TALKER CONTINUITY FACILITATES SPEECH PROCESSING INDEPENDENT OF LISTENERS' EXPECTATIONS

Yaminah D. Carter, Sung-Joo Lim, and Tyler K. Perrachione

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

ABSTRACT

Recognizing speech is faster when listening to speech from a single continuous talker than mixed talkers. Does facilitation from talker adaptation depend on listeners' expectations about talker continuity? Here, we measured response times during a speeded word identification task for pairs of words. We manipulated listeners' expectations about hearing the same talker by varying the probability of talker continuity within word pairs across two conditions: high probability of same-talker trials with rare different-talker trials or vice-versa. Word recognition was faster for same-talker trials, regardless of listeners' expectations about talker continuity. In a follow-up experiment, we measured listeners' response times to pairs of words spoken by one talker in three conditions that manipulated expectations about whether the word itself would be repeated. Here, only expected word repetitions led to faster response times. These results suggest that talker adaptation is a feed-forward process, occurring automatically during speech perception.

Keywords: talker adaptation, priming, expectation, prediction, speech perception

1. INTRODUCTION

Efficiently extracting linguistic messages from the enormous acoustic-phonetic variability across talkers is challenging due to the lack of invariance between speech acoustics and listeners' abstract phonemic representations [10]. Yet listeners recognize speech from one continuous talker faster than speech from mixed talkers, suggesting that talker adaptation makes speech processing more efficient [4,12,13].

The efficiency gains from listening to one talker's speech are thought to result because listeners can reduce the degrees of freedom in mapping between speech acoustics and phonemic representations, thereby making perceptual decisions more efficient [14]. Models of speech processing formalize this by testing reduced numbers of potential interpretations of a speech signal based on expectations about its source [9]. Expectations about the perceptual world in general can facilitate processing by reducing the neurocomputational demands supporting perception [17]. Correspondingly, expected repetitions of a talker elicit smaller neural responses than unexpected repetitions [1], suggesting that neural signatures of talker adaptation may in part reflect fulfilled perceptual expectations [3,16,20].

However, research showing adaptation-related efficiency gains in speech perception primarily compare listening to long blocks of a single continuous talker vs. long blocks of mixed, unpredictable talkers [4,12,16,20]. Thus, it is impossible to distinguish how listeners' expectations of talker continuity contribute to talker adaptation effects. Only one study probed the role of expectations in talker adaptation: expecting to hear different talkers can impede speech processing [11]. However, it is unknown whether efficiency *gains* come from expected vs. unexpected talker continuity, as hypothesized by both speech-specific [9] and domain-general [17] models of perception.

Could the faster and more accurate speech processing from talker adaptation [14] be understood as the result of fulfilled top-down perceptual expectations? Here, we investigated how word identification speed is affected by listeners' expectations about the upcoming talker. We manipulated listeners' expectations to hear speech from a continuous talker vs. different talkers by having them identify word pairs when there was a high probability of repetition vs. high probability of change. If top-down expectations facilitate talker adaption [11], we would expect to see faster speech recognition when the expectations about the talker are met compared to when they are violated.

Finding only feed-forward, not expectation-based, effects of talker continuity from a first experiment, we ran a second, control experiment to confirm that speech processing efficiency could be affected by expectations, now at the lexical level.

EXPERIMENT 1: EXPECTATION OF TALKER CONTINUITY

2.1. Methods

2.1.1. Participants

Native American English speakers (N=20; 14 female, 6 male; age 18–26 years) participated in this

experiment. All participants had a self-reported history free from speech, hearing, or language disorders. Participants provided written informed consent. All experi-mental procedures were approved by the Institutional Review Board at Boston University.

2.1.2. Stimuli

The target words "boot" and "boat" were recorded by four native speakers of American English (2 female). These target words were selected because the acoustics of the /u/-/o/ contrast are highly variable across talkers [7], leading to heightened processing interference in a mixed-talker setting [4]. Word durations were on the order of 227 ± 15 ms (mean ± s.d.).

2.1.3. Procedure

Participants performed a speeded word identification task, in which they indicated whether each word was "boot" or "boat" as quickly and accurately as possible. Trials were organized around pairs of words, such that two words were played in succession with a 2-s stimulus onset asynchrony. The two words in a trial were either spoken by different talkers (talker change) or a single talker (talker repetition). Each trial (pair of words) was followed by a 2-s silent interval (**Fig. 1**), which created the temporal structure of trials that allowed us to manipulate listeners' expectations about repetition vs. change within word pairs.

Participants performed this task in two conditions that varied their expectations as to whether the same talker would say both words in a trial: (1) in the *expect change* condition, the first and second words in a trial were predominately spoken by different talkers (80% of trials), with infrequent trials (20%) in which the same talker produced both the first and second words; and (2) in the *expect repeat* condition, the first and second words in a trial were predominately spoken by the same talker (80% of trials), with infrequent trials (20%) in which the second word was produced by a different talker than the first. Similar procedures have been widely used to induce perceptual expectancy effects in participants [1,18].

Each condition was completed across four, 60-trial blocks (240 trials / condition). The more probable trial structure (e.g., talker-repeats) occurred in 192 trials, while the less probable structure (e.g., talker-changes) occurred in 48 trials. Stimuli were pseudorandomized such that the same word was not presented for more than three consecutive trials.

Written directions remained on a monitor throughout the experiment, instructing participants to report the word they heard by pressing the corresponding key on a number pad ("boot" = 1, "boat" = 2).

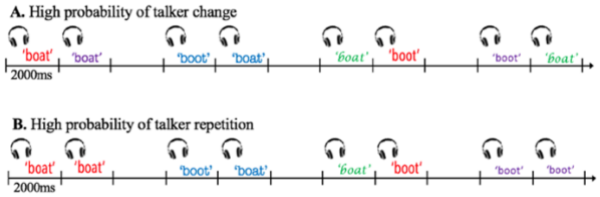
2.1.4. Data analysis

The focus of the study was to examine whether word identification speed is affected (*i*) by continuity in the source of speech across words in an auditory stream (i.e., whether a word was said by the same talker as the previous word), (*ii*) by listeners' expectations about whether there would be talker continuity, and (*iii*) by an interaction between talker continuity and listeners' expectations of continuity.

To this end, we analyzed the response times (RTs) in identifying the second target word in the word pair (i.e., from the onset of the second word) using a linear mixed-effects model (*lme4* in R v.3.3.3). The model included two fixed factors and their interaction: *talker repetition* in word pairs (talker repetition vs. talker change) and listeners' *expectations* about talker repetition (expected repetition vs. expected change), as well as random intercept and slopes by participant. Significant effects were determined based on Type-II Wald χ^2 tests (*car* in R). Any significant effects found in the model were followed up by testing pairwise differences of least-squares means (*difflsmeans* in R).

Trials where participants incorrectly identified the second word in the trial or with RTs exceeding three standard deviations from the participant's mean were excluded from analysis (~4.7% of trials). RTs were log-transformed to conform to normality.

Figure 1: Trial structure for the **(A)** high probability of talker change and **(B)** high probability of talker repeat conditions. Target words were presented in pairs for 2-s each, with trials separated by 2-s silent intervals. Colors denote different talkers.



2.2. Results

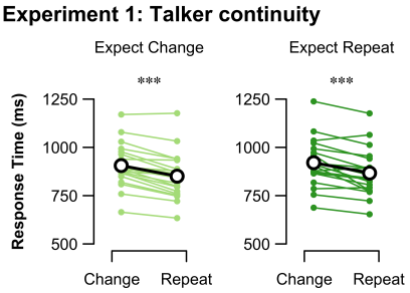
The linear mixed-effects analysis on RTs revealed a significant main effect of talker repetition ($\chi^2(1) = 62.70$, $p < 0.001$), but no significant main or interaction effects related to listeners' expectation about the upcoming talker (expectation: $\chi^2(1) = 1.14$, $p = 0.29$; talker repetition × expectation: $\chi^2(1) = 0.12$, $p = 0.73$). Pairwise differences tests revealed that RTs were significantly faster when two words in the pair were spoken by the same talker than different talkers

regardless of listeners' expectation about the talker repetition (**Fig. 2**; talker-repeat: 861ms; talker-change: 908ms; $\beta = -0.027$, $t = 7.92$, $p < 0.001$).

2.3. Discussion

Both expected and unexpected repetitions of a talker led to faster word identification compared to trials where the talker changed, and there was no difference in the magnitude of this facilitation based on listeners' expectations. This result suggests that the efficiency gains realized from talker adaptation [4,12] may not result from fulfillment of top-down expectations based on which talker a listener anticipates hearing. Here, talker adaptation appears to be a feed-forward process, where coherence in the auditory stream rapidly and automatically tunes the auditory system for more efficient processing of speech from that source [2,15,19]. This may suggest that the computational processes underlying talker adaptation, viz., reducing the number of possible interpretations for acoustic-phonemic mappings [9] may also be a feed-forward, rather than feed-back process. Alternatively, it may challenge the idea that such model selection is happening at all.

Figure 2: Mean RTs to the second target words in the trial pair. Thin lines and points denote RTs for individual participants; bold lines with large white circles indicate group mean RT. *** $p < 0.001$.



However, because Experiment 1 failed to induce expectation-related differences in word identification speed at all, we wanted to confirm that such expectation-related differences could be obtained under this paradigm. Consequently, we ran a second experiment under an identical trial structure, but where we varied the probability and expectation of repeating the word itself, rather than who said it.

3. EXPERIMENT 2: EXPECTATION OF LEXICAL REPETITION

3.1. Methods

3.1.1. Participants

A new sample of 20 native American English speakers (13 female; age 18–26 years) was recruited. Participation criteria were the same as in Experiment 1.

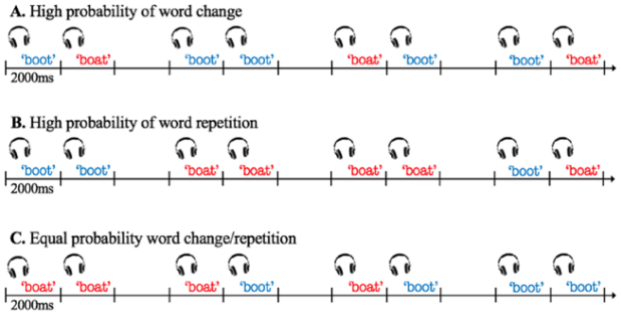
3.1.2. Stimuli

Stimuli consisted of the same target words "boot" and "boat" from Experiment 1, but recordings from only one male talker were used.

3.1.3. Procedure

The task was identical to that of Experiment 1. On each trial, participants identified each word (i.e., "boot" or "boat") in the word pair (**Fig. 3**). The two words in each trial were either different (word change) or the same (word repetition). Participants performed this task under three conditions: (1) *expectation of word change*, (2) *expectation of word repetition* and (3) *no expectation of word change or repetition*. Each condition of the task was completed across two, 60-trial blocks (120 trials / condition). In the expect-change or expect-repeat conditions, the higher probability (expected) trial structure occurred in 96 trials (80%) and the less probable (unexpected) structure occurred in 24 trials (20%). In the equal probability condition, the paired words were equally likely to change or repeat (50%), with trials presented in a pseudo-random order. The same written directions as Experiment 1 were displayed on the monitor throughout the experiment.

Figure 3: Trial structure for the **(A)** high probability of word change, **(B)** high probability of word repeat, and **(C)** equal probability conditions. Target words (shown here by color) were presented in pairs, with trials separated by 2-s silent intervals. Recordings from one talker were used throughout.



3.1.4. Data analysis

The focus of Experiment 2 was to examine whether word identification speed is affected (*i*) by repetition of a target word itself, (*ii*) by listeners' expectations about whether the word would be repeated or change, and (*iii*) by an interaction between word repetition and listeners' expectations of repetition or change.

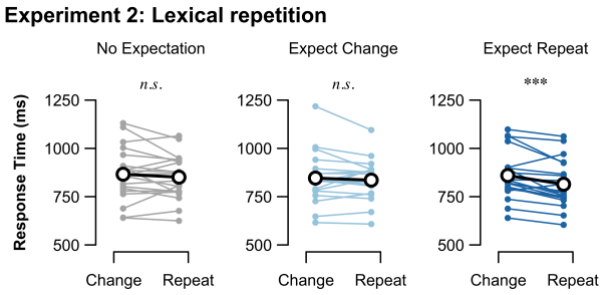
As in Experiment 1, we analyzed RTs for correct identifications of the second word in each pair (i.e., from the onset of the word). We used a linear mixed-effects model to examine whether listeners' RTs were influenced by *word repetition* (word-repeats vs. word-changes) and by listeners' *expectations* about the word repetition (expect word to repeat, expect word to change, and no expectation). In addition to these fixed factors and their interactions, the model included random intercepts and slopes by participant. Any significant effects determined by the Wald χ^2 test were followed up by pairwise differences of least-squares means tests (*difflsmeans*). Prior to the analysis, RTs were log-transformed and log-RTs greater than 3 SDs from each participant's mean were excluded (~1.8% of trials).

3.2. Results

3.2.1. Effects of word repetition and expectation

The linear mixed-effects model revealed significant effects of word repetition, listeners' expectation, and their interactions on RT (**Fig. 4**; word repetition: $\chi^2(1) = 4.70$, $p = 0.030$; expectation: $\chi^2(2) = 7.94$, $p = 0.019$; word repetition × expectation: $\chi^2(2) = 16.14$, $p < 0.001$). Subsequent pairwise tests revealed that listeners were significantly faster to identify the target word only when they expected a repetition and their expectations were met ($\beta = -0.025$, $t = 4.08$, $p < 0.001$). However, faster RTs associated with word repetition (vs. change) were not observed when listeners had no expectation ($\beta = -0.0073$, $t = 1.31$, $p = 0.20$), nor when they expected the word to change ($\beta = -0.0016$, $t = 0.27$, $p = 0.79$) Likewise, RTs were not faster when listeners' expectations of word change were fulfilled ($\beta = 0.0077$, $t = 1.06$, $p = 0.30$).

Figure 4: Mean RTs to target words in the second position of the trial pair. Thin lines and points denote RTs for individuals, bold lines with large white circles indicate mean RTs across participants. A significant repetition × expectation effect was found; this suggests that listeners were faster to identify the same word repeated in the trial only when they expected the repetition. ***$p < 0.001$; *n.s.*, not significant.



3.2.2. Comparison to Experiment 1

Experiment 1 revealed faster RTs for talker-repetition trials even when listeners expected the talker to change, whereas Experiment 2 results did not show any difference in RTs when listeners expected the word to change. Thus, we used a new linear mixed-effects model of RTs to directly compare whether the impact of *repetition* (vs. change) of either the talker (Experiment 1) or word (Experiment 2) differed based on which of these *dimensions* (talker vs. word) listeners expected to change.

We did not find a significant main effect of *dimension* on RTs ($\chi^2(1) = 0.52$, $p = 0.47$), but the model revealed a significant main effect of *repetition* ($\chi^2(1) = 30.59$, $p < 0.001$) and a significant *dimension* × *repetition* interaction ($\chi^2(1) = 15.75$, $p < 0.001$), suggesting that the feed-forward facilitatory effects of repetition differed depending on whether the unexpectedly repeated dimension was talker or word. Post hoc pairwise tests revealed that the effect of unexpected dimension repetition was significant only when the talker unexpectedly repeated (Experiment 1: talker repeats vs. expected changes; $\beta = -0.027$, $t = 6.80$, $p < 0.001$), but not when the word itself repeated (Experiment 2: word repeats vs. expected changes; $\beta = 0.0018$, $t = 0.35$, $p = 0.73$).

3.3. Discussion

Only expected repetition of a word led to faster word identification compared to trials on which the word changed. When listeners had no expectations about word repetition, or when they expected the word to change, repetition was not facilitatory. This suggests

that faster processing of repeated speech content results from fulfilled top-down expectations.

In Experiment 2, when a word change was expected, only one other word (boot/boat) was possible (in contrast to Experiment 1, where correctly expecting a change in talker still left uncertain which of the other three talkers would be heard next). Surprisingly, correctly expecting a word change also did not result in faster word identification. It appears that only fulfilled expectations of word repetition are facilitatory for speech processing, whereas expected changes or unexpected repetitions have no effect compared to having no expectations at all.

4. CONCLUSIONS

The results of these experiments contrast starkly: Listeners were significantly faster identifying words spoken by the same talker even while expecting to hear a different talker, but an unexpected word repetition did not expedite processing. These results suggest that talker adaptation-related efficiency gains in speech processing may reflect bottom-up mechanisms that interpret speech in the context of preceding speech [6,8,15]. Speech processing efficiency appears to primarily be driven by continuity in the speech source [2,19], not top-down expectations [9], suggesting obligatory talker adaptation in speech processing [4].

5. REFERENCES

[1] Andics, A., Gál, V., Vicsi, K., Rudas, G., Vidnyánszky, Z. 2013. FMRI repetition suppression for voices is modulated by stimulus expectations. *NeuroImage* 69, 277–283.

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

[3] Chandrasekaran, B., Chan, A.H., Wong, P.C.M. 2011. Neural processing of what and who information in speech. *J. Cog. Neurosci*. 23, 2690–2700.

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

[5] 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.

[6] Heald, S.L.M., Van Hedger, S.C., Nusbaum, H.C. 2017. Perceptual plasticity for auditory object recognition. *Front. Psychol.* 8:781.

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

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

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

[11] Magnuson, J., Nusbaum, H. 2007. Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *J. Exp. Psychol—Human*, 33, 391–409.

[12] Mullennix, J., & Pisoni, D. 1990. Stimulus variability and processing dependencies in speech perception. *Percept. Psychophys.* 47, 379–390.

[13] Mullennix, J., Pisoni, D., Martin, C. 1989. Some effects of talker variability on spoken word recognition. *J. Acoust. Soc. Am.* 85, 365–378.

[14] Nusbaum, H., Magnuson, J. 1997. Talker normalization: Phonetic constancy as a cognitive process. In K. Johnson & J.W. Mullennix (Eds.), *Talker variability in speech processing*, 109–132.

[15] Nusbaum, H.C., Morin, T.M. (1992). Paying attention to differences among talkers. In Y. Tohkura, Y. Sagisaka, E. Vatikiotis-Bateson (Eds.), *Speech perception, production, and linguistic structure*. Tokyo: Ohmasha Publishing, 113–134.

[16] Perrachione, T.K., Del Tufo, S.N., Winter, R., Murtagh, J., Cyr, A., Halverson, K., Ghosh, S.S., Christodoulou, J.A., Gabrieli, J.D.E. 2016. Dysfunction of rapid neural adaptation in dyslexia. *Neuron*, 92, 1383–1397.

[17] Summerfield, C., de Lange, F. 2014. Expectation in perceptual decision making: neural and computational mechanisms. *Nat. Rev. Neurosci*. 15, 745–756.

[18] Summerfield, C., Trittschuh, E., Monti, J., Mesulam, M-M., Egner, T. 2008. Neural repetition suppression reflects fulfilled perceptual expectations. *Nat. Neurosci*. 11, 1004–1006.

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

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