment 1

Methods

Stimuli. We recreated the 19 phonetically-balanced monosyllabic words used by Magnuson and Nusbaum (2007) with the DECtalk synthesizer, with a mean F0 of 150 Hz, and also created pitch-shifted variants (with mean F0 shifted to 160 Hz). As in Magnuson and Nusbaum, the full set of stimuli consisted of the words *ball*, *bluff*, *cad*, *cave*, *cling*, *depth*, *done*, *dime*, *gnash*, *greet*, *jaw*, *jolt*, *knife*, *lash*, *reek*, *romp*, *park*, *priest* and *tile*.

We also synthesized a monologue (where pitch variation was not intended to cue a talker change) and a dialogue (where pitch variation corresponded to a change in talker). These were used to build listener expectations for one or two talkers, as described below.

Participants. Subjects were recruited from the University of Connecticut community and completed the experiment in the lab. All individuals were at least 18 years of age and self-identified as monolingual native speakers of American English with no history of neurological, speech, hearing or language deficits. All procedures were approved by the University of Connecticut Institutional Review Board, and subjects provided informed consent prior to participating.

Given that accuracy was very high in the Magnuson and Nusbaum (2007) experiments, we decided *a priori* to exclude participants from analyses if they had accuracy levels below 90 percent. We collected data until we had a sample of 88 participants who met this criterion (44 participants in each group); a total of 5 participants were excluded from analyses for failing to meet the 90% criterion. A power analysis indicated that with the effect size estimated from the original dataset from Magnuson and Nusbaum (2007), our sample size would exceed power of 0.90 at an α of 0.05.

Procedure. Participants were assigned to two groups. One group listened to a monologue (for the *one-voice expectation* group) and the other to a dialogue (for the *two-voice expectation* group); the text of the monologue and dialogue are presented in Figure 1. In both the monologue and dialogue, pitch changed from sentence to sentence. Participants hearing the monologue were told that there was pitch variation for the purpose of trying to make our low-quality synthetic speech sound more natural, but all sentences were produced by one character.

Monologue	**Dialogue**
I have a ton of homework tonight. I'm not sure if I'm going to make it to practice. But if I don't make it to tonight's practice, then I won't be able to play in the game on Saturday. I don't want to miss the first game of the season, but I know that if I don't do my Spanish project, I may not get a passing grade on my progress report. Why did I wait until the last minute to do the project? I knew that I'd be benched for the rest of the season if I got a failing grade. Well, I guess I'll just have to miss practice to get the project done and wait until next week's game to play. And I should really try harder to get my grades up. My team needs me on the field.	**Bill:** Joe, I have a ton of homework tonight. I'm not sure if I'm going to make it to practice. **Joe:** But Bill, if you don't make it to tonight's practice, then you won't be able to play in the game on Saturday. **Bill:** I don't want to miss the first game of the season, Joe, but I know that if I don't do my Spanish project, I may not get a passing grade on my progress report. **Joe:** Bill, why did you wait until the last minute to do the project? You knew that you'd be benched for the rest of the season if you got a failing grade. **Bill:** Well, Joe, I guess I'll just have to miss practice to get the project done and wait until next week's game to play. **Joe:** Yeah, Bill, and you should really try harder to get your grades up. Your team needs you on the field.

Figure 1. *A monologue and dialogue were used to establish listeners' expectations that they would hear one or two talkers, respectively.*

Following this, participants completed the word monitoring task. On every trial, participants were shown a written target word (either *ball*, *tile*, *cave* or *done*, depending on the trial) and then heard a sequence of 16 words, with the target appearing four times in the sequence. Target items could not appear in the first or final positions of the sequence, and a target could not appear in two consecutive positions. Non-target items were selected (with replacement) from the remaining stimulus words; note that, consistent with previous studies, an item that served as a target item on one trial could therefore serve as a non-target item on other trials.

Participants were instructed to press the spacebar as soon as they heard the target word, and instructions emphasized speed. Every subject received two types of trial: In *blocked* trials, participants only heard words produced by one of the two talkers (i.e., with a constant mean pitch

within the block); in *mixed* trials, words were produced by both talkers (i.e., with varying mean pitch within the block). There were 48 trials of each type, and we followed the randomization procedures described by Magnuson and Nusbaum (2007). The experiment was programmed in OpenSesame (Mathôt, Schreij, & Theeuwes, 2012).

Results

Overall results are summarized in Table 1.

Table 1
Accuracy and response time data from Experiment 1 (Speeded monitoring)

Expectations	**Accuracy (%)** Mean (SD)	**Response Time (ms)**	
		Blocked Trials Mean (SD)	*Mixed Trials* Mean (SD)
One Voice Instructions	97.9 (1.5)	484 (96)	482 (100)
Two Voice Instructions	97.7 (2.1)	483 (95)	482 (94)

All analyses were conducted in R (R Core Team, 2019). Models were implemented using the *mixed* function in the "afex" package (Singmann, Bolker, Westfall, & Aust, 2018); this function interfaces with the *lmer* function of the "lme4" package (Bates, Maechler, Bolker, & Walker, 2015) but provides results in an ANOVA-like format, using chi-square tests to evaluate the significance of each fixed effect.

In general, accuracy in the monitoring task was high (mean: 97.8%, SD: 1.8%). Accuracy data were analyzed using a logit mixed effects model with two fixed effects (Expectations x Blocking). We used a backward stepping procedure to identify the most parsimonious random effects structure, an approach that is useful for maximizing power and reducing Type I error (Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). Such a procedure involves comparing each potential random effects structure to a simpler structure and favoring the simpler structure if there

is no significant loss of model fit to the data. This procedure identified a random effects structure that consisted of random by-subject intercepts and random slopes for Blocking; note that this constitutes the maximal random effect structure. There were no significant effects of Expectations or Blocking on accuracy (Expectations, $\chi^2(1) = 0.09$, $p = 0.77$; Blocking, $\chi^2(1) = 0.26$, $p = 0.61$; Expectations x Blocking, $\chi^2(1) = 0.18$, $p = 0.67$).

Response times were measured from word onset. Following Magnuson and Nusbaum (2007), responses that occurred less than 150 ms after stimulus onset were counted as responses to the previous word. Response time data (Figure 2) were analyzed with a generalized linear mixed effects model using the same stepping procedure as above to identify the appropriate random effects structure. Following the recommendation of Lo and Andrews (2015), we employed a model with an identity link function, and a chi-square test was used to determine whether specifying a gamma or inverse Gaussian distribution in the model would allow for a better approximation of the response time distribution; both models fit the data equally well, so we opted to use a gamma distribution. The critical interaction between Expectations and Blocking was not significant, $\chi^2(1) = 0.00$, $p = 0.97$. Neither the main effect of Expectations, $\chi^2(1) = 0.00$, $p = 0.98$, nor the main effect of Blocking, $\chi^2(1) = 1.24$, $p = 0.27$, was significant.

For direct comparison with Magnuson and Nusbaum (2007), we also analyzed our data using a two-way Expectations x Blocking ANOVA. We observed the same pattern of results as with the mixed effects model, with a non-significant effect of Expectations, $F(1, 172) = 0.00$, $p = 0.99$, a non-significant effect of Blocking, $F(1, 172) = 0.00$, $p = 0.97$, and a non-significant interaction between Expectations and Blocking, $F(1, 172) = 0.01$, $p = 0.95$.

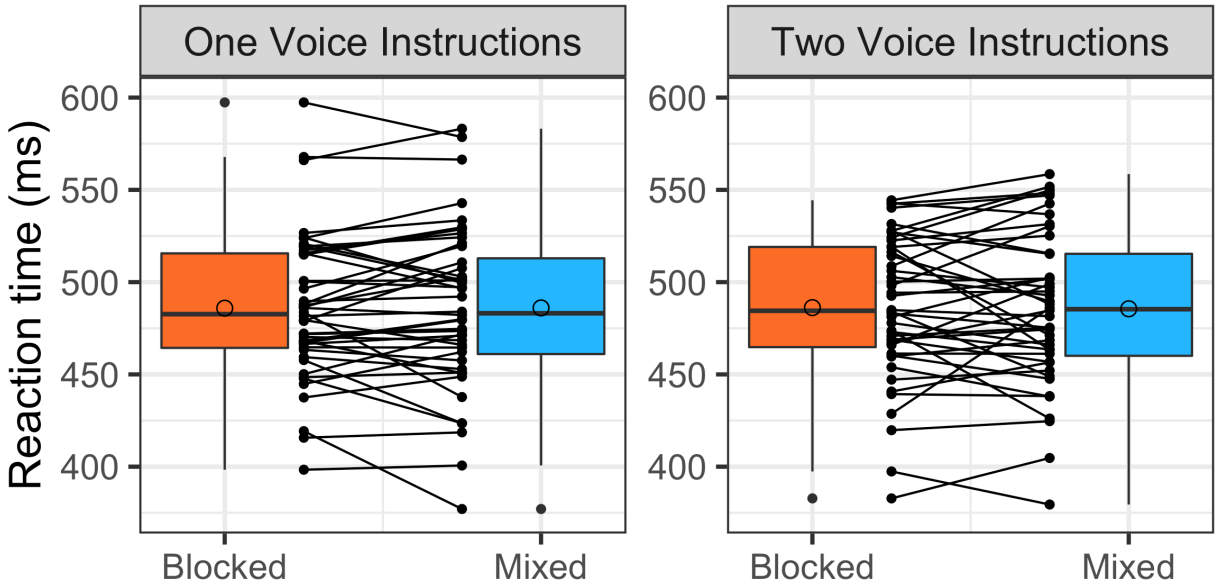

Figure 2. *Results of Experiment 1. Box-and-whisker plots indicate the distribution of data in each group. In these plots, the median is represented by a horizontal line in the box, and the mean as a circle. The box height is defined by the first and third quartiles, and the whiskers extend to the minimum and maximum values that are no more than 1.5 times the distance between the first and third quartiles. Horizontal / diagonal line segments show the effect of Blocking for each subject.*

Discussion

In Experiment 1, we attempted a well-powered, pre-registered replication of Experiment 4 from Magnuson and Nusbaum (2007). We did not find evidence that the emergence of a multi-talker processing cost could be modulated by listeners' expectations about whether they would hear one voice or two, suggesting that the previous finding was likely spurious. Alternatively, the slight changes in materials and audio equipment could have disrupted the effect, although this would suggest that, at best, the effect is quite fragile.

While Magnuson and Nusbaum (2007) used a speeded monitoring paradigm to test for effects of expectations in accommodating talker variability, recent work suggests that the processing costs in this paradigm may not reflect talker normalization *per se*. In particular, the standard monitoring paradigm requires subjects to monitor for one unique token during Blocked

trials (e.g., “ball” produced by a male talker) but to monitor for two tokens during Mixed trials (“ball” produced by a male talker and by a female talker). Thus, Blocked and Mixed trials differ not only in the presence or absence of talker variability but also in the number of unique tokens to which participants must respond. Saltzman et al. (under review) found that when the number of target tokens was equated between Blocked and Mixed trials – specifically, when Mixed trials contained word-to-word changes in talker but all the targets in a given trial were produced by one talker – no multi-talker processing cost was elicited.

The implications of this outcome are complex. Having to monitor for two different tokens was an explicit feature of the design, intended to increase attentional demands in order to detectably disrupt the (theorized-to-be) attention-demanding process of talker normalization. It is possible that the asymmetric attentional demands in blocked- vs. mixed-talker conditions may themselves substantially drive multi-talker processing costs in this paradigm. But even so, two physically distinct target tokens should increase attentional demands and slow processing.

Given these uncertainties, we conducted a second experiment using a speeded word identification paradigm instead. Of interest is whether expectations can influence the size of multi-talker processing costs elicited in this task.

Experiment 2

In Experiment 2, we investigated whether expectations could modulate the emergence of multi-talker processing costs in a speeded classification paradigm, wherein participants had to decide whether they heard the word “buy” or “pie” on every trial. Speeded classification paradigms have been increasingly used to study talker normalization (e.g., Choi & Perrachione, 2019a,

2019b), with one recent study observing multi-talker processing costs even when identifying the talker was not strictly necessary for resolving word identity (Choi et al., 2018).

Methods

Stimuli. The words "buy" and "pie" were synthesized following the same procedures as in Experiment 1. We also used the same monologue and dialogue stimuli that had been used in the previous experiment.

Participants. Subjects were recruited from the University of Connecticut community and met the same eligibility requirements as in Experiment 1. As before, all procedures were approved by the University of Connecticut Institutional Review Board, and subjects provided informed consent prior to participating.

Because of limitations on in-person data collection related to the COVID-19 pandemic, participants completed the experiment remotely from their personal computers. Of note, recent empirical data suggest that despite variability in the particular devices participants may use, online platforms offer reasonably precise measurements of response times (Anwyl-Irvine, Dalmaijer, Hodges, & Evershed, 2020), and a recent study found no difference between in-lab and online reaction time effects for several cognitive tasks (Miller, Schmidt, Kirschbaum, & Enge, 2018). Further, response time data collected online do not tend to be more variable than data collected in person; rather, to the extent that there are differences between the two environments, they tend to manifest as shifts in the distribution of overall response times (de Leeuw & Motz, 2016). Thus, our goal of measuring multi-talker processing costs should not be compromised by the fact that Experiment 2 was conducted online; multi-talker processing costs reflect a difference between (within-subject) conditions, so even if an online participant might be slower overall, the size of the

multi-talker processing cost should not be affected by the online environment. Finally, even if the amount of variability in response times differs across environments, results of a simulation study suggest that the impact on statistical power should be negligible (Brand & Bradley, 2012). For these reasons, we opted to proceed with the pre-registered sample size for Experiment 2 rather than recruiting a larger sample.

As a consequence of conducting Experiment 2 online, the experimenters were not able to see whether participants were using headphones, as would have been possible in the lab. We therefore also required participants to pass a psychophysical headphone screening (Woods, Siegel, Traer, & McDermott, 2017) for their data to be included in analyses. In this task, pure tones are played in stereo, and listeners are asked to indicate on every trial which of three tones is quietest; because of phase cancellation in the stimuli, this task can be used to assess whether participants are listening to the stimuli over loudspeakers or via headphones. If participants failed the headphone screening once, they were reminded of the importance of wearing headphones and given a second opportunity to pass the screening.

We recruited 121 participants for Experiment 2. Thirty were excluded for failing the headphone screening twice, and an additional three were excluded for failing to meet the 90% accuracy criterion used in Experiment 1. This resulted in a sample of 88 participants, with 44 participants in each group.

Procedure. We adapted the paradigm from Choi et al. (2018). During an initial exposure phase, participants heard either the monologue or dialogue to establish expectations. Subsequently, each subject listened to two single-talker blocks and two mixed-talker blocks while completing a speeded word classification task. On every trial of the classification task, participants heard either the word “buy” or “pie” and indicated which they heard by making a button response as quickly

as possible. The classification task consisted of 4 blocks of 40 trials with a 2-second SOA, and the same item could not occur more than three times in a row. In a single-talker block, all the trials were produced by one talker (i.e., at one mean F0), with every participant exposed to both talkers across blocks (with order counterbalanced). In each mixed-talker block, 20 productions from each talker (i.e., each mean F0) were intermixed throughout the block, for a total of 40 trials.

After completing the experiment, participants completed a questionnaire in which they were asked the following questions: (1) Did you notice anything unusual about the experiment? (2) How many talkers did you notice during the entire experiment? (3) On a scale from 1-10, 10 being the most confident, what is your level of confidence in your answer to Question 2? While such a questionnaire was not administered in the original experiment by Magnuson and Nusbaum (2007), we implemented one here in order to verify the effectiveness of the monologue/dialogue manipulation. (Note that we had intended for this questionnaire to be included in Experiment 1 as well, as specified in our pre-registered methods, but the questionnaire was not included there due to a programming error.) Experiment 2 was programmed using the online experiment software Gorilla (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020).

Results

Overall results are summarized in Table 2.

Table 2

Accuracy and response time data from Experiment 2 (Speeded classification)

Expectations	**Accuracy (%)** Mean (SD)	**Response Time (ms)**	
		Blocked Trials Mean (SD)	*Mixed Trials* Mean (SD)
One Voice Instructions	97.9 (14.4)	743 (232)	736 (226)
Two Voice Instructions	97.9 (14.5)	709 (223)	714 (229)

In general, task accuracy was high (mean: 97.9%, SD: 14.5%) and comparable to the performance observed in Experiment 1. Accuracy data were analyzed following the same procedure as in Experiment 1. Specifically, we used a logit mixed effects model with fixed effects of Expectations and Blocking, random by-subject slopes for Blocking, and random by-subject intercepts. This represents the maximal random effects structure and provided a marginally better fit than a model without random slopes ($p = 0.07$). As in Experiment 1, there were no significant effects of Expectations or Blocking on accuracy (Expectations, $\chi^2(1) = 0.03$, $p = 0.87$; Blocking, $\chi^2(1) = 0.12$, $p = 0.73$; Expectations x Blocking, $\chi^2(1) = 0.00$, $p = 0.95$).

Response time data for correct responses (Figure 3) were analyzed using a generalized linear mixed effects model; this model considered fixed factors of Expectations and Blocking, and as in Experiment 1, we used a backward stepping procedure to identify the optimal random effects structure. This procedure led to adopting a model with random by-subject slopes for Blocking and random intercepts for each subject (i.e., the maximal random effects structure). As above, we used an identity link function and specified a gamma distribution in our model. Results indicated no significant effects of Expectations or Blocking (Expectations, $\chi^2(1) = 0.68$, $p = 0.41$; Blocking, $\chi^2(1) = 0.00$, $p = 0.96$; Expectations x Blocking, $\chi^2(1) = 0.27$, $p = 0.61$).

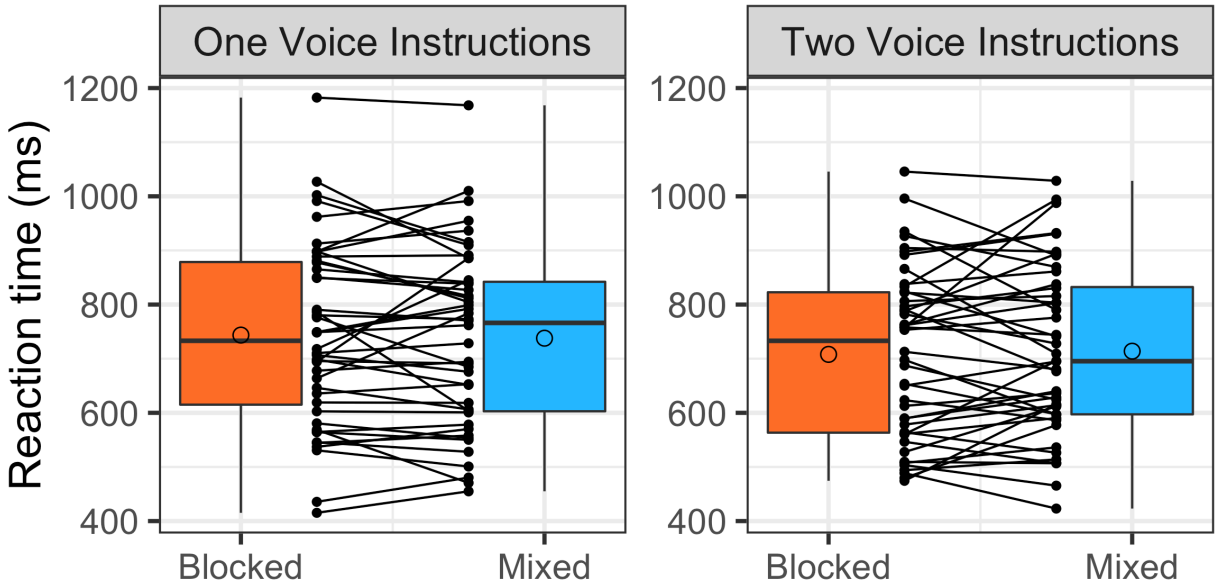

Figure 3. *Results of Experiment 2. Box-and-whisker plots indicate the distribution of data in each group. In these plots, the median is represented by a horizontal line in the box, and the mean as a circle. The box height is defined by the first and third quartiles, and the whiskers extend to the minimum and maximum values that are no more than 1.5 times the distance between the first and third quartiles. Horizontal / diagonal line segments show the effect of Blocking for each subject.*

Finally, we assessed the effectiveness of the expectations manipulation through a post-task questionnaire. Most participants indicated that they did not notice anything unusual about the experiment, and no participants correctly guessed the purpose of the experiment. We also asked participants to indicate how many talkers they heard in the entire experiment and their confidence in each. Results for these two questions are provided in Table 3. In examining these data, it is striking that the distribution of responses did not differ dramatically between groups – regardless of the audio heard during the instructions phase, approximately half the participants said they heard one talker, approximately one-quarter said they heard two, and approximately one-quarter provided some other response (e.g., 3 or 4). However, listeners who heard the dialogue were significantly less confident that they had only heard one talker, as assessed by a two-tailed *t* test, $t(42) = 2.52$, $p = 0.016$.

	How many talkers did you notice during the entire experiment?			**On a scale from 1-10, 10 being the most confident, what is your level of confidence in your response?**		
	Number of subjects who said "one"	*Number of subjects who said "two"*	*Number of subjects with other response*	*Average response for subjects who said "one"*	*Average response for subjects who said "two"*	*Average response for subjects with other response*
One Voice Instructions	24	12	8	7.13	5.58	4.00
Two Voice Instructions	20	13	11	5.15	5.77	5.82

Table 3. *Answers to debriefing questions in Experiment 2.*

Discussion

In Experiment 2, we examined whether expectations could modulate the emergence of multi-talker processing costs in a speeded word classification task. Following Magnuson and Nusbaum (2007), we attempted to manipulate expectations by presenting participants with either a monologue or a dialogue during the instructions phase, cueing them to interpret a 10 Hz change in mean F0 as either within-talker variability or a change in talker. As was also the case in Experiment 1, we did not observe multi-talker processing costs in either group. Results from a post-task questionnaire suggest that the expectations manipulation was only partially effective, as a similar number of participants in each group reported hearing only one talker throughout the experiment; nonetheless, those listeners who heard the dialogue and reported hearing only one talker were significantly less confident in their response than listeners who heard the monologue and reported hearing only one talker.

General Discussion

In a previous study, Magnuson and Nusbaum (2007) found that listeners' expectations about whether they would hear one or two talkers modulated the emergence of multi-talker processing costs in a speeded monitoring task. However, their experiment was underpowered and the result may have been spurious. In this registered report, we attempted to replicate their findings, conducting a well-powered experiment that closely followed their methodology (Experiment 1). However, we did not observe multi-talker processing costs in our sample. We also conducted a second experiment using a speeded classification task (Experiment 2) and similarly found no evidence that expectations could modulate the emergence of multi-talker processing costs.

One possibility is that our null results may have been driven by methodological details. To be consistent with the previous experiment by Magnuson and Nusbaum (2007), the current work used synthetic stimuli generated by the DECtalk synthesizer. The earlier study used DECtalk materials originally developed in the 1990s as a convenient way to alter pitch without modifying other talker characteristics, and listeners in the earlier study may have been more accustomed to lower-quality speech synthesis. Contemporary listeners may have engaged differently with this low-quality speech, and as a result, the monologue and dialogue may not have sufficiently shaped participants' expectations for how many voices they would hear. Alternatively, subtle differences in our (recreated) stimuli or audio equipment might have diluted the impact of the monologue and dialogue contexts. Results from the post-task questionnaire in Experiment 2 seem to support such a possibility, as a comparable number of listeners in each of the two groups reported having heard only one talker throughout the experiment. Furthermore, the fact that we did not observe multi-talker processing costs in general (i.e., there was no effect of Blocking) in either experiment suggests that listeners may have had trouble mapping the 10 Hz difference in fundamental

frequency onto a talker difference. In future work, it would be informative to test whether effects of expectations can be seen when more naturalistic stimuli are used instead. However, we might expect talker normalization effects to be even *less* pronounced with natural speech, as previous work on speaking rate normalization has shown stronger effects with synthetic speech than with more natural stimuli (Toscano & McMurray, 2012).

Alternatively, it may simply be the case that expectations cannot modulate the emergence of multi-talker processing costs. If that is the case, then it becomes necessary to rethink the claim that talker normalization can be modulated by expectations (Magnuson & Nusbaum, 2007). Instead, normalization may proceed in a fairly automatic fashion, with listeners passively estimating and adjusting for differences in vocal tract length between talkers – and by and large, this is how talker normalization has been characterized in the literature (Joos, 1948; Ladefoged & Broadbent, 1957; Nearey, 1989; Weatherholtz & Jaeger, 2016). However, it is worth underscoring that phonetic differences across talkers are driven not only by the physics of the vocal tract (a formula adjusting for vocal tract size could be applied fairly reflexively), but also by layers of talker identity such as sexual orientation, gender, and regional dialect (Munson, 2007; Munson, McDonald, DeBoe, & White, 2006: Johnson, Strand, & D'Imperio, 1999; Labov, Ash, & Boberg, 2006). For instance, prepubescent boys and girls, do not differ in vocal tract length, yet they approximate the differences in the formant structure of adult men and women (Johnson, 2008). Previous work has also shown that listeners interpret vowels differently depending on the visually perceived gender of a talker and even depending on their expectations of whether they will hear a male or female talker (Johnson, Strand, & D'Imperio, 1999). This latter piece of evidence suggests that to the extent that listeners use sociophonetic information, talker normalization processes are penetrable at some level to effects of expectation.

Even more compellingly, as we reviewed earlier, Nusbaum and Morin's (1992) finding that talker variability interacts with cognitive load implies a resource- and attention-demanding process. Taken together, these findings constitute evidence that listeners are not simply normalizing speech on the basis of information that is recovered automatically from the speech signal. Rather, a listener's interpretation of the speech signal must also be shaped by their inferences about the talker they are hearing.

A more radical alternative is that normalization may not be needed at all to accommodate talker variability. In episodic theories, recognition of spoken words is achieved through resonance between the incoming signal and acoustically rich, detailed speech episodes maintained in memory (Goldinger, 1998; Pufahl & Samuel, 2014). In a seminal study, Goldinger (1998) described how such a theory can account for listener tendencies to spontaneously imitate their conversational partner. Because recent episodes are activated more strongly, a listener will have strongly activated episodic traces for words that their conversational partner has just produced, and through the coupling of the perceptual and production systems, a listener's speech output will often resemble that of their interlocutor. Imitation is particularly marked for low-frequency words, which are associated with relatively few episodic traces; by contrast, when a listener activates a high-frequency word, recognition is determined by resonance with a large number of traces, and the aggregate may not strongly resemble any particular individual episode.

On the basis of their finding that expectations could modulate the emergence of multi-talker processing costs, Magnuson and Nusbaum (2007) argued against a particular form of the episodic theory in which the episodes maintained in memory are unanalyzed, unparameterized auditory objects. Analysis of the signal, the authors argued, is critical for speech perception; in order for expectations to modulate the size of multi-talker processing costs, listeners need to be

able to decompose the speech signal into particular auditory dimensions and to change how much they attend to certain dimensions (depending on whether they expect to hear one voice or two). Though we were not able to replicate this key finding from Magnuson and Nusbaum, we believe there are still a number of reasons to disfavor nonanalytic episodic models. First, Magnuson, Nusbaum, Akahane-Yamada, and Saltzman (2021) found a talker-change cost even for talkers with whom a listener has extensive experience (family members). As they argue in detail, this is consistent with a parallel-contingent relation (Turvey, 1973) between voice characteristics and phonetic identification; a talker's vocal characteristics condition phonetic realization (and vice-versa, in many cases; Remez, Fellowes, & Rubin, 1997). Following a talker change, it seems that listeners must hear enough speech to detect that the talker is familiar before they can exploit past experience to facilitate speech perception. As Magnuson et al. (2021) discuss, it is unclear how a talker-change cost would emerge from wholly non-analytic episodic theories. Second, a listener's interpretation of the speech signal can be shaped by contextual factors, such as coincident printed text (Frost, Repp, & Katz, 1988), the visible movements of the articulators (McGurk & MacDonald, 1976), and one's expectations of talker gender (Johnson et al., 1999). Furthermore, several studies indicate that listeners can quickly adapt to the idiosyncratic way that a particular talker produces their speech sounds (Kraljic & Samuel, 2005; Luthra, Mechtenberg, & Myers, 2021; Maye, Aslin, & Tanenhaus, 2008; Norris, McQueen, & Cutler, 2003; Saltzman & Myers, 2021). In order to perceive speech in such a flexible manner, listeners must be able to attend more or less to certain stimulus dimensions, which is not possible in a nonanalytic episodic model. However, we clarify that we do not object to the core tenet of episodic models – that recognition involves resonance with an aggregate of episodes – and that this represents a potential alternative to talker normalization.

In closing, the present study suggests that the previous finding by Magnuson and Nusbaum (2007) that expectations can influence talker normalization should be regarded with skepticism. However, we suggest that future work is needed to more definitively establish the degree to which talker normalization is permeable by high-level expectations.

Acknowledgments

We thank five anonymous reviewers for helpful feedback on our proposed methods and an earlier draft of our manuscript.

Open Practices Statement

All stimuli, data and analysis code are publicly available at https://github.com/disaltzman/TalkerTeam-Expectations. The experimental approach was approved by this journal prior to data collection.

References

Anwyl-Irvine, A., Dalmaijer, E. S., Hodges, N., & Evershed, J. K. (2020). Realistic precision and accuracy of online experiment platforms, web browsers, and devices. *Behavior Research Methods*. Advance online publication.

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388–407.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Brand, A., & Bradley, M. T. (2012). Assessing the effects of technical variance on the statistical outcomes of web experiments measuring response times. *Social Science Computer Review*, *30*(3), 350–357.

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.

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, *48*(1), 1–12.

Fenn, K. M., Shintel, H., Atkins, A. S., Skipper, J. I., Bond, V. C., & Nusbaum, H. C. (2011). When less is heard than meets the ear: Change deafness in a telephone conversation. *Quarterly Journal of Experimental Psychology*, *64*(7), 1442–1456.

Frost, R., Repp, B. H., & Katz, L. (1988). Can speech perception be influenced by simultaneous presentation of print? *Journal of Memory and Language*, *27*(6), 741–755.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Johnson, K. A. (2008). Speaker normalization in speech perception. *The Handbook of Speech Perception*, 363–389.

Johnson, K. A., Strand, E. A., & D'Imperio, M. (1999). Auditory-visual integration of talker gender in vowel perception. *Journal of Phonetics*, *27*, 359–384.

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

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

Kraljic, T., & Samuel, A. G. (2005). Perceptual learning for speech: Is there a return to normal? *Cognitive Psychology*, *51*(2), 141–178.

Labov, W., Ash, S., & Boberg, C. (2006). *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin, Germany: Walter de Gruyter.

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

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*, 1–16.

Luthra, S., Mechtenberg, H., & Myers, E. B. (2021). Perceptual learning of multiple talkers requires additional exposure. *Attention, Perception & Psychophysics*. Advance online publication.

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accomodation 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, & Psychophysics.* Advance online publication.

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.

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.

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The weckud wetch of the wast: Lexical adaptation to a novel accent. *Cognitive Science*, *32*(3), 543–562.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*, 746–748.

Miller, R., Schmidt, K., Kirschbaum, C., & Enge, S. (2018). Comparability, stability, and reliability of internet-based mental chronometry in domestic and laboratory settings. *Behavior Research Methods*, *50*(4), 1345–1358.

Mullennix, J. W., & Pisoni, D. B. (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*(4), 379–390.

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.

Munson, B. (2007). The acoustic correlates of perceived masculinity, perceived femininity, and perceived sexual orientation. *Language and Speech*, *50*(1), 125–142.

Munson, B., McDonald, E. C., DeBoe, N. L., & White, A. R. (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics*, *34*(2), 202–240.

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

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology*, *47*(2), 204–238.

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).

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46.

Palmeri, T. J., Goldinger, S. D., & Pisoni, D. B. (1993). Episodic Encoding of Voice Attributes and Recognition Memory for Spoken Words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(2), 309–328.

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.

Pufahl, A., & Samuel, A. G. (2014). How lexical is the lexicon? Evidence for integrated auditory memory representations. *Cognitive Psychology*, *70*, 1–30.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Remez, R. E., Fellowes, J. M., & Rubin, P. E. (1997). Talker identification based on phonetic information. *Journal of Experimental Psychology: Human Perception & Performance*, *23*, 651–666.

Saltzman, D., Luthra, S., Myers, E. B., & Magnuson, J. S. Attention, task demands, and multi-talker processing costs in speech perception. Under review.

Saltzman, D. I., & Myers, E. B. (2021). Listeners are initially flexible in updating phonetic beliefs over time. *Psychonomic Bulletin & Review*. Advance online publication.

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). afex: Analysis of Factorial Experiments. R package version 0.21-2. https://CRAN.R-project.org/package=afex.

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*(6), 1284-1301.

Turvey, M. T. (1973). On peripheral and central processes in vision: Inferences from an information-processing analysis of masking with patterned stimuli. *Psychological Review*, *80*, 1–52.

Weatherholtz, K., & Jaeger, T. F. (2016). Speech perception and generalization across talkers and accents. In *Oxford Research Encyclopedia of Linguistics*.

Woods, K. J. P., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, and Psychophysics*, *79*(7), 2064–2072.