ATTENTIONAL REORIENTATION EXPLAINS PROCESSING COSTS ASSOCIATED WITH TALKER VARIABILITY

Sung-Joo Lim, Jessica A. A. Tin, Allen Qu, and Tyler K. Perrachione

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

ABSTRACT

Speech processing is slower when listening to multiple talkers versus one continuous talker. Is this difference due to facilitation from perceptual adaptation to one talker's speech or interference from sudden switches between talkers? In two experiments, we examined how speech recognition speed depends on ongoing exposure to a talker. Listeners performed a speeded word identification task, in which they heard words from one talker for 2–7 consecutive trials before the talker switched. Word identification was slowest on trials where the talker switched and faster after a single exposure to a talker. However, additional exposure to a talker did not further expedite word identification. Furthermore, more frequent talker switches led to slower speech processing. Our findings suggest that speech processing efficiency does not depend on listeners becoming perceptually adapted to a talker's speech over time; rather, slower speech processing after a change in talker results from cognitive costs of attentional reorientation.

Keywords: talker adaptation, speech perception, auditory streaming, attentional reorientation

1. INTRODUCTION

Processing speech from multiple talkers introduces substantial phonetic variability [7], which creates further ambiguity in the nondeterministic mapping between speech acoustic and listeners' phonemic categories [12]. Correspondingly, listeners are less efficient at recognizing speech spoken by multiple talkers compared to one consistent talker [5,15,16].

It has been proposed that listeners' perception becomes rapidly adapted to a talker's speech [9,11], with perceptual tuning to talker-specific phonetic features facilitating speech processing by reducing the demands in resolving acoustic-phonemic ambiguity [11,17]. However, an alternative explanation of the relative interference from talker variability is that abrupt discontinuity in stimulus features disrupts listeners' attentional focus during speech processing. Discontinuity in the source of speech (i.e., a change in talker) imposes a cost to switch attention to the newly encountered source [1,3]. Thus, processing speech from multiple talkers may disrupt listeners' ability to efficiently form a coherent stream of speech [6,18].

The impact of talker variability has been mostly investigated by comparing processing speech from a single talker against processing mixed-talker speech where talkers constantly switched. Thus, it is unclear whether there is a *cost* associated with processing phonetically-variable, attentionally-disruptive speech from mixed talkers vs. a *benefit* of adaptation to speech from one continuous talker. Also, studies that compare processing under single- vs. mixed-talker conditions can reveal little about how processing speech from one continuous talker unfolds over time.

Here, we examined how continued exposure to a talker affects speech recognition using a speeded classification task. We investigated whether the duration of exposure to a single, continuous talker influences (*i*) listeners' response time for identifying subsequent words from that talker, and (*ii*) the magnitude of processing interference when switching to a new talker. We varied the number of consecutive trials of speech from a single talker prior to switching to a new talker.

If listeners become perceptually adapted to a talker over time, additional exposure to that talker should lead to faster speech processing. Furthermore, processing interference from an abrupt talker switch should be greater after more adaptation (i.e., be larger for less frequent talker switches). However, if talker discontinuity disrupts attentional focus to the source of a speech stream, word recognition speed should be affected only when the talker switches, and greater interference incurred with more frequent switches.

2. EXPERIMENT 1

2.1. Methods

2.1.1. Participants

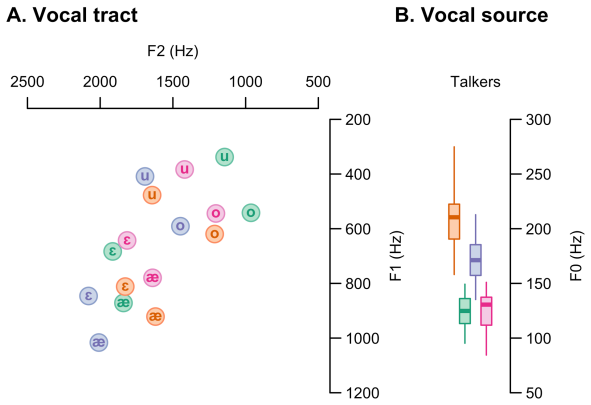
Native speakers of American English (N=20; age 18–33 years) with normal hearing were recruited. Participants gave written informed consent, approved by the Institutional Review Board at Boston University.

2.1.2. Stimuli and procedure

Two words (*boot* and *boat*) were recorded by four native speakers of American English (2 female) (Fig. 1), and were normalized to equivalent RMS amplitude.

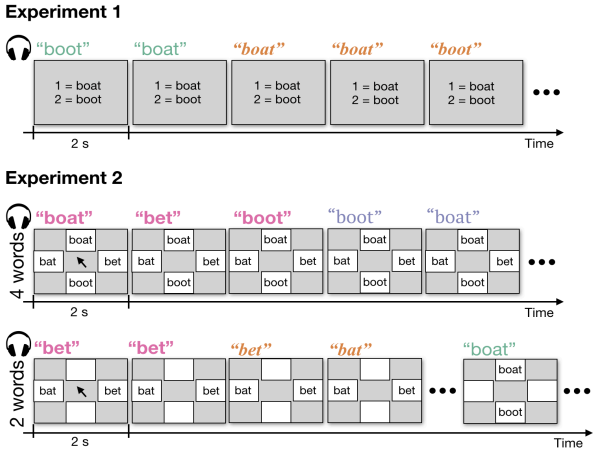
These words were chosen due to the high degree of acoustic-phonemic ambiguity across talkers [5].

Figure 1: Phonetic variability across talkers for the target words "boot," "boat," "bet," and "bat."



Listeners performed a speeded word identification task. On every 2-s trial, listeners identified the spoken word as quickly and accurately as possible using a keypad (Fig. 2). We parametrically varied the number of trials in a row with words from a single talker; listeners heard words from one talker for spans of 2–7 consecutive trials before the talker switched for the subsequent span. All span lengths were equally attested across talkers, and the length of spans preceding a talker switch was balanced across talkers. Transition probabilities between the words and across the talkers were equated throughout the experiment. Participants completed four, 326-trial blocks. The experiment was conducted in a sound attenuated chamber using PsychoPy (v.1.8.1). Stimuli were delivered with Sennheiser HD-380 pro headphones.

Figure 2: Illustrations of the word identification tasks. Colors and fonts denote different talkers.



2.1.3. Data analysis

The main dependent measure was response time (RT) on correct trials (accuracy: 96.2 ± 5.4%). Trials in which participants' log-transformed RT was greater than 3 standard deviation from their mean were excluded (<1% of correct trials). Prior to analysis, RT was log-transformed to ensure normality. Analyses were carried out using two separate linear mixed-effects models (*lme4* in R v3.3.3).

In the first model, we analysed how listeners' word identification speed changed across successive trials of a span after first encountering a new talker. We modeled the *number of consecutive encounters* of a talker as a fixed factor with 7 levels (0–6 trials; 0 being the first encounter of a new talker); we specified contrasts that tested the pairwise differences between successive increases in number of trials.

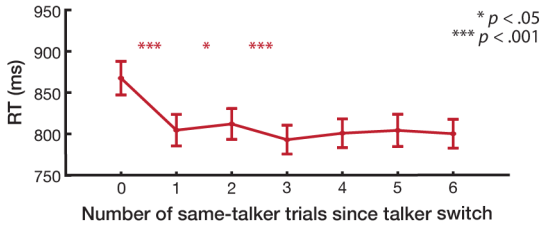
The second model examined whether the *amount of prior exposure* to a talker affected listeners' subsequent word identification when the talker switched. The length of the prior span before a talker switch was entered in the model as a factor (6 levels; 2–7 trials), with contrasts testing the differences between successive increases in the number of preceding trials.

Both models included random intercepts by participant. The significance of each factor was determined based on Type-II Wald χ^2 tests (*car* in R).

2.2. Results

First, we analysed how listeners' word identification speed changed across successive trials of a span from the first encounter of a talker. We found a significant effect of the number of repeated encounters of a talker ($\chi^2(6) = 564.25$; $p \ll 0.0001$); RTs were slowest upon first encountering a talker (Fig. 3), but a single repetition of the talker led to significantly faster responses (trial 0 vs. trial 1; $\beta = -0.033$, $t = 18.46$, $p \ll 0.0001$). On the second repeat, RT was slightly slower (trial 1 vs. trial 2; $\beta = 0.0041$, $t = 2.22$, $p = 0.026$), but became faster again on the third repeat (trial 2 vs. trial 3; $\beta = -0.0094$, $t = 4.54$, $p \ll 0.0001$) and plateaued for further exposure to the talker (all ts < 1.51, ps > 0.13).

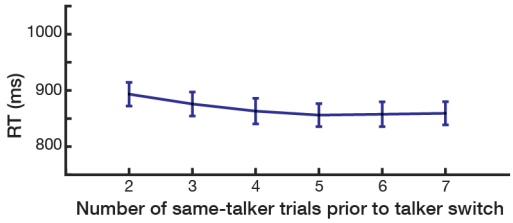
Figure 3: Mean RT for word identification on successive trials from a single talker. Error bars indicate ±1 standard error of the mean (SEM).



Next, we analysed whether the amount of prior exposure to a talker affected listeners' RTs when talker switched. We found that this factor had a significant

effect on RT ($\chi^2(5) = 23.69$; $p < 0.00025$; Fig. 4); listeners' responses were slower after less exposure to a talker (2 vs. 3 repetitions: $\beta = 0.008$, $t = 1.78$, $p = 0.075$; 3 vs. 4 repetitions: $\beta = 0.007$, $t = 1.66$, $p = 0.096$). Longer exposures did not affect listeners' RTs at switch (all ts < 0.52, ps > 0.61).

Figure 4: Word identification RT on talker switch trials as a function of the number same-talker preceding trials. Error bars indicate ±1 SEM.



2.3. Discussion

The results indicate that a single exposure to a talker was sufficient to maximally facilitate speech processing; there was no additional benefit from longer exposure to the same talker. Furthermore, hearing a new talker was not more disruptive after longer exposure to a preceding talker's speech; rather, *briefer* exposure to one talker before switching to a new one increased processing interference. That is, more frequent talker switches resulted in slower word identification. This pattern is inconsistent with perceptual adaptation to a talker. Instead, our findings are in line with an alternative explanation—that talker *discontinuity* interferes with speech processing by disrupting listeners' attentional focus and auditory streaming.

However, the simplicity of the decision may limit our ability to detect more graded changes in facilitation with sustained exposure to a talker. Could the rapid plateau in word identification speed after a single exposure to a talker merely reflect their ability to detect any change in acoustics from the previous trial of the same talker [8,13]? In Experiment 2, we increased the number of target words to test whether more graded perceptual adaptation might be evident with more degrees of freedom in responses.

3. EXPERIMENT 2

3.1. Methods

3.1.1. Participants

New participants (N=13, age 18–24 years) met the same inclusion/exclusion criteria as Experiment 1.

3.1.2. Stimuli and procedure

Stimuli comprised amplitude-normalized recordings of natural productions of four English words (*boot*, *boat*, *bet*, *bat*) spoken by four native speakers of American English (2 male, 2 female; Fig. 1).

Listeners identified a target word every 2 s by clicking a mouse button on the matching word displayed on the screen (Fig. 2). On each trial, listeners either heard a word spoken by the same talker as the previous trial—for spans of 2–7 consecutive same-talker trials—or spoken by a different talker.

Listeners completed four, 444-trial blocks. In two blocks, listeners were presented with all four target words (4AFC). In the other two blocks, listeners heard only two of the four words (2AFC), divided into six sub-blocks for all six possible pairs of words.

As in Experiment 1, the number of spans of each length were equated across talkers, as were the transitions between the number of successive trials prior to switching to each talker. Transition probabilities between the four target words and across the four talkers, and the number of presentations of each stimulus were equated. The order of 2AFC and 4AFC blocks were counterbalanced across participants.

3.1.3. Data analysis

RTs of correct trials were the main dependent measure (accuracy: 97.4 ± 2.8%). Log-transformed RTs were analysed using two separate linear mixed-effects models.

Following Experiment 1, the first model examined the effect of exposure to a talker within a span (7 levels; 0–6 trials) on identifying speech by that talker. The second model examined the effect of the amount of prior exposure to a preceding talker's speech (6 levels; 2–7 trials) on the subsequent identification of a word spoken by a new talker.

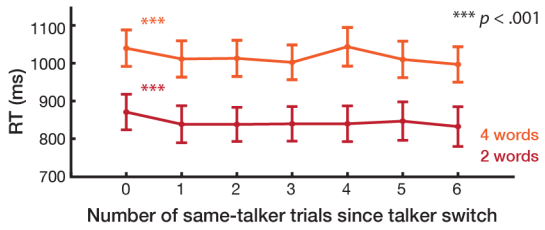
In both models, we also examined whether these effects differed depending on the number of response options (i.e., 2AFC vs. 4AFC) serving as a fixed factor in the model. Both models included random intercepts and slopes by participant. The significance of factors was determined by Type-II Wald χ^2 tests.

3.2. Results

First, we analysed whether the number of successive exposures to a talker affected subsequent identification of the same talker's speech, and whether it differed between the 4- vs. 2-word choice trials (Fig. 5). This model of listeners' RTs revealed significant effects of the number of repeated encounters of a talker ($\chi^2(6) = 90.92$; $p \ll 0.0001$) and the number of response options ($\chi^2(1) = 112.40$; $p \ll 0.0001$), and their interactions ($\chi^2(6) = 13.00$; $p = 0.043$). Listeners were significantly slower at identifying words in 4AFC than 2AFC trials across the sequence of same-talker trials ($\beta = 0.091$, $t = 10.55$, $p \ll 0.0001$). Furthermore,

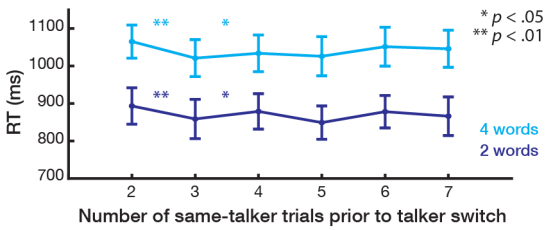
a single repeated exposure to a talker after a talker switch significantly expedited the RTs (trial 0 vs. 1; $\beta = 0.015$, $t = 5.81$, $p \ll 0.0001$) and the degree of RT reduction was similar for 2AFC vs. 4AFC decision trials ($t = 0.80$, $p = 0.42$). However, longer exposure to the same talker's speech did not lead to faster RTs (trial 1 vs. 2, 2 vs. 3, etc; all ts < 0.51, ps > 0.61).

Figure 5: Listeners' RTs for repeated exposure to a talker from the first encounter with the talker. Orange and red lines indicate RTs in 4AFC and 2AFC trials, respectively. Error bars indicate ± 1 SEM. (Note that the y-axis range differs from Fig. 3.)



In the second model, we examined whether listeners' RTs on talker-switch trials were affected by the amount of exposure to the preceding talker, and whether this effect depended on the number of response options (Fig. 6). The model on RTs at talker-switch trials revealed significant effects of the amount of exposure to the preceding talker ($\chi^2(5) = 23.54$; $p < 0.0003$) and the number of response options ($\chi^2(1) = 107.58$; $p \ll 0.0001$), but no significant interaction between these factors ($\chi^2(5) = 1.96$; $p = 0.85$). Listeners were significantly slower when the preceding talker was heard for two compared to three trials ($\beta = 0.021$, $t = 3.25$, $p = 0.0012$). RTs were slightly slower with an additional repetition of the preceding talker (4 vs. 3 trials; $\beta = 0.015$, $t = 2.28$, $p = 0.023$), but additional exposure to the preceding talker did not affect RT (all ts < 1.86, ps > 0.063). While listeners were slower when choosing from among four words than two at switch trials ($\beta = 0.088$, $t = 10.37$, $p \ll 0.0001$), the effect of the number of prior exposures at switch was similar for both conditions (all ts < 1.18, ps > 0.24).

Figure 6: Word identification RTs at talker-switch trials across preceding talker span lengths. Light and dark blue lines indicate 4AFC and 2AFC trials, respectively. Error bars indicate ± 1 SEM.



3.3. Discussion

Listeners were slower at identifying words when the number of alternative choices increased. As in Experiment 1, a single exposure to a talker immediately facilitated identification of words spoken by the same talker, even with the increased number of alternative choices. Also, there was no additional facilitation from longer exposure to the talker, even when there were more response choices. Furthermore, our findings suggest that shorter amounts of exposure to a preceding talker leads to greater interference in processing speech from a new talker regardless of the number of choices that listeners have to make. Thus, we again found that talker discontinuity disrupts listeners' speech processing, and this disruption is exacerbated with increasingly frequent talker switches.

4. CONCLUSION

In two experiments, we examined why speech processing efficiency differs for single- vs. mixed-talker speech. After the first encounter with a new talker, further exposure to their speech did not make speech processing more efficient. Likewise, processing interference when encountering a new talker was greatest after shorter exposure to a prior talker.

Prior studies have variously described this difference as the *benefit* of talker adaptation or the *cost* of interference from multiple talkers [3,5,15,17]. Our results suggest an account of speech processing where listeners' attention is disrupted by a change in talker, rather than one where they become perceptually adapted to a talker's speech over time. Processing interference from mixed-talker speech appears to result from attentional disruption that impairs listeners' ability to form a coherent auditory stream [2,18].

It is possible that uncertainty about the upcoming talker can influence speech processing efficiency [14], as we observed slower responses when listeners encountered the same talker's speech for three consecutive trials than a single repeat of a talker (Fig. 3). However, recent work indicates that not only does a change in talker consistently interferes with speech processing efficiency, but that repetition of a talker is always facilitatory, regardless of listeners' expectation about the upcoming talker [4,10]. Furthermore, the structural certainty/uncertainty about the upcoming talker was identical in both Experiments 1 and 2, but we did not observe the same pattern of RT changes between repetition trials 2 and 3 in these two experiments. Thus, rather than the top-down expectation about the talker, bottom-up change in the source of speech stream may make the greater contribution to processing interference from mixed-talker speech.

5. REFERENCES

[1] Best, V., Ozmeral, E. J., Kopčo, N., Shinn-Cunningham, B. G. 2008. Object continuity enhances selective auditory attention. *Proc. Natl. Acad. Sci. U.S.A.* 105, 13174–13178.

[2] Bregman, A. S. 1990. *Auditory Scene Analysis.* Cambridge, MA: MIT Press.

[3] Bressler, S., Masud, S., Bharadwaj, H., Shinn-Cunningham, B. 2014. Bottom-up influences of voice continuity in focusing selective auditory attention. *Psychol. Res.* 78, 349–360.

[4] 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).

[5] 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.

[6] Darwin, C. J., & Carlyon, R. P. 1995. Auditory Grouping. In B. C. Moore (Ed.), *Hearing Handbook of Perception and Cognition.* Academic Press, 387–424.

[7] Hillenbrand, J., Getty, L. A., Clark, M. J., Wheeler, K. 1995. Acoustic characteristics of American English vowels. *J. Acoust. Soc. Am.* 97, 3099–3111.

[8] Holt L. L., Lotto A. J., Kluender K. R. 2000. Neighboring spectral content influences vowel identification. *J. Acoust. Soc. Am.* 108, 710–722.

[9] Johnson, K. 1990. The role of perceived speaker identity in F0 normalization of vowels. *J. Acoust. Soc. Am.* 88, 642-654.

[10] Kapadia, A.M. & Perrachione, T.K. 2019. Processing costs associated with talker variability do not scale with number of talkers. *19th International Congress of Phonetic Sciences* (Melbourne, August 2019).

[11] 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.

[12] Liberman, A. M., Cooper, F. S., Shankweiler, D. P., Studdert-Kennedy, M. 1967. Perception of the speech code. *Psychol. Rev.* 74, 431–461.

[13] Lotto, A.J., Kluender K. R. 1998. General contrast effects of speech perception: Effect of preceding liquid on stop consonant identification. *Percept. Psychophys.* 60, 602–619.

[14] Magnuson, J. S., Nusbaum, H. C. 2007. Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *J. Exp. Psychol. Hum. Percept. Perform. 33*, 391–409.

[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. 1997. Talker normalization: Phonetic constancy as a cognitive process. In K. A. Johnson & J. W. Mullennix (Eds.), *Talker variability and speech processing.* New York, NY: Academic Press, 109–132.

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