PROCESSING COSTS IMPOSED BY TALKER VARIABILITY DO NOT SCALE WITH NUMBER OF TALKERS

Alexandra M. Kapadia and Tyler K. Perrachione

Department of Speech, Language & Hearing Sciences, Boston University, USA
tkp@bu.edu

ABSTRACT

Phonetic variability across talkers imposes additional processing costs during speech perception, often measured by performance decrements between single- and mixed-talker conditions. However, it is unclear whether greater phonetic variability (i.e., more talkers) imposes greater processing costs. Here, we measured response times in a speeded word identification task, in which we manipulated the number of talkers (1, 2, 4, 8, or 16) in each block. Word identification was slower in every mixed-talker condition compared to the single-talker condition, but the magnitude of this performance decrement was not affected by the number of talkers. Furthermore, in a condition with uniform transition probabilities between two talkers, word identification was faster when the talker was the same as the prior trial compared to when the talker switched between trials. These results are consistent with an auditory-streaming model of talker adaptation, where processing costs associated with changing talkers result from attentional reorientation.

Keywords: speech perception; phonetic variability; processing cost; talker adaptation; auditory streaming

1. INTRODUCTION

Variation in the acoustic realization of speech across talkers is a major source of phonetic variability in speech signals [8]. Listeners are nonetheless highly successful in extracting stable phonemic information from talkers' speech despite the acoustic-phonetic inconsistency across talkers. In the presence of speech from multiple talkers, the possibility of more than one interpretation of an acoustic signal imposes additional processing demands as listeners must accommodate trial-by-trial variation in order to maintain phonetic constancy [9,17]. The costs of processing speech from multiple talkers, known as the *interference effect*, are reflected in listeners' slower response times during speech processing tasks [15,17].

Current models of talker variability in speech processing suggest that processing efficiency depends on the number of possible competing interpretations of a speech signal. Foremost among these models, the *ideal adapter framework* posits that reducing the number of possible interpretations of an acoustic signal makes speech processing more efficient [10]. A prediction of this framework is that processing speech from a limited number of potential talkers (e.g., 2 or 4) should be more efficient than a larger number of talkers (e.g., 8 or 16) because the possible interpretations of the acoustic signal are more constrained.

The interference effect of processing mixed-talker speech was notably measured by Mullennix and Pisoni [15]. They investigated processing dependencies between linguistic content and talker voice contingent on the amount of variability in the stimulus set for each variable (number of different talkers and number of different words). Their results have frequently been used to assert that the interference effect of processing speech from multiple talkers does not depend on the number of talkers, a conclusion at apparent odds with predictions of the ideal adapter framework.

However, examination of the prior data [15] reveals that these results cannot be convincingly read to support the received wisdom that the interference effect of mixed talkers is constant across increasing number of talkers. In the first of their two experiments, greater interference was reported under conditions that simultaneously increased both the number of talkers and the number of word choices, making it impossible to dissociate the effects of talker- and word-variability on processing costs. While their second experiment manipulated each variable independently, mixed-talker speech resulted in null to minimal interference effects on speech processing efficiency. Given the historical prominence of this study, the minimal mixed-talker interference measured in its second experiment is surprisingly inconsistent with the large interference effects shown by both prior and subsequent studies [6,9,16-18].

No empirical evidence from other sources exists to evaluate these predictions, since the canonical interpretation of the original report [15] has led researchers not to parametrically vary the number of talkers in their experiments, opting instead to employ mixed-talker conditions with a fixed number of talkers (usually between two and ten) [6,14,23,25].

Due to (*i*) the inconsistency between the received wisdom concerning fixed costs of talker-variability on speech processing efficiency and the predictions of contemporary speech processing models, (*ii*) the surprising lack of mixed-talker interference in the data this interpretation is founded on, and (*iii*) the paucity of other empirical work addressing this question, we set out to investigate whether mixed-talker interference varies as a function of the amount of talker variability, operationalized as the number of talkers. We attempted to parsimoniously replicate the original study [15] by having listeners perform a word identification task with a single, minimal lexical pair across mixed-talker conditions in which we parametrically manipulated only the number of talkers.

2. METHODS

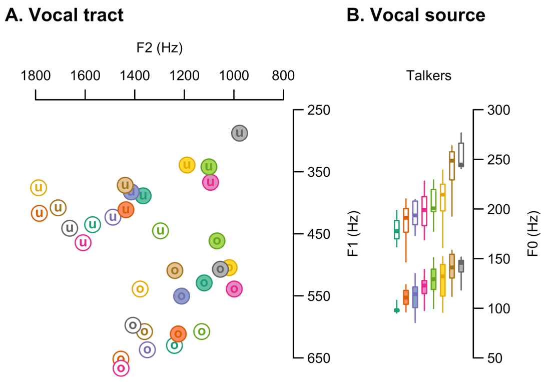
2.1. Participants

Native speakers of American English ($N = 24$; 19 female, 5 male; age 18-25, mean = 20.2 years) participated in this study. All participants had a self-reported history free from speech, language, or hearing disorders. Participants provided written informed consent, approved and overseen by the Institutional Review Board at Boston University.

2.2. Stimuli

Stimuli consisted of two naturally spoken English words, “boot” and “boat”. These words were chosen because they share the same onset and coda but differ in their vowel nucleus (/u/ vs. /o/) on a phonological contrast with a great deal of potential acoustic-phonemic ambiguity across talkers [4,8] (Fig. 1). These words were recorded by eight male and eight female native speakers of American English in a sound-attenuated booth with a Shure MX153 microphone and Roland Quad Capture sound card sampling at 44.1 kHz and 16 bits. Stimuli were RMS amplitude normalized to 65 dB SPL using Praat [2].

Figure 1: Phonetic variability across all 16 talkers for the stimuli “boot” and “boat.” **(A)** Filled (male talkers) and empty circles (female talkers) with “u” and “o” indicate the location of each talker’s vowel in F1-F2 space. **(B)** Box plots show the fundamental frequency (f_0) range for these stimuli across talkers. Colors correspond to individual talkers.



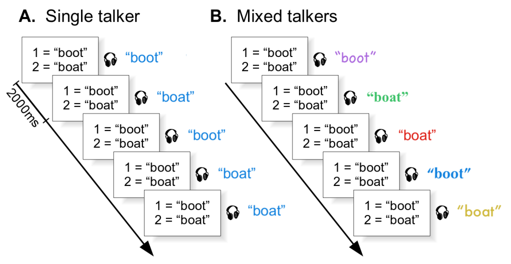
2.3. Procedure

Participants performed a speeded word identification task across six 64-trial blocks in which we parametrically varied the number of talkers (1, 2, 4, 8, or 16). Each target word was presented 32 times per block in a pseudo-random order, with the constraints that the same word not repeat more than three times in a row and that the same talker not repeat in adjacent trials during all (but one) mixed-talker conditions.

Stimulus order for n = 2 mixed talkers presented a unique challenge: if a talker could not repeat in adjacent trials, the two talkers would have to alternate predictably on each successive trial. Listeners could thus anticipate with perfect certainty which talker they would hear on the subsequent trial, potentially reducing the interfering effect of talker variability [10]. Alternatively, the talkers could be ordered randomly, in which case the same talker could occur for multiple trials in a row, reducing trial-by-trial phonetic variability [18]. Because we were uncertain how these two stimulus orders would affect speed of processing, we chose to investigate both. In the *2-talker alternating* condition, the talker switched on every trial; in the *2-talker uniform* condition, the probability of each possible talker transition was equal on every trial (e.g., after hearing A, the probability of hearing A or B was equal on every trial, and vice-versa).

Participants were instructed to listen to the stimuli and indicate which word they heard as quickly and accurately as possible by pressing a corresponding number key (Fig. 2). Written directions at the beginning of each block informed participants of the number of talkers in that block. Talkers were randomly selected for each participant, such that there were an equal number of female and male voices in each condition (the single talker block was split into a female talker half and a male talker half). Conditions were presented in a random order. Stimulus delivery was controlled using PsychoPy2 (v1.83.03) [20] with presentation via Sennheiser HD-380 Pro headphones.

Figure 2: Schematic representation of stimulus delivery. Participants performed a speeded word identification task while listening to speech produced by **(A)** a single talker or **(B)** mixed talkers. Mixed talker conditions included 2, 4, 8, or 16 talkers.



2.4. Data analysis

Response times (RTs) were measured from the onset of the target word on each trial; only RTs from correct trials were included in the analysis (accuracy was at ceiling: 96.0% ± 3.6%). RTs were analyzed in R using linear mixed-effects models implemented in the package *lme4* using maximal fixed and random effects structure [1]. Fixed factors included either number of talkers (1, 2, 4, 8, or 16), talker variability (1-talker, 2-uniform repeats, 2-uniform changes, or 2-alternating), or trial-by-trial gender variability (same gender or different gender). Random effects included by-participant slopes and intercepts and by-stimulus intercepts. Significance of effects was determined at $\alpha = 0.05$, with *p*-values for model terms based on the Satterthwaite approximation of the degrees of freedom obtained from the package *lmerTest*.

3. RESULTS

3.1. Amount of talker variability

We assessed how the amount of talker variability affected RTs (Fig. 3) using a linear mixed effects model with *number of talkers* (1, 2, 4, 8, or 16) as the fixed factor and the random factors described above. In this analysis, we collapsed the data from the two versions of the 2-talker condition. Two contrasts on the model were run separately: (*i*) for each mixed-talker level against the single-talker baseline, and (*ii*) for pairwise differences between increasing number of talkers.

RTs in all four mixed-talker conditions were significantly slower than the single-talker condition (Table 1). Increasing the number of talkers beyond the first introduction of variability (from one talker to two talkers) had no further effect on RTs (Table 2).

3.2. Talker continuity

In our design, two conditions involved continuous speech from a single talker across two or more successive trials: the single-talker condition and the 2-talker-uniform condition. While listeners could expect the talker to repeat throughout the single-talker condition, they could not have anticipated whether the talker would continue across any particular successive trials of the 2-talker-uniform condition. To this end, we explored (*i*) the effect of listener expectation by comparing RTs on these “repeat” trials (A$\underline{\text{A}}$) in the 2-talker-uniform condition with RTs in the single-talker condition, and (*ii*) the effect of talker repetition by comparing RTs on “repeat” trials (A$\underline{\text{A}}$) with “change” trials (A$\underline{\text{B}}$) in the 2-talker-unifom condition.

Finally, we compared RTs on trials from the uniform condition with unpredictable talker changes to RTs on trials from the alternating condition with predictable changes to ascertain whether ability to predict the upcoming talker expedited speech processing [10], even when the talker differed from the preceding trial.

We analyzed RTs during the 1- and 2-talker conditions in a linear mixed effects model with *talker variability* (1-talker, 2-uniform repeats, 2-uniform changes, or 2-alternating) as the fixed factor and random factors as before. Model contrasts were the pairwise differences between conditions (Fig. 4).

RTs were significantly faster for anticipated talker continuity (the single-talker condition) compared to unanticipated talker continuity (trials in the uniform version where the talker was the same as the prior trial) (Table 3). RTs in the uniform version were also significantly faster on trials where the talker did repeat compared to those where the talker did change, notwithstanding listeners’ inability to have anticipated any such repetition. However, listeners’ ability to reliably anticipate the change in talker did not affect their word recognition speed compared to trials where the talker change could not be anticipated (2-alternative vs. 2-uniform changes).

Table 1: Additional processing costs are imposed by mixed talkers vs. a single continuous talker.

Contrast	*β*	*s.e.*	*t*	*p*
2 vs. 1	0.064	0.011	5.94	5×10^{-6}
4 vs. 1	0.057	0.010	5.48	2×10^{-5}
8 vs. 1	0.064	0.008	8.28	3×10^{-8}
16 vs. 1	0.057	0.008	6.82	7×10^{-7}

Table 2: No additional processing costs are imposed by increasing amounts of talker variability.

Contrast	*β*	*s.e.*	*t*	*p*
2 vs. 1	0.064	0.011	5.94	5×10^{-6}
4 vs. 2	–0.006	0.010	–0.66	0.52
8 vs. 4	0.007	0.010	0.69	0.50
16 vs. 8	–0.008	0.006	–1.19	0.25

Figure 3: Mean RT as a function of number of talkers. Significance of pairwise contrasts are indicated above the line. Error bars indicate ± 1 SEM.

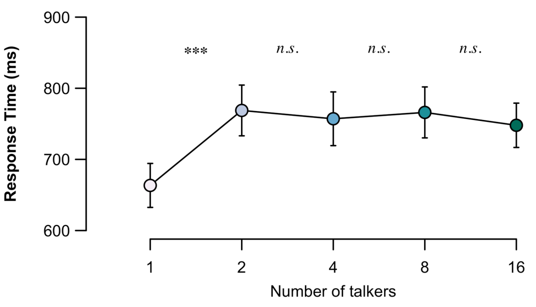


Figure 4: Mean RT as a function of talker continuity and listeners' ability to anticipate talker change. *$p < 0.01$; ***$p < 0.0001$; *n.s.* not significant.

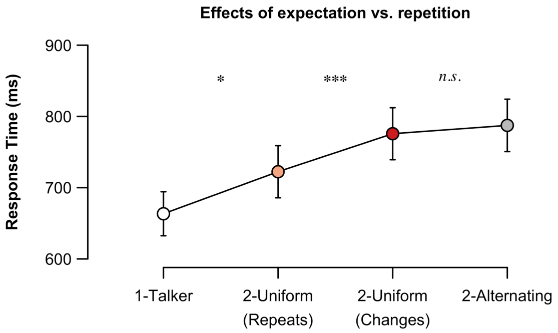


Table 3: Processing costs associated with anticipated vs. unanticipated talker continuity or change.

Contrast	*β*	*s.e.*	*t*	*p*
Single talker vs. unanticipated continuity	0.039	0.013	2.92	0.008
Unanticipated continuity vs. change	0.033	0.005	6.11	1×10^{-8}
Change vs. anticipated change (alternating)	0.003	0.009	0.35	0.73

3.3. Talker gender

During mixed-talker conditions with 4, 8, or 16 talkers, the talker changed between every trial. We investigated whether the degree of talker change between trials affected word recognition speed.

We compared RTs on trials where there was a greater change in a talker's phonetic characteristics from those of the preceding trial's talker (across-gender talker changes) to trials with smaller magnitude changes (within-gender talker changes) for each condition using a linear mixed effects model with a fixed factors of *number of talkers* (4, 8, or 16) and *previous talker gender* (same or different) and random factors as above. In a Type III analysis of variance (ANOVA) on this model, we found that RTs were significantly slower for a talker change across genders compared to within the same gender, regardless of the total number of talkers in the condition (Table 4).

Table 4: Processing costs associated with talker change across gender vs. within gender do not vary with number of total talkers in the condition.

Effect	*F*	*df(n, d)*	*p*
Number of talkers (4 vs. 8 vs. 16)	0.83	(1, 23.4)	0.45
Previous talker gender (same vs. different)	5.66	(1, 22.9)	0.026
Number of talkers × previous talker gender	1.87	(2, 4291.3)	0.15

4. DISCUSSION

We investigated whether the processing costs of mixed-talker speech varied with increasing number of talkers. We also explored whether trial-by-trial factors (unanticipated talker repetition, predictable talker change, and degree of between-talker phonetic differences) affected word identification speed.

Increasing the number of talkers, and therefore the amount of phonetic variability, did not further increase processing costs beyond those added by *any* talker variability (i.e., 2 talkers). Word identification was equally slow with 16 talkers as with just two. This result convincingly confirms the received wisdom that processing costs do not scale with the number of different talkers [15]. Moreover, this result requires us to revisit the idea that acoustic-phonemic mappings are made more efficient by reducing the decision space of possible interpretations of the acoustic signal [10]. Instead, our observations support a view of speech processing interference that arises when talker discontinuity disrupts listeners' ability to form a coherent speech stream from a consistent source [3,5,11,12,21].

Word recognition was likewise faster when the same talker spoke on two consecutive trials, even if the continuity was not predicable. This result suggests a feedforward, facilitatory effect of talker continuity on speech processing efficiency [3,5,12], consistent with predictions of feedforward auditory streaming models [11,21]. Even when listeners could perfectly anticipate which other talker would speak on the next trial, word recognition was not faster than when the next talker was unpredictable. This suggests that the facilitatory effects of talker continuity are feedforward, not feedback, consistent with streaming [3,5, 11,12,21,22,24], but not decision-space models [10].

Finally, the magnitude of talker-specific phonetic variation between trials did affect processing costs, suggesting that the degree of trial-to-trial difference in phonetic characteristics may contribute to the magnitude of processing costs incurred when auditory streaming is disrupted [13,19], thereby perhaps providing insight into how basic mechanisms of auditory adaptation [e.g., 7] may underlie talker adaptation in speech processing.

Together, these results (*i*) confirm the canonical interpretation of [15] that variability-induced processing costs do not scale with more talkers, (*ii*) are consistent with a view that interference effects result from disruption of a coherent auditory stream such as that afforded by talker continuity [5,21,24], and (*iii*) challenge the notion that top-down expectations can guide model selection in accounting for acoustic-phonetic correspondences across talkers [10].

5. REFERENCES

[1] Barr, D.J., Levy, R., Scheepers, C., & Tily, H.J. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J. Mem. Lang.* 68, 255-278.

[2] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5, 341-345.

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

[4] Choi, J.Y., Hu, E.R., Perrachione, T.K. 2018. Varying acoustic-phonemic ambiguity reveals that talker normalization is obligatory in speech processing. *Atten. Percept. Psychophys.* 80, 784-797.

[5] Choi, J.Y., Perrachione, T.K. Time and information in perceptual adaptation to speech. Submitted.

[6] Green, K.P., Tomiak, G.R., Kuhl, P.K. 1997. The encoding of rate and talker information during phonetic perception. *Percept. Psychophys.* 59, 675-692.

[7] Herrmann, B., Henry, M.J., Fromboluti, E.K., McAuley, J.D., Obleser, J. 2015. Statistical context shapes stimulus-specific adaptation in human auditory cortex. *Am. J. Physiol.-Heart C.* 113, 2582-2591.

[8] Hillenbrand, J., Getty, L.A., Clark, M.J., Wheeler, K. 1995. Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97, 7099-3111.

[9] Johnson, K. 2005. Speaker normalization in speech perception. In: D.B. Pisoni, R.E. Remez (eds), *The Handbook of Speech Perception*. Malden: Blackwell, 363-389.

[10] Kleinschmidt, D.F., Jaeger, T.F. 2015. Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychol. Rev.* 122, 148-203.

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

[12] Lim, S.-J., Shinn-Cunningham, B.G., Perrachione, T.K. in press. Effects of talker continuity and stimulus rate on auditory working memory. *Atten. Percept. Psychophys.*

[13] Magnuson, J.S., Nusbaum, H.C. 2007. Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *J. Exp. Psychol. Human* 33, 391-409.

[14] Morton, J.R., Sommers, M.S., Lulich, S.M. 2015. The effect of exposure to a single vowel on talker normalization for vowels. *J. Acoust. Soc. Am.* 137, 1443-1451.

[15] Mullennix, J.W., Pisoni, D.B. 1990. Stimulus variability and processing dependencies in speech perception. *Percept. Psychophys.* 47, 379-390.

[16] Mullennix, J.W., Pisoni, D.B., Martin, C.S. 1989. Some effects of talker variability on spoken word recognition. *J. Acoust. Soc. Am.* 85, 365-378.

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

[18] Nusbaum, H.C. and Morin, T.M., 1992. Paying attention to differences among talkers. *Speech perception, production and linguistic structure*, 113-134.

[19] Palmeri, T.J., Goldinger, S.D., Pisoni, D.B. 1993. Episodic encoding of voice attributes and recognition memory for spoken words. *J. Exp. Psychol. Learn* 19*,* 309-328.

[20] Peirce, J.W. 2007. PsychoPy: Psychophysics software in Python. *J. Neurosci. Meth.* 162, 8-13.

[21] Shinn-Cunningham, B.G. 2008. Object-based auditory and visual attention. *Trends Cogn. Sci.* 12, 182-186.

[22] Sjerps, M.J., Mitterer, H., McQueen, J.M. 2011. Constraints on the processes responsible for the extrinsic normalization of vowels. *Atten. Percept. Psychophys*. 73, 1195-1215.

[23] Sommers, M.S., Kirk, K.I., Pisoni, D.B. 1997. Some considerations in evaluating spoken word recognition by normal-hearing, noise-masked normal-hearing, and cochlear implant listeners. I: The effects of response format. *Ear Hearing* 18, 89.

[24] Winkler, I., Denham, S.L., Nelken, I. 2009. Modeling the auditory scene: predictive regularity representations and perceptual objects*.* *Trends Cogn. Sci.* 13, 532-540.

[25] Wong, P.C.M., Nusbaum, H.C., Small, S. L. 2004. Neural bases of talker normalization. *J. Cognitive Neurosci.* 16, 1173-1184.