Attention, Perception, & Psychophysics
https://doi.org/10.3758/s13414-019-01684-w

PERCEPTUAL/COGNITIVE CONSTRAINTS ON THE STRUCTURE OF SPEECH COMMUNICATION: IN HONOR OF RANDY DIEHL

Effects of talker continuity and speech rate on auditory working memory

Sung-Joo Lim [1,2] · Barbara G. Shinn-Cunningham [2] · Tyler K. Perrachione [1]

Abstract
Speech processing is slower and less accurate when listeners encounter speech from multiple talkers compared to one continuous talker. However, interference from multiple talkers has been investigated only using immediate speech recognition or long-term memory recognition tasks. These tasks reveal opposite effects of speech processing time on speech recognition – while fast processing of multi-talker speech impedes immediate recognition, it also results in more abstract and less talker-specific long-term memories for speech. Here, we investigated whether and how processing multi-talker speech disrupts working memory maintenance, an intermediate stage between perceptual recognition and long-term memory. In a digit sequence recall task, listeners encoded seven-digit sequences and recalled them after a 5-s delay. Sequences were spoken by either a single talker or multiple talkers at one of three presentation rates (0-, 200-, and 500-ms inter-digit intervals). Listeners' recall was slower and less accurate for sequences spoken by multiple talkers than a single talker. Especially for the fastest presentation rate, listeners were less efficient when recalling sequences spoken by multiple talkers. Our results reveal that talker-specificity effects for speech working memory are most prominent when listeners must rapidly encode speech. These results suggest that, like immediate speech recognition, working memory for speech is susceptible to interference from variability across talkers. While many studies ascribe effects of talker variability to the need to calibrate perception to talker-specific acoustics, these results are also consistent with the idea that a sudden change of talkers disrupts attentional focus, interfering with efficient working-memory processing.

Keywords Talker adaptation · Speech perception · Auditory working memory · Recall efficiency · Aditory streaming

Introduction

Prior research has revealed that talker variability can interfere with speech processing. Compared to processing speech from one consistent talker, listening to speech spoken by a series of multiple talkers leads to slower and/or less accurate speech recognition (Choi, Hu, & Perrachione, 2018; Mullennix & Pisoni, 1990; Mullennix, Pisoni, & Martin, 1989; Nusbaum & Magnuson, 1997). Since the phonetic realization of speech varies greatly depending on who said it and in what context (Hillenbrand, Getty, Clark, & Wheeler, 1995; Peterson & Barney, 1952), speech phonetics do not have a deterministic relationship with abstract phonemic representations (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). Therefore, some authors have interpreted the interference effect of listening to speech from multiple talkers as indicating that there is a processing cost to recalibrating listeners' perception to accommodate differences in phonetic-phonemic correspondence across talkers and tokens (Magnuson & Nusbaum, 2007; Nearey, 1998; Nusbaum & Morin, 1992). However, when the source of a speech stream suddenly switches between subsequent tokens, there is also a disruption of featural continuity in the auditory stimulus. An alternative explanation for the interference effect from multiple talkers is that such bottom-up, stimulus-driven discontinuity may disrupt listeners' attentional focus and lead to a perceptual cost when listeners must switch attention between different sources and features encountered in the speech stream (Best, Ozmeral, Kopčo, & Shinn-Cunningham, 2008; Bressler, Masud, Bharadwaj, & Shinn-Cunningham, 2014). Thus, talker discontinuity may impede a listener's ability to integrate auditory events over time (Shinn-Cunningham,

✉ Sung-Joo Lim
sungjoo@bu.edu

[1] Department of Speech, Language, and Hearing Sciences, Boston University, 635 Commonwealth Ave, Boston, MA 02215, USA

[2] Biomedical Engineering, Boston University, Boston, MA, USA

Published online: 08 February 2019

2008), interfering with auditory stream formation (Darwin & Carlyon, 1995; Sussman, Horváth, Winkler, & Orr, 2007).

While talker-specific characteristics of speech sounds are putatively irrelevant to the linguistic content of speech itself, knowledge about talkers can affect how listeners perceive linguistic messages (e.g., Johnson et al., 1999; Niedzielski, 1999). Furthermore, talker-specific information is indeed encoded in long-term memory for speech (Geiselman & Bellezza, 1977). For instance, listeners recognize words less accurately when they are spoken by a different talker during encoding versus recognition (Bradlow, Nygaard, & Pisoni, 1999; Craik & Kirsner, 1974; Palmeri, Goldinger, & Pisoni, 1993). Furthermore, familiarity with a talker's voice enhances intelligibility of that talker's speech under adverse listening conditions (Nygaard & Pisoni, 1998; Nygaard, Sommers, & Pisoni, 1994; Souza, Gehani, Wright, & McCloy, 2013). However, inclusion of talker-specific information in long-term memory for speech appears to require a suitable amount of time or effort during speech encoding, as slower or more effortful processing is more likely to yield talker-specific long-term memories (Luce & McLennan, 2005; but see Theodore, Blumstein, & Luthra, 2015). According to this view, fast processing of speech only allows encoding of abstracted content of speech information while discarding talker-specific details of speech. However, this notion is inconsistent with the fact that fast processing of multiple talkers' speech interferes with listeners' immediate recognition of speech (Choi et al., 2018; Green, Tomiak, & Kuhl, 1997; Mullennix et al., 1989; Mullennix & Pisoni, 1990).

To date, the cost of listening to multi-talker speech has been investigated in the context of either immediate perceptual processing (e.g., Choi et al., 2018; Mullennix et al., 1989; Mullennix & Pisoni, 1990) or episodic long-term memories for speech (e.g., Goldinger, Pisoni, & Logan, 1991; Martin, Mullennix, Pisoni, & Summers, 1989; Nygaard et al., 1994; Nygaard, Sommers, & Pisoni, 1995; Palmeri et al., 1993). It remains unknown whether and how talker variability affects an intermediate processing stage that lies between perception and long-term memory – that is, the processing and short-term maintenance of speech information in working memory. Furthermore, while both immediate speech recognition and long-term memory for speech exhibit interference effects from multi-talker speech, the rate of processing multi-talker speech appears to influence these two levels of processing in opposite ways. Correspondingly, nothing is known about how the time for processing multi-talker speech during encoding affects working memory representations for speech – particularly whether the effect of processing time on speech working memory is similar to immediate speech recognition or long-term memory for speech. The classic working memory model asserts that the contents of working memory are abstract representations of linguistic units (Baddeley, 1992), which would suggest processing time should affect working memory and long-term memory similarly. However, recent studies have demonstrated that stimulus-specific acoustic details (e.g., syllable pitch) are also maintained in working memory (Lim, Wöstmann, & Obleser, 2015; Lim, Wöstmann, Geweke, & Obleser, 2018). This raises the possibility that talker variability during perceptual processing will also influence auditory working memory representations, and will do so in a manner similar to how it affects immediate speech recognition. The present study examines whether and how variability in task-irrelevant indexical features (i.e., speech spoken by multiple talkers) of speech impacts the speed, accuracy, and efficiency of recall from auditory working memory.

Here, we investigated whether processing speech from multiple talkers interferes with working memory performance compared to processing speech from one consistent talker. Furthermore, we tested how any such cost changed as a function of the time available to process individual speech tokens. Using a delayed recall of digit span task, we calculated listeners' working memory[1] recall *efficiency* (an integrated statistic combining processing speed and accuracy; Townsend & Ashby, 1978) in conditions manipulating talker variability (single talker vs. multiple talkers) and time to encode each digit (0-, 200-, or 500-ms inter-stimulus intervals; ISIs). We hypothesized that talker variability would increase cognitive cost and/or attentional disruption, leading listeners to be less accurate and efficient in recalling digits spoken by multiple talkers compared to a single talker. If talker-specific details in speech are discarded at fast speech rates (e.g., McLennan & Luce, 2005), we would expect the processing cost to be reduced when hearing multiple talkers at faster rates. Conversely, the emergence of auditory streaming (i.e., the integration of discrete acoustic tokens into one coherent auditory object) decreases as ISI increases (van Noorden, 1975); therefore, if the cost of listening to multiple talkers is incurred by a disruption of streaming (and attentional focus), we would instead expect greater processing costs at faster speech rates, when streaming of single-talker speech is strongest and the effect of a disruption of streaming the greatest.

[1] It is a matter of debate whether the psychological construct assessed by forward digit span tasks is better operationalized as *short-term memory* instead of *working memory*, the latter of which reflects the additional cognitive demands of sustained attention to short-term memory representations (e.g., Conway et al., 2002; Cowan, 2008; Engle et al., 1999; Kane et al., 2001). However, unlike conventional digit span tasks that involve immediate recall, the current task imposes an additional 5-s retention period, during which participants must actively maintain the memory representations prior to recall. Thus, while not requiring executive manipulation of the short-term memory store, as in backwards digit span or sequencing tasks, the current task does introduce additional demands for focused attention to short-term memory representations, as in working memory tasks. Furthermore, the principal goal of the current study was to investigate the *nature* of the mental representations of speech maintained in short-term memory storage, which is inherent to both working- and short-term memory systems (Engle et al., 1999; Cowan, 2008). Thus, here we operationalize our task as one of *working memory*, with the caveat that it is, at least, an instrument for measuring representations in the short-term memory system.

Methods

Participants

Twenty-seven native English-speaking listeners (19 female; mean age = 21.2 ± 2.45, 18–29 years) with normal hearing participated in the study. The sample size was determined based on the number of permutations needed to counterbalance experimental conditions. All participants were recruited through the Boston University online job advertisement system. All experimental procedures were approved by the Boston University Institute Review Board. Participants provided written informed consent and were compensated (US$15/h) for participating.

Stimuli

Natural recordings of the digits 1–9 produced by eight native American-English speakers (four female and four male, one token of each digit by each talker) were used in the study (Fig. 1). Digits were recorded in a sound-attenuated chamber at a sampling rate of 48 kHz at 16-bit resolution. In order to prevent temporal asynchrony in processing digit sequences, each digit recording was resynthesized to 550 ms in duration, using the *pitch-synchronous overlap and add* (PSOLA; Moulines & Charpentier, 1990) algorithm in Praat to maintain natural sound quality. Linear 10-ms onset and 30-ms offset ramps were applied to all stimuli, and stimuli were normalized to equivalent root-mean-squared amplitude (65 dB SPL).[2]

The digit sequence for each trial was constructed by concatenating recordings of seven digits. The sequence construction was pseudorandom; digits appeared in any position in the sequence with equal probability, with the constraint that no adjacent repetitions of the same digit appeared in any sequence.

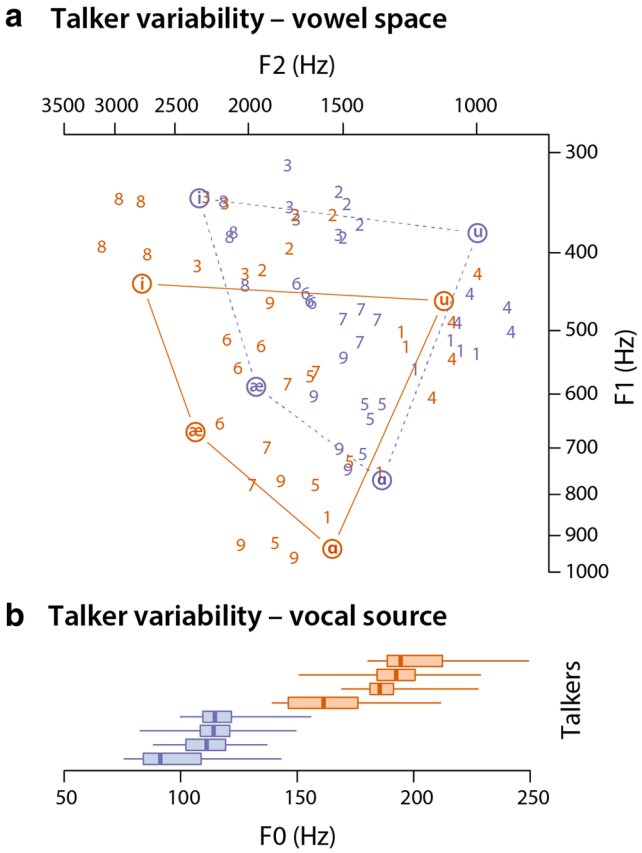

Fig. 1 Acoustic variability of the spoken digit stimuli. (**A**) The vowel space (first and second formants; F1 × F2; note log scale) of each stimulus. Each point indicates the mean F1 and F2 of the sonorant portion of each stimulus. The digit identity of a stimulus is marked as a number (e.g., 9 = "nine"). Orange and blue indicate the four female and four male talkers, respectively. For reference, the acoustic measurements of the current stimuli are situated against the canonical acoustics of the four English point vowels (circled vowels: /i/, /u/, /æ/, and /ɑ/) reported by Hillenbrand et al. (1995). (**B**) The distribution of vocal pitch (fundamental frequency; F0) of each talker's digit recordings. F0 was sampled at every 15 ms of the sonorant portion of each stimulus. The median, interquartile range, and extrema of the F0 of each talker are displayed

Experimental task and procedure

Participants performed a delayed digit sequence recall task (Fig. 2) in a 2 × 3 design manipulating conditions of *talker* (single talker vs. multiple talkers) and stimulus *presentation rate* (0-, 200-, and 500-ms ISIs) during digit sequence encoding. On each trial, participants heard a sequence of seven randomly selected digits, either all spoken by a single talker or each digit spoken by a different, random talker (i.e., no repeated talkers in a multi-talker sequence). After a 5-s retention period, participants recalled the sequence in the order of its presentation during encoding. A number pad appeared on the computer screen following the 5-s retention, and participants used a computer mouse to select the digits in order. Throughout the digit encoding and retention periods, participants fixated on a center dot on the screen. Participants were instructed not to speak the digits out loud.

Participants completed six blocks of 24 trials each. In each block, there was an equal number of trials with digit sequences spoken by a single talker or multiple talkers. The order of trials was semi-randomized; three trials of the same talker condition (e.g., three single-talker trials) were presented in a row. To facilitate the effect of talker continuity, the same talker produced digits in each of the three consecutive single-talker trials. Within each block, digit sequences were presented with the same ISI, and the order of ISI blocks was counterbalanced across participants using Latin-square permutations.

Prior to the start of the experiment, participants completed a brief practice session (five trials) in which the digit sequence was spoken by one male talker; this talker's speech was used in only the practice session, and never presented during the

[2] All stimuli are available at https://open.bu.edu/handle/2144/16460

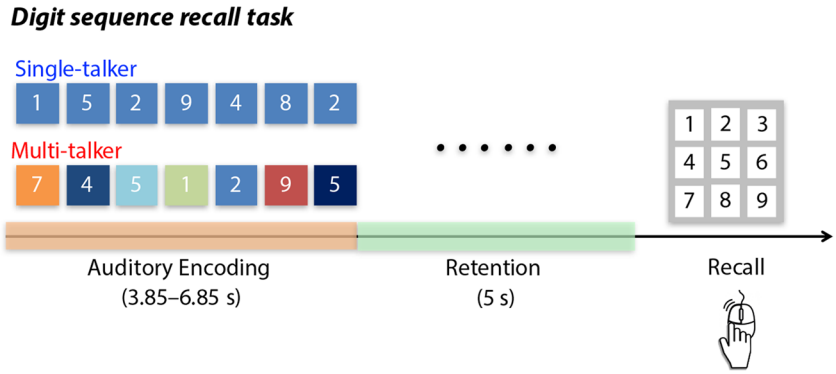

Fig. 2 Illustration of the digit sequence recall task. Participants heard digit sequences spoken either by a single talker or with each digit in the sequence spoken by a different talker (i.e., seven talkers). Each digit sequence was presented at a rate of 0-, 200-, or 500-ms inter-digit stimulus delay. After a 5-s retention period, participants recalled the digit sequence, in order, using a mouse to select digits from a visual display

main experimental task. Throughout the experiment, the number of presentations of each stimulus (i.e., each talker's spoken digits) was balanced across both talker conditions. The experiment was controlled by Psychtoolbox-3 (MATLAB). No feedback was provided. The experiment was conducted in a sound-attenuated booth, and sounds were delivered through Sennheiser HD-380 pro headphones.

Data analysis

Accuracy was determined based on participants' recall of the correct digit at each position in the sequence on each trial. Response time (RT) was quantified as the time delay between the appearance of the number pad screen and participants' selection of the first digit in the sequence. For calculating RT, any trial in which either participants' first digit response was incorrect or their log-transformed RT was greater than 3 standard deviations from their mean in that condition was excluded. Prior to statistical analysis, RT data were log-transformed to ensure normality.

The main interest of the current study was to examine whether talker variability disrupts recall of speech from memory and, if it does, how this cost varies with stimulus presentation rate. To this end, we analyzed the recall accuracy and onset RT using logistic and linear mixed-effects models, respectively, implemented in *lme4* in R (v3.3.3). Fixed factors included *talker* (single- vs. multi-talkers) and *stimulus rate* (0-, 200-, and 500-ms ISIs), with participants as a random factor. An additional fixed factor, digit position (1–7), was included in analyzing recall accuracy using the logistic mixed-effects model. We used forward iterative model comparisons; starting from a null model (i.e., an intercept), we added fixed and interaction effects, as well as subject-wise random slopes of the fixed factors, in a stepwise manner. Model fitting was assessed based on maximum likelihood estimation, and models were compared based on log-likelihood ratio tests (Chambers & Hastie, 1992). Any significant effects found in the best model were followed by *post hoc* testing using differences of least-squares means (*lsmeans* in R).

To gain further insight into the results, we tested the effects of talker and ISI conditions on participants' overall *memory recall efficiency*. The efficiency measure is a composite score, combining both the accuracy and RT measures (cf. inverse efficiency, adapted from Bruyer & Brysbaert, 2011; Townsend & Ashby, 1978, 1983). The efficiency measure is computed as an individual's average recall accuracy divided by their mean RT (i.e., proportion correct / RT) across trials. We conducted a repeated-measures ANOVA on the log-transformed recall efficiency with two within-subject factors: *talker* (single vs. multiple talkers) and *stimulus rate* (0-, 200-, and 500-ms ISIs). Any significant effects were followed up by *post hoc* paired-samples *t*-tests.

Results

Figure 3 illustrates the average accuracy of recall performance by digit position across levels of the talker and ISI conditions. As commonly observed in the serial memory recall tasks, the logistic mixed-effects model analysis revealed a robust main effect of digit position ($\chi^2(1) = 7.64$, $p = 0.0057$): accuracy was highest for recalling digits in the initial and final positions of the sequence (i.e., primacy and recency effects; Fig. 3A).

Analysis also revealed a significant main effect of talker ($\chi^2(1) = 18.015$, $p = 2.2 \times 10^{-5}$), but a marginal effect of ISI ($\chi^2(1) = 3.087$, $p = 0.079$). As shown in Fig. 3B and C, participants recalled the digits less accurately when the digit sequence was spoken by multiple talkers than by a single talker ($M_{diff} = -3.77$; $t_{26} = 4.13$, $p = 0.00034$) consistently across ISIs. However, addition of any interactions among the fixed factors did not improve the model fit (all $ps > 0.40$).

We tested whether participants' RTs differed between the single versus multiple talker conditions as a function of the

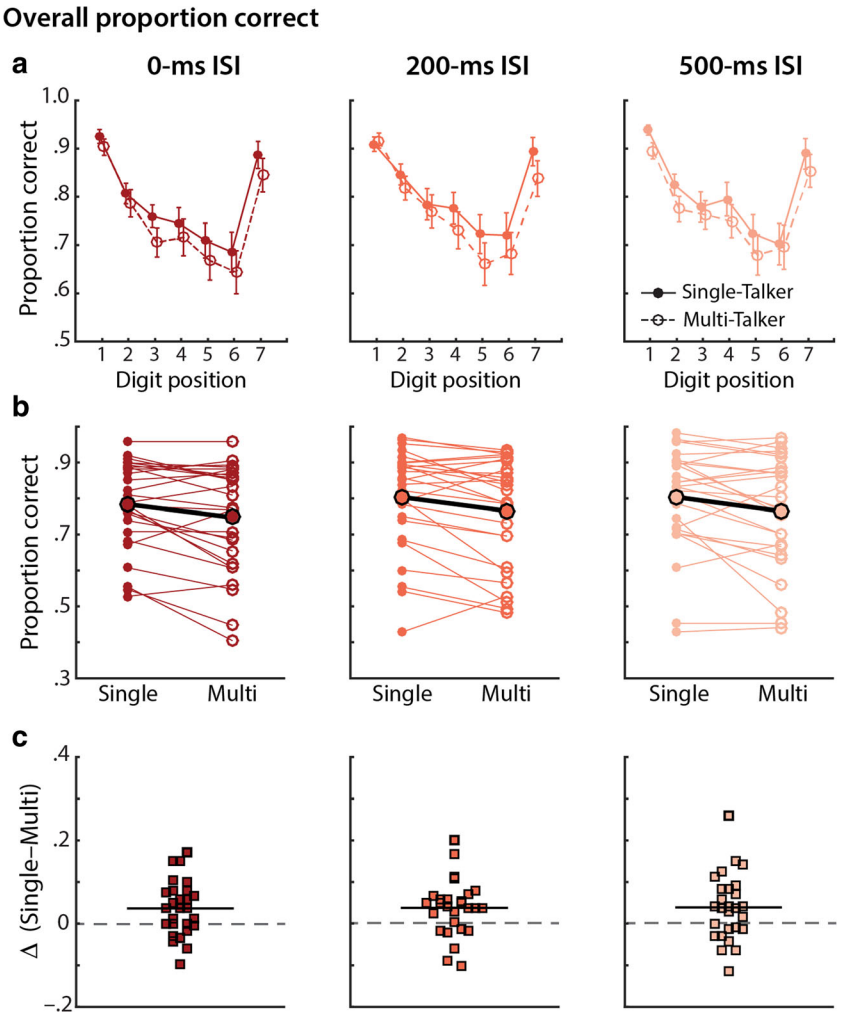

Fig. 3 Average proportions of correctly recalled digits in each of 2 (talkers) × 3 (stimulus rate; ISI) conditions. (**A**) The mean proportion of correctly recalled digits as a function of digit position. Error bars indicate the ±1 standard error of the mean (SEM) across participants. (**B**) Mean proportion of correct digit recalls by talker condition. The thin colored lines connect each individual participant's performance in the single- and multi-talker conditions. Group mean performance is indicated by the black circles connected by a bold line. (**C**) Dot density plots of individual participant's differences (single- vs. multi-talker) in mean proportion correct in each ISI condition. The solid line indicates the mean difference across participants

stimulus presentation rate (ISI). The linear mixed-effects modeling analysis of the log-transformed RTs revealed a significant talker × ISI interaction ($\chi^2(1) = 4.59$, $p = 0.032$) and a main effect of talker ($\chi^2(1) = 5.89$, $p = 0.015$), but a marginal effect of ISI ($\chi^2(1) = 2.98$, $p = 0.084$). As illustrated in Fig. 4, we found that participants recalled digit sequences significantly slower when they were spoken by multiple talkers than by a single talker (M_{diff} = 48.66 ms). However, recalling digits spoken by multiple talkers was slower than for a single talker only when the digits were presented at the fastest rate (0-ms ISI: M_{diff} = 128.99 ms; $t_{26} = 2.96$; $p = 0.0065$); the talker effect did not hold for rates of 200-ms ISI (M_{diff} = 36.62 ms, t_{26} = 1.04; $p = 0.31$) or 500-ms ISI (M_{diff} = –19.63 ms, $t_{26} = 0.044$, $p = 0.97$) (Fig. 4B).

Next, we examined the effect of talker as a function of ISI on the overall memory recall efficiency measure – a composite score of recall accuracy and RT (Fig. 5). A 2 (single vs. multiple *talkers*) × 3 (0-, 200-, 500-ms *ISIs*) repeated-measures ANOVA on the log-transformed efficiency measure revealed a significant effect of talker ($F_{1,26}$ = 16.25, p = 0.00043, η^2_p = 0.38) and a significant talker × ISI interaction ($F_{1,26}$ = 5.41, p = 0.028, η^2_p = 0.17), but no significant effect of ISI ($F_{1,26}$ = 0.018, p = 0.89; η^2_p = 0.00071). *Post hoc* t-tests revealed that recalling digits spoken by multiple talkers compared to a single talker was less efficient, especially when digits were presented at a fast rate (0-ms ISI: M_{diff} = –0.15, t_{26} = 3.50, p = 0.0017). However, the difference in recall efficiency caused by multiple talkers decreased as a function of presentation rates (Fig. 5B; 200-ms ISI: M_{diff} = –0.083, t_{26} = 2.10, p = 0.046; 500-ms ISI: M_{diff} = –0.042, t_{26} = 1.35, p = 0.19).

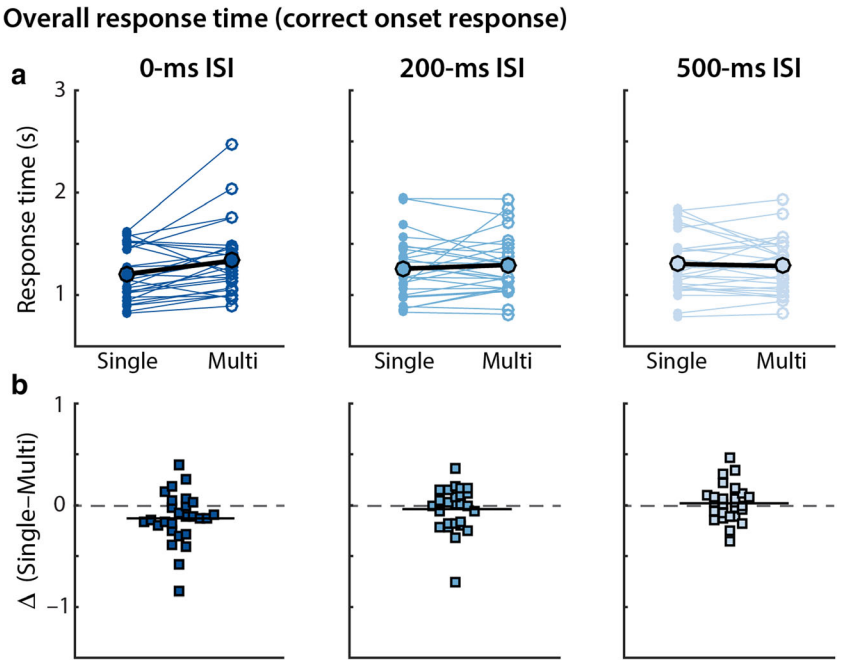

Fig. 4 Response times of the onset of digit sequence recall by talker condition across stimulus presentation rates (ISIs). (**A**) Thin colored lines connect each individual participant's performance in the single- and multi-talker conditions. Group means across participants are indicated by the black circles connected by bold lines. (**B**) Dot density plots of individuals' differences (single–multi-talker) in mean response time across ISIs. The solid line indicates the mean difference (single–multi-talker) across participants

Discussion

Do discontinuities in the talker's voice during speech encoding influence how well listeners process and store speech information in working memory? And does any such effect of talker discontinuity depend on the rate of incoming speech? We investigated these questions by using a delayed-recall of digit span task, in which

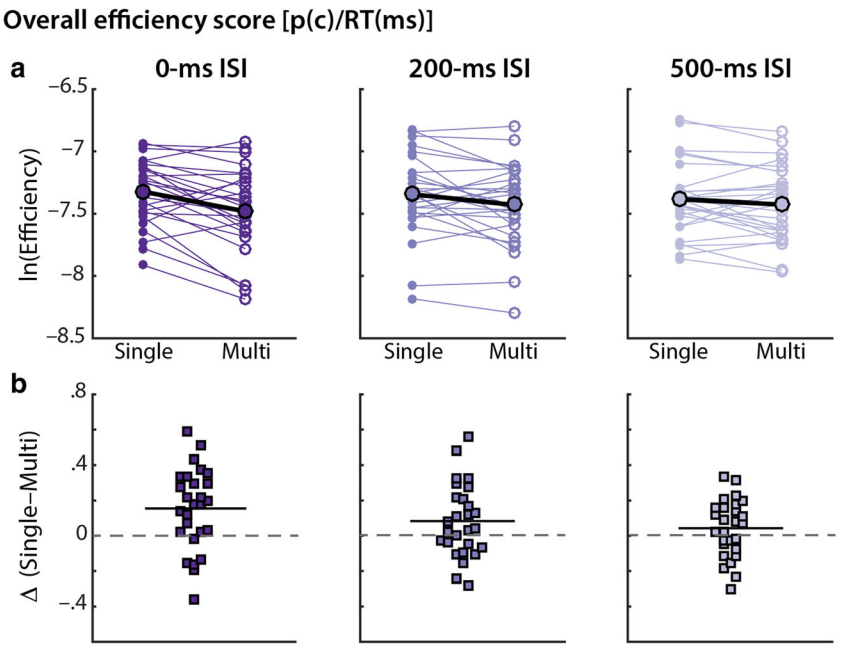

Fig. 5 Efficiency of digit sequence recall by talker condition across stimulus presentation rates (ISIs). (**A**) Mean log-transformed efficiency score of the digit sequence recall. Data points connected by thin colored lines indicate individual participants' performance in the single- vs. multi-talker conditions in each ISI condition. Group means across participants are indicated by the black circles connected by bold lines. (**B**) Dot density plots of the individuals' differences (single–multi-talker) in average recall efficiency in each ISI condition. The solid line indicates the mean difference (single–multi-talker) across participants

participants maintained and recalled a serial order of digits spoken by one continuous talker or multiple talkers at three different speech rates.

We found that talker variability in speech interfered with serial recall of working memory. Compared to a sequence of digits spoken by one talker, listeners were less accurate and slower – thereby, less efficient – in memory recall for digits spoken by multiple talkers. Moreover, temporal demands on encoding (speech rate) impacted the extent to which talker discontinuity disrupted listeners' recall speed, hence efficient working memory for speech. For the fastest speech rate, the disruption of memory recall was greatest (Fig. 5). Conversely, when listeners were given additional time to process and encode each speech token, talker variability had a smaller effect on memory recall efficiency. Thus, our results suggest that speech at faster rates leads to greater interference from talker variability on recalling speech information in working memory.

Interestingly, this finding is in contrast to what would be predicted by a framework that emphasizes the role of timing in determining the representational detail of encoded speech (Luce et al., 2003; McLennan & Luce, 2005). According to this view, encoding and retrieving fine-grained details of speech, such as talker-specific features, require additional time beyond that devoted to parsing the abstract lexical or phonological content. As such, slower processing time for speech (including due to greater task difficulty) amplifies talker-specificity effects in long-term memory recognition tasks (Mattys & Liss, 2008; McLennan & González, 2012; McLennan & Luce, 2005). Conversely, faster timing is thought to prevent encoding, and thus retrieval, of fine-grained acoustic details, thereby emphasizing processing of abstract speech representations, and reducing talker-specificity effects. However, this view is also inconsistent with prior findings that demonstrated robust talker variability/specificity effects on immediate speech recognition, regardless of timing and task difficulty; even with fast speech presentation rates and under relatively low task difficulty, talker variability interferes with speech perception as well as memory recall (e.g., Choi et al., 2018; Goldinger et al., 1991).

A recent study has suggested that rather than timing per se, explicit attention during speech processing might be important in the encoding and retrieval of talker-specific information from speech. Theodore et al. (2015) recently showed that a talker-specificity effect emerged during speech recognition only when listeners' attention was directed to indexical speech features during encoding. Nonetheless, talker variability has repeatedly been shown to influence how rapidly listeners can recognize spoken words, even without directed attention to talker (e.g., Choi et al., 2018; Mullennix & Pisoni, 1990; *inter alia*). In the present study, listeners' working memory recall was also affected by talker variability even though this task did not require attention to talker identity during speech processing.

There are two potential explanations of how talker variability can disrupt working memory for speech as shown in the present study. One possibility is that working memory representations of speech are not maintained as purely abstracted phonological/lexical items independent from sensory-specific processing (cf. Baddeley, 1992, 2003), but rather that these working memory representations retain the indexical variability of speech encountered during encoding. That is, as stimulus-specific details of speech are internally represented in memory (Lim et al., 2015, 2018), speech spoken by one consistent talker might be more efficiently represented in working memory than speech spoken by different talkers, even if the speech carries the same linguistic message (see below for a discussion of potential differences in memory representations). This possibility is consistent with an emerging neurobiological model of speech working memory, which posits that maintenance of speech information in memory relies on the very same brain networks involved in speech perception and production (Hickok, 2009; Jacquemot & Scott, 2006; Perrachione, Ghosh, Ostrovskaya, Gabrieli, & Kovelman, 2017). Thus, the speech working memory store might remain susceptible to variability-related interference such as that observed during immediate recognition.

The other intriguing possibility is that discontinuities in talkers disrupts listeners' attentional focus during speech processing (e.g., Bressler et al., 2014; Darwin & Hukin, 2000; Shinn-Cunningham, 2008). Such disruption in turn would directly degrade the fidelity of speech encoding, so that working memory would inherit a degraded representation of speech. Of course, these two possibilities are not mutually exclusive; degradations in memory performance may be due to disruptions in speech encoding, to inefficient maintenance of working memory representations of speech, or to both. Our data suggest that talker variability does more than just disrupt speech encoding, as we still observe a significant effect of talker variability in recall accuracy even at the slowest speech rate (i.e., 500 ms), when listeners have ample time to encode and/or switch attention to each speech token (Fig. 3, right; $\chi^2(1) = 6.48$, $p = 0.011$).

Our results are in line with the idea that talker continuity contributes to binding discrete speech tokens (here, spoken digits) into a coherent auditory object (Bregman, 1990; Shinn-Cunningham, 2008). In particular, when speech tokens are encountered close in time, talker continuity in speech can give rise to the perception of a

continuous auditory stream (Best et al., 2008; Bressler et al., 2014). The advantage of auditory streaming for speech perception is that an entire stream is processed as a single perceptual object (Joseph, Kumar, Husain, & Griffiths, 2015; Macken, Tremblay, Houghton, Nicholls, & Jones, 2003; Maddox & Shinn-Cunningham, 2012; Mathias & Kriegstein, 2014; Shinn-Cunningham, 2008; Sussman et al., 2007). Within psychoacoustics, it is well accepted that the sequence order of items within one stream are stored as part of the identity of the single auditory object (Bizley & Cohen, 2013; Griffiths & Warren, 2004); however, if streaming breaks down and tokens are perceived to be, and stored in memory as, distinct streams, it becomes difficult to judge the temporal order of the items (Bregman & Campbell, 1971; Vliegen, Moore, & Oxenham, 1999). Temporal gaps between speech tokens break down automatic streaming (Best et al., 2008; Bressler et al., 2014), so that individual tokens are processed more independently, even if spoken by the same talker. Thus, it may be that disruptions of streaming – either by switches in talkers or longer silent delays – reduce processing efficiency and impair the ability to recall a sequence of speech tokens in order, as the speech tokens are maintained in working memory as multiple, separate auditory objects, rather than one coherent object.

Parsing speech has been shown to rely upon the acoustic regularities of the preceding speech context (Evans and Iverson, 2004; Huang & Holt, 2012; Kleinschmidt & Jaeger, 2015; Ladefoged & Broadbent, 1957; Liberman et al., 1956; Mann, 1986; see Heald, van Hedger, & Nusbaum, 2017 for a review). The facilitation afforded to speech processing by a continuous talker has been previously attributed to biasing effects of this preceding context. That is, listeners can build and use a talker-specific acoustic-to-phonemic mapping from the immediately preceding speech of a talker in order to reduce the computational demands in disambiguating subsequent speech from the same talker (e.g., Kleinschmidt & Jaeger, 2015; Magnuson & Nusbaum, 2007; Nusbaum & Magnuson, 1997). While context precedence may explain a general facilitation afforded by talker continuity, it cannot fully account for our result that there are differences in talker-specificity effects across speech-processing timing. Under a model in which only preceding context matters, speech from one talker should be uniformly beneficial compared to speech from multiple talkers, regardless of the presentation rate, because the amount of preceding speech from a specific talker is equal when subsequent speech tokens are encountered. However, by showing the largest amount of facilitation from talker continuity at the fastest speech rate, our results extend and refine this view of context-specific speech processing by providing a novel, mechanistic insight into how talker continuity facilitates perception and memory for connected speech. Here, we propose that talker continuity-related facilitation – and thereby the context precedence effect – can be understood as a process of auditory streaming and attention. That is, both featural (i.e., source) and temporal continuities in speech sounds may be crucial in forming a coherent auditory stream that binds preceding context and subsequent speech into a coherent auditory object. Streaming may facilitate speech processing by allowing listeners to efficiently allocate attention to specific acoustic features within a coherent auditory object (without a need to redirect attention to a new stream; Lakatos et al., 2013; Shinn-Cunningham, 2008; Winkler et al., 2009; Woods and McDermott, 2015), allowing for more efficient extraction of the phonemic content of the speech making up that stream. Conversely, in the case of encountering different talkers, attention is disrupted, which interferes with speech processing.

While we have emphasized the bottom-up, stimulus-driven aspects of talker continuity versus discontinuity effects, it is also important to consider the role of top-down processing (i.e., volitional allocation of attention and expectation) in driving this effect. Prior studies have demonstrated that, even with identical speech sounds, perceived discontinuity in speech – manipulated by modulating listeners' expectation that words will be spoken by multiple talkers rather than by a single talker – alone can induce interference effects like those caused by processing speech from truly different talkers (e.g., Magnuson & Nusbaum, 2007). Similarly, in the auditory scene analysis framework, listeners' processing strategies can determine whether they perceive a sequence of discrete acoustic tokens as either one integrated stream or two segregated streams of sounds. But importantly, this line of work also emphasizes how temporal proximity among the discrete sounds plays an important role in grouping sounds into one integrated auditory stream (e.g., Bregman, 1990; van Noorden 1975). This pattern indicates that bottom-up, temporal contiguity can constrain the top-down strategy of processing multiple auditory events. Although the current study did not systematically bias listeners' expectations about the talkers in the speech sequence (e.g., Magnuson & Nusbaum, 2007), it would be of interest for future studies to examine the influence of prior expectation of the perceived continuity versus discontinuity in talkers on the emergence of auditory streaming, and its cascading impact on speech processing efficiency and memory representations for speech.

It is of note that behavioral measurements of working memory performance alone, as in the present study, may

not unambiguously dissociate exactly when and how variability-related interference effects arise. Such questions could be addressed by measuring neurophysiological responses (e.g., electro-/magneto-encephalography) and by employing neuroimaging techniques (e.g., fMRI). These techniques will elucidate the neural responses and brain regions involved separately in the speech-encoding and memory-retention stages of task performance (e.g., Lim et al., 2015; Wöstmann, Lim, & Obleser, 2017). For instance, prior fMRI studies demonstrated that recognition of words spoken by multiple talkers compared to a single talker was manifested as greater recruitment of cortical regions responsible for speech and language processing (Chandrasekaran, Chan, & Wong, 2011; Perrachione et al., 2016), as well as higher-order cognitive regions associated with selective attention (Wong, Nusbaum, & Small, 2004). Extending this line of research in future studies employing these techniques will be necessary to investigate how and whether talker variability distinctly influences speech encoding and memory maintenance phases.

In conclusion, talker variability in speech affects cognitive operations ranging from perception and attention to memory maintenance (see Heald & Nusbaum, 2014, for a review). Compared to processing a single talker's speech, processing multiple talkers' speech creates more opportunity for interference, which has a cascading influence on both accuracy and efficiency of memory processes. Many prior studies have asserted that talker variability increases the computational demands on listeners for resolving ambiguity in mapping between acoustics and linguistic representations during perception, and to recalibrate their perception to talker-specific acoustics (Antoniou & Wong, 2015; Martin et al., 1989; Mullennix & Pisoni, 1990; Nusbaum & Magnuson, 1997; Nusbaum & Morin, 1992; Perrachione, Lee, Ha, & Wong, 2011). The current study provides evidence supporting an alternative underlying mechanism through which talker variability interferes speech processing. Under an auditory attention framework a change in talker disrupts listeners' attentional processing, impairing their ability to form a coherent auditory stream, which ultimately interferes with the efficient formation of auditory objects to be maintained in working memory.

Acknowledgements This work was supported by NIH grant R03DC014045 and a Brain and Behavioral Research Foundation NARSAD Young Investigator grant to TKP and NIH grant R01DC009477 to BGSC. SJL was supported by NIH training grant T32DC013017. We thank Yaminah Carter for her assistance.

References

Antoniou, M., & Wong, P. C. M. (2015). Poor phonetic perceivers are affected by cognitive load when resolving talker variability. *The Journal of the Acoustical Society of America*, *138*(2), 571–574. https://doi.org/10.1121/1.4923362

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Baddeley, A. (2003). Working memory: looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Best, V., Ozmeral, E. J., Kopčo, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences*, *105*(35), 13174–13178. https://doi.org/10.1073/pnas.0803718105

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, *14*(10), 693–707. https://doi.org/10.1038/nrn3565

Bradlow, A. R., Nygaard, L. C., & Pisoni, D. B. (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, *61*(2), 206–219. https://doi.org/10.3758/BF03206883

Bregman, A. S. (1990). Auditory scene analysis. Cambridge, MA: MIT Press.

Bregman, A. S., & Campbell, J. C. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, *89*(2), 244–249.

Bressler, S., Masud, S., Bharadwaj, H., & Shinn-Cunningham, B. (2014). Bottom-up influences of voice continuity in focusing selective auditory attention. *Psychological Research*, *78*(3), 349–360. https://doi.org/10.1007/s00426-014-0555-7

Bruyer, R., & Brysbaert, M. (2011). Combining speed and accuracy in cognitive psychology: Is the inverse efficiency score (IES) a better dependent variable than the mean reaction time (RT) and the percentage of errors (PE)? *Psychologica Belgica*, *5*(1), 5–13.

Chambers, J. M., & Hastie, T. J. (1992). Statistical models in S. Pacific Grove, CA: Wadsworth.

Chandrasekaran, B., Chan, A., & Wong, P. C. M. (2011). Neural processing of what and who information in speech. *Journal of Cognitive Neuroscience*, *23*(10), 2690–2700.

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, & Psychophysics*, *80*, 784–797.

Conway A. R. A., Cowan N., Bunting M. F., Therriault D. J., Minkoff S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*, 163–184.

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research, 169*, 323–338.

Craik, F. I. M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *The Quarterly Journal of Experimental Psychology*, *26*(2), 274–284. https://doi.org/10.1080/14640747408400413

Darwin, C. J., & Carlyon, R. P. (1995). Auditory grouping. In B. C. Moore (Ed.), Hearing handbook of perception and cognition (pp. 387–424). Elsevier. https://doi.org/10.1016/B978-012505626-7/50013-3

Darwin, C. J., & Hukin, R. W. (2000). Effectiveness of spatial cues, prosody, and talker characteristics in selective attention. *The Journal of the Acoustical Society of America*, *107*(2), 970–977. https://doi.org/10.1121/1.428278

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid

intelligence: a latent-variable approach. *Journal of Experimental Psychology: General*, *128*(3), 309-331.

Evans B. G., & Iverson, P. (2004). Vowel normalization for accent: An investigation of best exemplar locations in northern and southern British English sentences. *The Journal of the Acoustical Society of America, 115* (1), 352-361.

Geiselman, R. E., & Bellezza, F. S. (1977). Incidental retention of speaker's voice. *Memory & Cognition*, *5*(6), 658–665. https://doi.org/10.3758/BF03197412

Goldinger, S. D., Pisoni, D. B., & Logan, J. S. (1991). On the nature of talker variability effects on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(1), 152–162.

Green, K. P., Tomiak, G. R., & Kuhl, P. K. (1997). The encoding of rate and talker information during phonetic perception. *Perception and Psychophysics, 59* (5), 675-692.

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, *5*, 887–892.

Heald, S. L. M., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, 1–15. https://doi.org/10.3389/fnsys.2014.00035/abstract

Heald, S. L.M., Van Hedger, S. C., & Nusbaum, H. C. (2017). Perceptual plasticity for auditory object recognition. *Frontiers in Psychology, 8*: 781. https://doi.org/10.3389/fpsyg.2017.00781

Hickok, G. (2009). The functional neuroanatomy of language. *Physics of Life Reviews*, *6*(3), 121–143. https://doi.org/10.1016/j.plrev.2009.06.001

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America*, *97*(5), 3099–3111. https://doi.org/10.1121/1.411872

Huang, J., and Holt, L. L. (2012). Listening for the norm: adaptive coding in speech categorization. *Frontiers in Psychology, 3*: 10.

Jacquemot, C., & Scott, S. K. (2006). What is the relationship between phonological short-term memory and speech processing? *Trends in Cognitive Sciences*, *10*(11), 480–486. https://doi.org/10.1016/j.tics.2006.09.002

Johnson, K., Strand, E. A., and D'Imperio, M. (1999). Auditory–visual integration of talker gender in vowel perception. *Journal of Phonetics, 27*(4), 359-384.

Joseph, S., Kumar, S., Husain, M., & Griffiths, T. D. (2015). Auditory working memory for objects vs. features. *Frontiers in Neuroscience*, *9*, 20738. https://doi.org/10.3389/fnins.2015.00013

Kane, M. J., Bleckley, M. K., Conway, A. R., Engle, R. W. (2001). A controlled-attention view of working-memory capacity. *Journal of Experimental Psychology: General*, 130(2), 169-183.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148.

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

Lakatos, P., Musacchia, G., O'Connel, M. N., Falchier, A. Y., Javitt, D. C., & Schroeder, C. E. (2013). The spectrotemporal filter mechanism of auditory selective attention. *Neuron, 77*(4), 750–761.

Liberman, A. M., Delattre, P. C., Gerstman, L. J., and Cooper, F. S. (1956). Tempo of frequency change as a cue for distinguishing classes of speech sounds. *Journal of Experimental Psychology, 52*(2): 127-37.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431–461. https://doi.org/10.1037/h0020279

Lim, S.-J., Wöstmann, M., & Obleser, J. (2015). Selective Attention to Auditory Memory Neurally Enhances Perceptual Precision. *The Journal of Neuroscience*, *35*(49), 16094–16104. https://doi.org/10.1523/JNEUROSCI.2674-15.2015

Lim, S.-J., Wöstmann, M., Geweke, F., & Obleser, J. (2018). The benefit of attention-to-memory depends on the interplay of memory capacity and memory load. *Frontiers in Psychology*, *9*, 146. https://doi.org/10.3389/fpsyg.2018.00184

Luce, P. A., & McLennan, C. T. (2005). Spoken word recognition: The challenge of variation. In D. B. Pisoni & R. E. Remez (Eds.), Handbook of speech perception (pp. 591–609). Maldon, MA: Blackwell.

Macken, W. J., Tremblay, S., Houghton, R., Nicholls, A. P., & Jones, D. M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of Experimental Psychology. Human Perception and Performance*, *29*(1), 43–51.

Maddox, R. K., & Shinn-Cunningham, B. G. (2012). Influence of task-relevant and task-irrelevant feature continuity on selective auditory attention. *Journal of the Association for Research in Otolaryngology*, *13*(1), 119–129. https://doi.org/10.1007/s10162-011-0299-7

Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic differences, listener expectations, and the perceptual accommodation of talker variability. *Journal of Experimental Psychology. Human Perception and Performance*, *33*(2), 391–409. https://doi.org/10.1037/0096-1523.33.2.391

Mann, V. A. (1986). Distinguishing universal and language-dependent levels of speech perception: evidence from Japanese listeners' perception of English "l" and "r". *Cognition, 24*(3), 169-196.

Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. V. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(4), 676–684.

Mathias, S. R., & Kriegstein, von, K. (2014). Percepts, not acoustic properties, are the units of auditory short-term memory. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(2), 445–450.

Mattys, S. L., & Liss, J. M. (2008). On building models of spoken-word recognition: When there is as much to learn from natural "oddities" as artificial normality. *Perception & Psychophysics*, *70*(7), 1235–1242. https://doi.org/10.3758/PP.70.7.1235

McLennan, C. T., & González, J. (2012). Examining talker effects in the perception of native- and foreign-accented speech. *Attention, Perception, & Psychophysics*, *74*(5), 824–830. https://doi.org/10.3758/s13414-012-0315-y

McLennan, C. T., & Luce, P. A. (2005). Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 306–321.

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5-6), 453–467. https://doi.org/10.1016/0167-6393(90)90021-Z

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. https://doi.org/10.1121/1.397688

Nearey, T. M. (1998). Static, dynamic, and relational properties in vowel perception. *The Journal of the Acoustical Society of America*, *85*(5), 2088–2113. https://doi.org/10.1121/1.397861

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Langauge and Social Psychology*, *18* (1), 62–85.

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 (pp. 109–132). New York, NY: Academic Press

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* (pp. 113–134). Tokyo.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, *60*(3), 355–376.

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1994). Speech perception as a talker-contingent process. *Psychological Science*, *5*(1), 42–46. https://doi.org/10.1111/j.1467-9280.1994.tb00612.x

Nygaard, L. C., Sommers, M. S., & Pisoni, D. B. (1995). Effects of stimulus variability on perception and representation of spoken words in memory. *Perception & Psychophysics*, *57*(7), 989–1001. https://doi.org/10.3758/BF03205458

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. https://doi.org/10.1037//0278-7393.19.2.309

Perrachione, T. K., Del Tufo, S. N., Winter, R., Murtagh, J., Cyr, A., Chang, P., et al. (2016). Dysfunction of rapid neural adaptation in dyslexia. *Neuron*, *92*(6), 1383–1397. https://doi.org/10.1016/j.neuron.2016.11.020

Perrachione, T. K., Ghosh, S. S., Ostrovskaya, I., Gabrieli, J. D. E., & Kovelman, I. (2017). Phonological working memory for words and nonwords in cerebral cortex. *Journal of Speech, Language, and Hearing Research*, *60*(7), 1959–1979. https://doi.org/10.1044/2017_JSLHR-L-15-0446

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & Wong, P. C. M. (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*(1), 461–472. https://doi.org/10.1121/1.3593366

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. https://doi.org/10.1121/1.1906875

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Souza, P., Gehani, N., Wright, R., & McCloy, D. (2013). The advantage of knowing the talker. *Journal of the American Academy of Audiology*, *24*(8), 689–700. https://doi.org/10.3766/jaaa.24.8.6

Sussman, E. S., Horváth, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception & Psychophysics*, *69*(1), 136–152. https://doi.org/10.3758/BF03194460

Theodore, R. M., Blumstein, S. E., & Luthra, S. (2015). Attention modulates specificity effects in spoken word recognition: Challenges to the time-course hypothesis. *Attention, Perception, & Psychophysics*, *77*(5), 1674–1684. https://doi.org/10.3758/s13414-015-0854-0

Townsend, J. T., & Ashby, F. G. (1978). Methods of modeling capacity in simple processing systems. In J. Castellan & F. Restle (Eds.), Cognitive theory (Vol. 3, pp. 200–239). Hillsdale, NJ: Erlbaum.

Townsend, J. T., & Ashby, F. G. (1983). Stochastic modeling of elementary psychological processes. Cambridge: Cambridge University Press.

van Noorden, L. P. A. S. (1975). Temporal coherence in the perception of tone sequences (Vol. 3, pp. 1–129). Eindhoven, The Netherlands: Institute for Perceptual Research. https://doi.org/10.6100/IR152538

Vliegen, J., Moore, B. C. J., & Oxenham, A. J. (1999). The role of spectral and periodicity cues in auditory stream segregation, measured using a temporal discrimination task. *The Journal of the Acoustical Society of America*, *106*(2), 938–945. https://doi.org/10.1121/1.427140

Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: Predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences*, *13*(12), 532-40.

Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural bases of talker normalization. *Journal of Cognitive Neuroscience*, *16*(7), 1–13.

Woods, K. J. P., & McDermott, J. H. (2015). Attentive tracking of sound sources. *Current Biology*, *25*(17), 2238–2246.

Wöstmann, M., Lim, S.-J., & Obleser, J. (2017). The human neural alpha response to speech is a proxy of attentional control. *Cerebral Cortex*, *27*(6), 3307–3317. https://doi.org/10.1093/cercor/bhx074