Brain & Language 221 (2021) 104996

Contents lists available at ScienceDirect

Brain and Language

journal homepage: www.elsevier.com/locate/b&l

Talker discontinuity disrupts attention to speech: Evidence from EEG and pupillometry

Sung-Joo Lim [a,*], Yaminah D. Carter [a], J. Michelle Njoroge [a], Barbara G. Shinn-Cunningham [b], Tyler K. Perrachione [a,*]

[a] Department of Speech, Language, and Hearing Sciences, Boston University, United States
[b] Neuroscience Institute, Carnegie Mellon University, United States

ARTICLE INFO

Keywords:
Talker variability
Auditory streaming
Attentional reorientation
Pupillometry
EEG
Neural alpha oscillatory power
Speech perception
Working memory

ABSTRACT

Speech is processed less efficiently from discontinuous, mixed talkers than one consistent talker, but little is known about the neural mechanisms for processing talker variability. Here, we measured psychophysiological responses to talker variability using electroencephalography (EEG) and pupillometry while listeners performed a delayed recall of digit span task. Listeners heard and recalled seven-digit sequences with both talker (single- vs. mixed-talker digits) and temporal (0- vs. 500-ms inter-digit intervals) discontinuities. Talker discontinuity reduced serial recall accuracy. Both talker and temporal discontinuities elicited P3a-like neural evoked response, while rapid processing of mixed-talkers' speech led to increased phasic pupil dilation. Furthermore, mixed-talkers' speech produced less alpha oscillatory power during working memory maintenance, but not during speech encoding. Overall, these results are consistent with an auditory attention and streaming framework in which talker discontinuity leads to involuntary, stimulus-driven attentional reorientation to novel speech sources, resulting in the processing interference classically associated with talker variability.

1. Introduction

There is immense variability in the acoustics of speech across talkers and contexts. The lack of direct, one-to-one correspondences between speech acoustics and linguistic units (Hillenbrand et al., 1995) presents a challenge for listeners who must resolve this inherent variability in order to efficiently perceive and remember speech. Speech processing is less efficient when the source of the speech signal is variable: many prior studies show that listeners are slower and less accurate in processing speech spoken by a series of multiple, mixed talkers compared to processing a single, consistent talker's speech (e.g., Choi et al., 2018; Mullennix and Pisoni, 1990; Nusbaum and Morin, 1992). This highly replicated phenomenon has been a driving force behind many contemporary psycholinguistic models of speech perception, which attempt to account for how listeners reliably decode speech signals in the face of acoustic variability across talkers (Pierrehumbert, 2003; Kleinschmidt & Jaeger, 2015).

Compared to the behavioral literature, research into how the brain processes speech variability is in its infancy. Functional brain imaging studies have consistently shown greater neural activation in bilateral superior temporal cortex when listening to speech from mixed talkers compared to a single talker (Belin and Zatorre 2003; Wong et al., 2004; Chandrasekaran et al., 2011; Perrachione et al., 2016). These observations support the idea that processing mixed and discontinuous talkers' speech is less efficient (i.e., there is greater neural activation) than processing a single talker's speech (Wark et al., 2007; Grill-Spector et al., 2006). Despite the observation of greater superior temporal activation to mixed talker speech, there is no consensus as to why mixed-talker speech incurs additional processing costs. Among the prevailing psycholinguistic models that address the challenge of talker variability in speech processing, each posits a distinct cognitive mechanism to account for the added processing demands, including computational complexity (Nearey, 1989; Johnson, 2005), allocation of cognitive resources (Nusbaum and Magnuson, 1997; Magnuson and Nusbaum, 2007), accessing long-term memory representations (Kleinschmidt and Jaeger, 2015), and stimulus-driven allocation of auditory attention (Bressler et al., 2014; Choi and Perrachione, 2019; Kapadia and Perrachione, 2020). However, prior neuroimaging studies were not designed to test the predictions of different psycholinguistic models. Conversely, the psycholinguistic models were derived principally from behavioral data, and

* Corresponding authors.
E-mail addresses: sungjoo@binghamton.edu (S.-J. Lim), tkp@bu.edu (T.K. Perrachione).

https://doi.org/10.1016/j.bandl.2021.104996
Received 29 January 2021; Received in revised form 11 July 2021; Accepted 13 July 2021
Available online 3 August 2021

rarely informed by neural evidence or putative neurocomputational mechanisms.

In this paper, we draw upon a range of psychophysiological techniques to evaluate the neural signatures of talker variability with respect to those predicted by a stimulus-driven auditory attention account of talker-specific speech processing (Choi and Perrachione, 2019; Kapadia and Perrachione, 2020). This framework proposes that the increased processing costs for parsing mixed talkers' speech reflect the cognitive costs involved in switching attention from one auditory source to another (Best et al., 2008; Bressler et al., 2014; Mehraei et al., 2018). Featural discontinuities in the acoustics of an auditory stream—such as those introduced by a switch from one talker to another—disrupt the automatic build-up of a coherent perceptual sound "stream" (Bregman, 1990). Therefore, when listening to mixed talkers' speech, discontinuities in the source of speech (i.e., talker switches) require listeners to reorient their attention to the acoustic features of each newly encountered talker, which adds processing demands, reflected in lower accuracy and increased reaction time for identifying the linguistic content of speech. Briefly, parsing mixed talkers' speech interferes with efficient speech processing because talker discontinuity disrupts listeners' attentional focus, while listening to a coherent speech stream enhances attentional focus (Best et al., 2008, 2018; Bressler et al., 2014; Mehraei et al., 2018).

A growing body of behavioral research supports the view that the processing costs associated with talker variability are better explained by attentional disruption than by phonetic recalibration or increased working- or long term-memory demands. For instance, any discontinuities in talkers—irrespective of the number of talkers or listeners' expectation about the upcoming talker—interfere equally with speech processing efficiency (Carter et al., 2019; Kapadia and Perrachione, 2020). Furthermore, continuity of acoustic features in a stream of events produces an automatic build-up of attention over time (Shinn-Cunningham, 2008; Sussman et al., 2007; Winkler et al., 2009; Best et al., 2018); however, this build-up depends upon the timing of the discrete events. When events are temporally close, streaming of similar events is enhanced, leading to more efficient speech processing (Best et al., 2008; Bressler et al., 2014; Choi et al., 2018; Lim et al., 2019); conversely, interference from mixed talkers' speech is greater with temporally contiguous speech sounds, as talker switches become more disruptive when listeners must quickly reorient their attention to each newly encountered talker (Best et al., 2008; Lim et al., 2019). However, behavioral evidence alone does not reveal whether attention-related brain processes are impacted by talker discontinuity, and how such impact changes with temporal proximity of speech.

Beyond its deleterious effects on the immediate recognition of speech, little is known about how talker variability persists in short-term memory for speech information. Recent work has demonstrated that speech working memory retains not only the abstract content of speech, but also other stimulus-specific details (Lim et al., 2015, 2018, 2021), and that working memory for speech is supported by some of the same neural process responsible for speech perception (Jacquemot and Scott, 2006; Perrachione et al., 2017; Scott and Perrachione, 2019). Thus, the impact of acoustic featural continuity across time may help reveal not only how speech is immediately perceived, but also how it is encoded and maintained in working memory. In particular, if both talker and temporal continuity impact listeners' perception of speech as a coherent stream, discontinuities in these attributes might also interfere with working memory maintenance of speech. While prior behavioral evidence is consistent with this possibility (Martin et al., 1989; Best et al., 2008; Bressler et al., 2014; Lim et al., 2019), disentangling whether discontinuity-related interference arises only during speech encoding, or whether it persists during speech memory retention, is best addressed by physiological, not simply by behavioral, measures.

To evaluate the stimulus-driven attentional streaming hypothesis of processing talker variability, we examined whether the neurophysiological signatures of processing mixed-talker speech are consistent with those associated with disrupting and reorienting listeners' attentional focus. We also explored whether the added costs of processing mixed-talker speech persisted in the neural signatures of working memory maintenance. In order to gain insights into complementary aspects of these underlying cognitive processes, we utilized two independent electrophysiological techniques that are sensitive to different neural processes on different timescales. Specifically, we simultaneously recorded scalp electroencephalography (EEG) and pupillometry to measure several temporally resolved neurophysiological signatures that reflect distinct neural mechanisms for perception and cognition, including cortical evoked potentials, neural oscillatory power, and pupil dilation (see below for details). In the current experiment, participants performed a delayed recall of digit span task that manipulated two conditions: talker continuity in the speech stream (digits spoken by a single vs. mixed talkers) and temporal continuity in the speech stream (digit sequences presented either continuously with 0-ms intervals vs. with temporal gaps of 500-ms inter-stimulus intervals; ISIs). Our goals were to (i) assess the impact of talker discontinuity on these neurophysiological signals during task performance, (ii) examine how talker discontinuity-related effects change with temporal discontinuity in speech, and (iii) investigate whether talker discontinuity impacts the maintenance of speech information in working memory.

We hypothesized that several distinct neurophysiological signals would be affected by talker and temporal discontinuities during speech processing and working memory maintenance. First, if processing mixed-talker speech disrupts listeners' attentional focus and leads to a processing cost associated with reorienting attention to different sources and features, there are two distinct candidate neurophysiological signatures of attentional disruption that may be elicited. One relevant signature can be revealed through event-related potentials (ERP) in the EEG signal, and the other can be derived from phasic pupil dilations measured in pupillometry. The neurophysiological mechanisms behind these signals are distinct: ERP and phasic pupil dilation responses are not necessarily related to each other (Murphy et al., 2011; Kamp and Donchin, 2015; Hong et al., 2014). Thus, by investigating these two signals in parallel, we can assay diverse neurobiological systems that may play a role in processing talker discontinuity. One relevant ERP signature is the evoked component P3a, which is similar to the distractor positivity (Pd; Stewart et al., 2017). This evoked component arises around 250–300 ms after the onset of a stimulus event, and it is a prominent cortex-originated neural signature of involuntary attentional reorientation to a distracting (or deviant) stimulus (Comerchero and Polich, 1998, 1999; Polich, 2007; Donchin et al., 1997; Stewart et al., 2017). Because this component is phase locked to stimulus events, this signature provides a temporally sensitive index of how listeners' attentional state is impacted immediately after they encounter speech from a new talker. Prior work in the visual domain has shown that the P3a/Pd is elicited in response to confusable distractors that share features similar to the target (Hickey et al., 2009; Hilimire et al., 2011; Sawaki and Luck, 2010). Similarly, auditory EEG studies found that this ERP component was elicited when listeners were distracted by a change in task-irrelevant features of a sound (e.g., Goldstein et al., 2002; Gaeta et al., 2003; Stewart et al., 2017). Thus, we expected to observe a P3a and/or Pd-like evoked response to mixed talkers' speech if listeners' attentional focus is disrupted by changes in task-irrelevant acoustic features, such as those that occur when the talker switches. Furthermore, if the interference from talker discontinuity depends on the timing of speech, the extent of talker discontinuity-related P3a/Pd response should depend on temporal continuity of the speech stream.

The other potential neurophysiological signature of attentional reorientation, functionally discrete from P3a/Pd, is the phasic pupil dilation response (Murphy et al., 2011; Hong et al., 2014; Kamp and Donchin, 2015). Compared to ERPs, phasic pupil dilation responses unfold over a more prolonged timescale (~1–2 s after an onset of a stimulus event) and index the activity of the locus coeruleus norepinephrine (LC-NE) system that cannot be directly assessed by EEG. The

pupil-linked LC-NE system has been associated with overall arousal and vigilance (Aston-Jones and Cohen, 2005; Sara, 2009; Gilzenrat et al., 2010), as well as cognitive effort and working memory load allocated during task performance (Beatty and Lucero-Wagoner, 2000; Unsworth and Robison, 2015; Kahneman and Beatty, 1966; Heitz et al., 2008; Johnson,1971; Johnson et al., 2014; Peavler,1974). Recent studies also suggest that phasic LC-NE activity plays a role in interrupting an ongoing attentional process to support monitoring of the surrounding environment (Bouret and Sara, 2005; Dayan and Yu, 2006; Sara and Bouret, 2012). Through this mechanism, the pupil-linked LC-NE system is both sensitive to the saliency and surprisal of interrupting sounds (Bala and Takahashi, 2000; Huang and Elhilali, 2017; Liao et al., 2016a, 2016b). In particular, rather than gradual and predictive change, this system specifically underlies automatic switching of attention due to an abrupt and rapid bottom-up change in the auditory environment, irrespective of its behavioral relevance (Zhao et al., 2019). Consequently, if the subcortical noradrenergic neuromodulatory system underlies automatic attentional disruption caused from processing mixed-talker speech, we expected that pupil dilation responses would reflect the degree to which attentional focus is disrupted and reoriented by talker discontinuity in speech, as well as how the timing of the speech stream affects the degree of disruption.

Finally, to investigate how talker discontinuity impacts working memory maintenance of speech, we also examined how neural oscillatory power, specifically in the alpha frequency range (8–12 Hz), is affected by talker and temporal discontinuity in speech. Alpha oscillatory power has been shown to be a reliable index of the cognitive demands during task performance across various modalities (Jensen and Mazaheri, 2010; Foxe and Snyder, 2011; Wöstmann et al., 2015, 2016). Enhanced alpha power reflects greater demand on working memory resources and attentional control. Specifically, alpha power parametrically increases during memory retention as more items are held in working memory (Jensen et al., 2002; Tuladhar et al., 2007; Obleser et al., 2012), as well as with increasing demand on attentional control to functionally inhibit task-irrelevant processes (Thut et al., 2006; Klimesch et al., 2007; Wöstmann et al., 2015). Thus, the direction of alpha power modulation can give additional insight into how talker discontinuity impacts working memory maintenance.

First, if stimulus-specific details are maintained in speech working memory (Lim et al., 2015, 2018, 2021) that is potentially contingent upon speech perceptual process (Jacquemot and Scott, 2006; Perrachione et al., 2017; Scott and Perrachione, 2019), it is possible that the increased amounts of acoustic featural variability in mixed- vs. single-talker speech might also affect working memory representations. Thus, if mixed-talker speech increases working memory load (for instance, mixed-talker speech is maintained as multiple, discrete speech objects, whereas single-talker speech is stored as a single object), we would expect to observe enhanced alpha oscillatory power for maintaining mixed- relative to single-talker speech in memory. However, we would expect to observe lower alpha power for maintaining mixed- vs. single-talker speech in memory if mixed-talker speech impairs allocation of attention directed to working memory—for instance, because talker discontinuity leads to inefficient storage of information in memory. Furthermore, in addition to talker continuity, if temporal continuity plays a role in storing speech as a coherent vs. discrete objects in memory, we would expect to see higher alpha power in the 500-ms vs. 0-ms ISI condition even for single-talker speech. However, if temporal discontinuity disrupts efficient attentional allocation, we would expect to observe lower alpha power for retaining speech with the 500-ms ISI vs. 0-ms ISI in memory.

2. Methods

2.1. Participants

Native English speaking adults ($N = 24$; 16 female, 8 male; mean age: 21 years; age range: 18–30) participated in the study. Participants were recruited through the Boston University online job advertisement system. All participants had normal hearing (≤20 dB HL along 250–4000 Hz based on in-lab audiometric tests within 6 months) and reported normal vision. Participants provided written informed consent and were paid $15/hour for their participation. All experimental procedures were approved and overseen by the Boston University Institutional Review Board. One participant's pupillometry data were excluded due to faulty data recording, yielding $n = 23$ for all pupil dilation analyses.

2.2. Stimuli

The stimuli were natural productions of the digits 1–9, recorded from eight native speakers of American English (4 female; 4 male). Each recording was resynthesized to be 550 ms in duration via the pitch-synchronous overlap and add algorithm (PSOLA; Moulines and Charpentier, 1990) implemented in Praat in order to minimize temporal asynchrony in digit sequence presentations. All recordings were normalized to have equivalent root-mean-square amplitude of 65 dB SPL. On each trial, seven randomly selected digit recordings were concatenated to construct a digit sequence appropriate for the task condition. Digit sequences were constrained such that any digit could appear in any position of a sequence, but the same digit could not repeat in adjacent positions within a sequence.

2.3. Task design and procedure

Fig. 1 illustrates the delayed recall of digit span task, based on a previous behavioral study (Lim et al., 2019). This task manipulated the conditions of *talker discontinuity* (mixed-talker vs. single-talker sequences) and *temporal discontinuity* (500-ms vs. 0-ms ISIs) during digit sequence presentation. On each trial, participants heard a sequence of seven spoken digits and then, after a 5-s "hold" period, recalled the sequence in the order of presentation. The digit sequence was either spoken by one consistent talker (single-talker condition) or each of the seven digits was spoken by a different, randomly selected talker (mixed-talkers condition), so that there was no talker repetition within a given sequence. After the 5-s hold period, participants were prompted to recall the sequence, using a computer mouse to click on the digits (in order) on a number pad GUI that appeared on the screen. Participants were instructed to internally rehearse the digit sequence during the 5-s hold period but to refrain from speaking out loud. Throughout the digit encoding and hold periods, listeners were asked to fixate on a black center dot displayed on the computer screen to minimize artifactual ocular movement. In order to reduce artifacts in the EEG and pupillometry data, participants were also verbally instructed to withhold eye

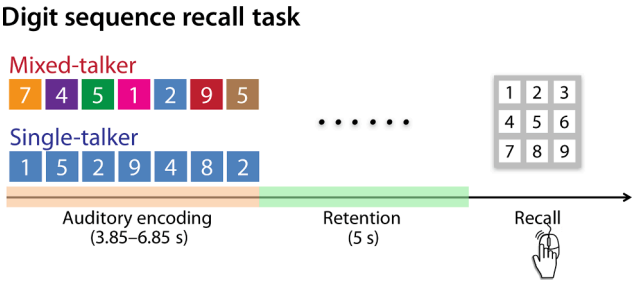

Fig. 1. Schematic illustration of the delayed digit span recall task. Listeners heard a sequence of seven digits, either spoken by one consistent talker (*single-talker*) or with each digit in the sequence spoken by a different talker (*mixed-talker*). The digits were presented either with 0-ms or 500-ms inter-digit intervals. After encoding, participants briefly retained the sequence, during a 5-s silent delay period. The appearance of a number pad display on the screen was participants' cue to begin recalling the digit sequence. Throughout the encoding and retention periods, a center fixation dot was displayed on the screen.

blinks as much as possible until the number pad appeared on the screen.

Each participant performed a total of 144 trials. Trials were organized in six runs of 24 trials. Temporal discontinuity conditions (500-ms vs. 0-ms ISIs) varied across runs; all trials in a run had an identical ISI, and the ISI alternated between runs. The order of runs was counterbalanced across participants using Latin-square permutation. Within each run, trials were blocked by talker condition, with three trials per block (i.e., 4 blocks of single-talker and 4 blocks of multi-talker per run). The order of blocks were randomized independently within each run and across participants. We ensured that all speech tokens of each talker were presented an equal number of times across the two talker conditions.

Prior to the experiment, participants completed three practice trials to become familiar with the task structure and with the eye tracker set up. Prior to start of each run, the participant's head position was stabilized using a head-chin rest, and their eye gaze position was calibrated using a five-point calibration controlled by EyeLink 1000 (SR Research). Participants were given a self-paced break after each trial, and a trial started only after the participant's eye gaze was stable at the fixation position on the screen; if the fixation drift check failed, eye gaze was re-calibrated before the experiment began. Each trial started 2 s after the eye gaze fixation was accepted, which was indicated to the participants via the color of the fixation dot displayed on the screen.

The experiment was controlled via Psychtoolbox-3 (MATLAB; Brainard, 1997; Kleiner et al., 2007). All sound stimuli were delivered through Etymotic ER-1 insert earphones in a darkened, electrically shielded, sound attenuated chamber. All visual information was displayed via a BenQ 1080p LED monitor (27″ diagonal) placed 57 cm from the participant's head position. The EyeLink 1000 eye-tracking system was placed below the computer monitor.

2.4. Data analysis

2.4.1. Behavioral analysis

A prior study using an identical task examined listeners' recall accuracy, response times, and a composite measure of recall efficiency (Lim et al., 2019). The current study focused on only the accuracy measure because task demands of simultaneous pupillometry and EEG recording required participants to suppress blinking until the end of the retention period, precluding them from providing their responses as quickly as possible. Here, we analyzed the accuracy of participants' memory recall for the digit sequence on each trial. Accuracy was quantified based on correct recall of the digit at each position in the sequence. Data were analyzed in R using a logistic mixed-effects model (*lme4* in R v3.3.3), with recall accuracy of each digit in the sequence as the dependent measure. The model's fixed-effects terms included the categorical factors of *talker discontinuity* (mixed- vs. single-talker) and *temporal discontinuity* (500-ms vs. 0-ms ISI), the continuous covariate *digit position* in the sequence (1–7) (mean-centered and unit scaled), and all two- and three-way interaction terms. Random-effects terms included by-participant intercepts and slopes for the fixed factors. Deviation (sum) coded contrasts were applied to the categorical factors. The significance of the effects was determined based on Type-III Wald χ^2 tests (at $p < 0.05$).

2.4.2. EEG data acquisition and preprocessing

EEG data were acquired from 64 active scalp channels in the standard 10/20 montage (Biosemi ActiveTwo system) at a sampling rate of 2048 Hz. Four additional electrodes were placed to record horizontal and vertical ocular eye movements, and two electrodes were placed on the earlobes for reference. Data were downsampled at 1024 Hz and band-pass filtered from 1 to 60 Hz using a zero-phase finite impulse response (FIR) filter (Kaiser windowed, Kaiser $\beta = 5.65$, filter length 7420 points).

The data were preprocessed using EEGLAB (Delorme and Makeig, 2004). An independent component analysis was performed on continuous data, and artifact components related to eye movements, heartbeat, or noise were identified and removed from the data. On average, 14.5 ± 6.3 (mean ± SD) of 64 components were removed per participant. Data from bad electrodes (2.1 ± 1.7) from seven participants were reconstructed using a spherical interpolation method implemented in EEGLAB. The continuous data were first divided into epochs of −2 to 13 s relative to the trial onset (i.e., onset of the first digit in the trial's digit sequence). Epochs were removed if any scalp electrodes showed an activity range greater than 200 μV or had infrequent electrical artifacts based on manual inspection within the data segments from −1 to 10 s and −1 to 13 s relative to trial onset for the 0-ms and 500-ms ISI condition trials, respectively. On average, 4.5% of trials were rejected per participant through this procedure; the proportion of remaining trials did not differ across the experimental conditions [repeated-measures ANOVA; *talker discontinuity*: $F_{1,23} = 2.07$, $p = 0.16$, $\eta^2_p = 0.01$; *temporal discontinuity*: $F_{1,23} = 0.47$, $p = 0.50$, $\eta^2_p = 0.014$; *talker* × *temporal discontinuity*: $F_{1,23} = 0.34$, $p = 0.56$, $\eta^2_p = 0.003$]. Data were analyzed using the FieldTrip toolbox (Oostenveld et al., 2011) and customized MATLAB scripts.

2.4.3. Event-related potentials

Prior to quantifying event-related potentials (ERPs), the single-trial EEG data were downsampled to 256 Hz. Because cortical auditory evoked responses are sensitive to the acoustic characteristics of speech sounds (e.g., Näätänen and Picton, 1987; Digeser et al., 2009; Khalighinejad et al., 2017; Tremblay et al., 2003), the high acoustic variability across the different digit recordings both within a talker and across the eight talkers can increase irrelevant noise in the evoked responses. In contrast to the evoked response to the first digit of a trial, which was preceded by a long period of silence, the evoked responses to the subsequent digits exhibit variable and non-canonical neural response patterns when time-locked to the onset of the audio recordings (see Fig. 3**A** and **Figure S1A**). This issue can be exacerbated when evoked responses to speech are estimated throughout the entire period of digit sequence encoding (Fig. 3**A** and **Figure S2**); the drifts in the evoked signals accumulate over time, which makes it even more difficult to isolate neural responses precisely time-locked (and phase-locked) to each stimulus event. To mitigate these concerns and to reduce potential noise, we therefore computed neural evoked responses time-locked to the voicing onsets of the spoken digits in the sequence (**Figure S1B**). First, we divided the single-trial EEG data into 600-ms epochs corresponding to each sequence digit from −0.1 to 0.5 s relative to the voicing onset of the specific digit. These single-trial EEG epochs were baseline corrected by subtracting the mean amplitude in the time window of −0.1 to 0 s relative to the stimulus voicing onset. On average, the onset of voicing varied by 113 ± 24 ms (mean ± SD) from the start of the stimulus recording.

The ERP for each condition was quantified as the average evoked responses to all spoken digits except the first digit (i.e., evoked responses averaged across digits positioned from 2 to 7). We excluded the response to the first digit because, unlike the other digits in the sequence, the first digit presentation followed a long silence (i.e., the pre-trial period), and the neural response to the first digit in the sequence is independent of the experimental manipulations of interests (i.e., talker discontinuity or temporal discontinuity), which manifest only upon the presentation of the second digit. Confirming these observations, the temporal profile of evoked responses to the first digit was qualitatively and quantitatively distinct from the that of the rest of the digits (Fig. 3**A; Figure S1**); also, evoked responses to the voicing onset of the first digit in the sequence did not differ across the conditions (all *p*s greater than 0.17 from permutation-based cluster tests).

For statistical analysis of ERPs to spoken digits, we used a cluster-based permutation *t*-test provided by Fieldtrip (Maris and Oostenveld, 2007). We conducted three planned analyses, which tested the main effects of *talker discontinuity* and *temporal discontinuity*, and their interaction during the time window of 0 to 0.5 s relative to the voicing onsets

of digits. First, to test the main effects of *talker discontinuity*, subject-wise differences between the mean ERPs in the mixed- and single-talker conditions (aggregated across the two temporal discontinuity conditions) were entered into the cluster-based, one-sample *t*-test against zero. Second, to test the main effect of *temporal discontinuity*, we contrasted the mean ERPs of the 500-ms vs. 0-ms ISI conditions (aggregated across the talker conditions) with cluster-based one-sample *t*-test against zero. To test the *talker* × *temporal discontinuity* interaction effect, we contrasted the ERPs of the talker condition trials (mixed- vs. single-talker) separately for each ISI; this ERP difference was entered into the paired *t*-test to contrast the effect of talker in the 500-ms vs. 0-ms ISI conditions. For each analysis, a permutation test (1000 Monte Carlo permutations) was performed with a cluster-based control at a type I error level of alpha = 0.05, implemented in Fieldtrip. The electrode neighborhood for this analysis was defined by neighboring distance of 17.5 mm, which resulted in on average of 6.88 neighbors for each electrode. The test resulted in time–electrode clusters exhibiting significant effects of the corresponding contrasts. For any of the significant clusters, we also examined the magnitude of the evoked response elicited in each condition (2 *talker* × *2 temporal discontinuity*) using a one-sample *t*-test against 0.

2.4.4. Time–Frequency analysis

Prior to performing the time–frequency analysis, EEG data were re-referenced to the average of all EEG electrodes. We obtained time–frequency representations (TFRs) of single-trial EEG data by convolving the data with a Hanning taper (fixed window length of 100 ms), which covered frequencies from 1 to 30 Hz with a resolution of 1 Hz. This procedure was applied to the whole epoch from −2 to 13 s relative to the onset of the trial (i.e., the onset of the first spoken digit) using a time step of 0.1 s. Single-trial power estimates were baseline corrected by subtracting the relative change with respect to the average oscillatory power across all four conditions during the pre-stimulus baseline (−0.5 to 0 s).

We focused on how talker discontinuity and temporal discontinuity affect neural oscillatory power during working memory retention after encoding digits. We adopted multi-level statistical analysis used in previous studies (Obleser et al., 2012; Lim et al., 2015; Wöstmann et al., 2017) for conducting three planned analyses, testing the main effects of *talker discontinuity* and *temporal discontinuity* as well as their interaction during the 5-s hold period during which listeners retained information in working memory. Single-subject-level statistical analysis was performed on single-trial data using an independent-samples regression *t*-test with contrast coefficients of −0.5 and 0.5 to test the effect of interest (i.e., *talker discontinuity:* mixed- vs. single-talker trials, *temporal discontinuity:* 500-ms vs. 0-ms ISI trials, respectively) while collapsing the other condition; this resulted in the individual subject's time–frequency–electrode beta weights of the corresponding contrast, which were entered into the group-level analysis.

To test the main effects of *talker* and *temporal discontinuity* at the group level, a cluster-based one-sample *t*-test against zero was performed from the frequency range from 3 to 20 Hz across the 5-s hold period. For testing the *talker* × *temporal discontinuity* interaction, we computed beta weights contrasting the talker conditions (mixed- vs. single-talker) on the first-level analysis separately for the 0-ms and 500-ms ISI condition trials; the resulting beta contrasts of the two speech rate conditions were tested at the group level using a paired-sample *t*-test. As when analyzing ERPs, the group-level analysis was conducted as using a cluster-based permutation test with 1000 random iterations (Maris and Oostenveld, 2007) with the same neighboring distance of 17.5 mm in defining the electrode neighborhood. Each test resulted in time–electrode–frequency clusters exhibiting significant effects of the corresponding conditions on the TFRs throughout the 5-s retention window.

In addition, we also examined the main effect of *talker discontinuity* on the neural oscillatory power across 3–20 Hz during the presentation of the spoken digit sequence, henceforth referred to as the encoding phase of the trial, using the multi-level analysis approach as described above. Due to the differences in the digit sequence durations across the two temporal discontinuity conditions, we contrasted the mixed- vs. single-talker conditions separately for the 0-ms and 500-ms ISI condition trials during this encoding phase (0-ms ISI: 0–3.85 s, 500-ms ISI: 0–6.85 s relative to trial onset).

2.4.5. Pupillometry preprocessing and analysis

Pupil diameters from the left and right eyes were continuously recorded with the EyeLink 1000 (SR Research) at a sampling rate of 500 Hz. The pupillometry data were epoched from −2 s to 8.85 s and −2 to 11.85 s relative to the trial onset (i.e., first spoken digit) for the 0-ms and 500-ms ISI trials, respectively. Preprocessing steps included a de-blinking procedure and removal of artifact trials. We used a combination of eye position, velocity, and acceleration filters to automatically detect eye blinks, shown as rapid drops in the pupil size traces. The blink data segment was replaced based on linear interpolation between the pupil sizes before and after each blink using customized MATLAB functions. We rejected any trials in which either more than 50% of the data consisted of blinks or that contained long blinks (duration greater than 500 ms) within the 1-s pre-trial window. Subsequently, trials with excessive noise were manually identified and removed from data analysis. On average, 17.35 (±15.49) out of 144 trials were rejected through this process. The pupil diameter of each trial was quantified as the mean of the pupil tracings of the two eyes. For trials in which the fixation of one eye was unstable during recording, the pupil tracing from the other stably fixated eye was used for the trial (3.65% of trials were a single pupil recording).

Pupil diameter during the whole trial period was baseline corrected by computing the relative change from the average pupil diameter during the pre-stimulus baseline (−0.5 to 0 s) of each trial. It is of note that there were no significant effects of the conditions or their interaction on the average pupil diameters during baseline [repeated-measures ANOVA; *talker discontinuity*: $F_{1,22} = 0.60$, $p = 0.45$; *temporal discontinuity*: $F_{1,22} = 1.17$, $p = 0.29$; *talker* × *temporal discontinuity*: $F_{1,22} = 1.16$, $p = 0.29$].

Our main question was whether pupil dilation responses during task performance were sensitive to talker and temporal discontinuity during the speech encoding and the memory retention phases of the trials. To assess pupil dilation responses during encoding of spoken digits, we quantified mean pupil diameter during the presentation of each digit in the sequence of each trial. In order to account for the temporal delay in pupillary response from stimulus onset (Hoeks and Levelt, 1993; Verney et al., 2004; Winn et al., 2015; 2018), the time windows in which pupil diameters were averaged extended 500 ms beyond the onset of each digit in the sequence. In order to quantify pupil responses during memory retention, we averaged pupil diameters during the 5-s memory retention phase of each trial.

The resulting trial-wise pupil diameter data were analyzed using linear mixed effects models. The fixed effects structure included categorical factors for *talker discontinuity*, *temporal discontinuity*, and their interaction. An additional continuous fixed factor for *digit position* (1–7; mean-centered and scaled), as well as its interaction with the other fixed factors, was entered into the model for analyzing pupil responses during encoding of spoken digits. The random effects structure included by-participant intercepts and slopes for the fixed factors.

3. Results

3.1. Talker discontinuity in speech interferes with speech working memory recall

We analyzed whether the 2 *talker* × 2 *temporal discontinuity* condition manipulations affected performance on the digit sequence recall task. The results of the logistic mixed-effects model of participants' memory

recall accuracy are listed in Table 1. The model revealed a significant main effect of *talker* [$\chi^2_1 = 16.03, p \ll 0.0001$], but no significant effects of *temporal discontinuity* [$\chi^2_1 = 0.47, p = 0.50$] or the interaction *talker* × *temporal discontinuity* [$\chi^2_1 = 0.0073, p = 0.93$]. These results indicate that performance was less accurate for digit sequences spoken by mixed talkers than by a single talker; however, the effect of talker discontinuity on memory recall accuracy did not depend on the ISI of spoken digits (Fig. 2).

The model also revealed a significant main effect of *digit position* [$\chi^2_1 = 47.45, p \ll 0.0001$]. As illustrated in Fig. 2, listeners' recall accuracy was greatest for recalling digits presented in the initial and final positions of the sequence, exhibiting typical primacy and recency effects of sequence recall. Importantly, the significant interaction between *temporal discontinuity* and *digit position* [$\chi^2_1 = 5.66, p = 0.017$] shows that the pattern of recall accuracy across sequences differs for the two ISIs: participants were significantly more accurate in recalling the first two digits in the sequence with 0-ms ISI compared to with 500-ms ISI [both $\beta > 0.095$, $z > 2.23$; $p < 0.026$].

3.2. Evoked responses during speech encoding are impacted by talker discontinuity and temporal discontinuity

Fig. 3**A** illustrates each condition's evoked response time course throughout the trial. Our main interest in this analysis was whether talker discontinuity in speech affected auditory evoked responses and whether the extent of the response differed across the temporally continuous vs. discontinuous speech stream. The first permutation-based cluster test examined the main effect of the talker condition on the average ERPs across the two ISI conditions. This test revealed one significant cluster exhibiting a mixed-talker > single-talker effect (Fig. 4), such that the mixed-talker condition exhibited a stronger positive potential compared to the single-talker trials in the time range of 215–348 ms following the voicing onsets of spoken digits ($p = 0.001$). This effect was widely distributed, but most pronounced in the fronto-central electrodes.

The other cluster test examining the main effect of temporal discontinuity yielded two significant clusters, both exhibiting stronger positivity in the 500-ms than the 0-ms ISI (**Figure S3**; Cluster #1: 0–63 ms; $p = 0.015$; Cluster #2: 105–500 ms; $p < 0.001$). However, the cluster test evaluating the *talker* × *temporal discontinuity* interaction effect did not yield any significant clusters. Thus, the pattern of results indicates that for both ISIs, mixed-talker trials exhibited a significantly larger positive potential around 300 ms after the onset of spoken digits than did single-talker trials (Fig. 4**C–D**). Based on this pattern, we further examined the magnitude of the response elicited by spoken digits of each condition using a one-sample *t*-test against 0. This test revealed that only the single-talker, 0-ms ISI condition did not yield significantly positive potential (Fig. 4**D**; mixed-talker speech at 500-ms ISI: $t_{23} = 6.49$, $p \ll 0.0001$; mixed-talker speech at 0-ms ISI: $t_{23} = 4.51$, $p = 0.00016$; single-talker speech at 500-ms ISI: $t_{23} = 5.23$, $p \ll 0.0001$; single-talker speech at 0-ms ISI: $t_{23} = 1.47$, $p = 0.16$).

Table 1
Mixed-effects logistic modeling results of the behavioral recall accuracies across digits in sequences.

Fixed factors	χ^2	*df*	*p*
Talker discontinuity	16.03	1	≪ 0.0001
Temporal discontinuity	0.47	1	0.50
Digit position	47.45	1	≪ 0.0001
Talker × Temporal discontinuity	0.0073	1	0.93
Talker discontinuity × Digit position	3.41	1	0.065
Temporal discontinuity × Digit position	5.66	1	0.017
Talker × Temporal discontinuity × Digit position	1.45	1	0.23

Note: Type III Wald χ^2 test

3.3. Alpha oscillatory power during speech working memory retention is sensitive to talker discontinuity

Our main question in analyzing oscillatory power was whether mixed- and single-talker speech produced different levels of neural oscillations during memory retention, and whether any such effect of talker condition differed across the two temporal discontinuity conditions. Fig. 3**B** shows the grand average oscillatory power time courses of the mixed-talker vs. single-talker conditions across the two ISI conditions. In all cases, there is a clear band of power in the alpha range that extends from the encoding phase and into the memory retention phase of the trials. We investigated the main effect of talker discontinuity (mixed- vs. single-talker) during the 5-s memory retention phase using a permutation-based cluster test. This analysis identified one cluster 1.95–3.55 s after the onset of the memory retention phase (Fig. 5**A**); this cluster exhibited significantly less oscillatory power in the alpha frequency range [8–11 Hz, $p = 0.001$] when listeners were maintaining digits spoken by mixed talkers compared to a single talker in both 0-ms and 500-ms ISI trials (Fig. 5**B**). The cluster test examining the *talker* × *temporal discontinuity* interaction effect did not yield any significant clusters (all *ps* ≥ 0.36). This pattern indicates that compared to maintaining single-talker speech, maintaining mixed-talker speech led to significantly less alpha power during the retention period across both temporal discontinuity conditions.

Although the main effect of *temporal discontinuity* (500-ms vs. 0-ms ISI) was not our main interest, we also noted that the temporal discontinuity of speech across the talker conditions had a significant effect on neural oscillatory power. As shown in **Figure S4**, a permutation-based cluster test examining the main effect of temporal discontinuity revealed a significant enhancement of alpha and lower beta power (10–16 Hz) when listeners were maintaining digit sequences presented at 0-ms ISI compared to 500-ms ISI [0–2.75 s; $p = 0.003$] in both talker conditions.

Lastly, we examined whether alpha oscillatory power differed when listeners were encoding speech spoken by mixed-talkers vs. a single-talker, as might be expected under cognitive resource-allocation based models of talker adaptation (Nusbaum and Magnuson, 1997; Magnuson and Nusbaum, 2007). The corresponding permutation-based cluster did not reveal any clusters exhibiting significant differences in neural oscillatory power in the range from 3 to 20 Hz during speech encoding of mixed- vs. single-talker digit sequences in either *temporal discontinuity* condition trials [0-ms ISI: all *ps* ≥ 0.11; 500-ms ISI: all *ps* ≥ 0.60].

3.4. Pupil dilation response during speech encoding is sensitive to talker and temporal discontinuities

We investigated whether task-evoked pupil dilation responses were related to encoding and maintaining mixed-talker vs. single-talker speech, and whether temporal discontinuity of speech affected pupil responses. Fig. 3**C** illustrates the time courses of the pupillary response during task performance in each experimental condition.

We analyzed the effects of *talker discontinuity* and *temporal discontinuity* on the pupillary responses for encoding each digit in the sequence. Linear mixed-effects model (Table 2) revealed a significant main effect of *digit position* [$F_{1,21.8} = 58.76, p \ll 0.0001$] and a significant interaction effect of *temporal discontinuity* × *digit position* [$F_{1,20310} = 9.88$; $p = 0.0017$]. As illustrated in Fig. 6, pupil dilations generally increased over the course of the encoding phase, and the amount that pupil dilation increased per digit was higher in the 0-ms than 500-ms ISI conditions. The same linear mixed-effects model (Table 2) did not reveal significant main effects of *talker discontinuity* [$F_{1,\ 21.8} = 0.19, p = 0.67$] or *temporal discontinuity* [$F_{1,\ 21.3} = 2.61, p = 0.12$], but there was a significant *talker* × *temporal discontinuity* interaction effect [$F_{1,\ 20312} = 5.45, p = 0.020$]. As shown in Fig. 6, this result shows that the pupil dilation increase for encoding digits spoken by mixed talkers vs. a single talker was higher when digits were presented at 0-ms ISI compared to 500-ms ISI [mean

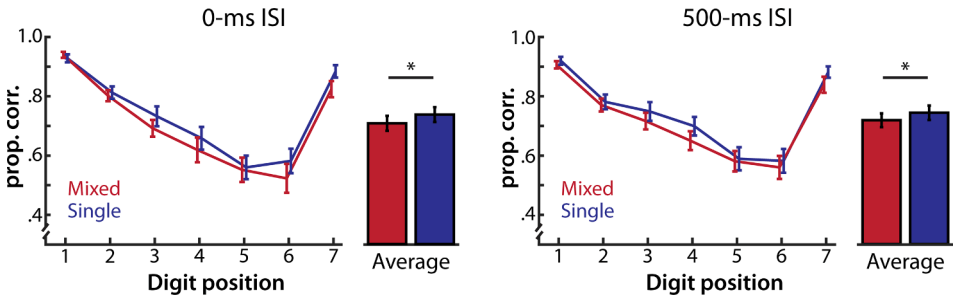

Fig. 2. Digit sequence recall performance in the talker and temporal discontinuity conditions. Mean recall accuracy for each digit position is illustrated in the mixed- vs. single-talker conditions in the 0-ms ISI (left) and 500-ms ISI (right) conditions. The bar graphs show mean performance across all digit positions. The error bars indicate ± 1 standard error of mean (SEM) across participants. * $p < 0.05$.

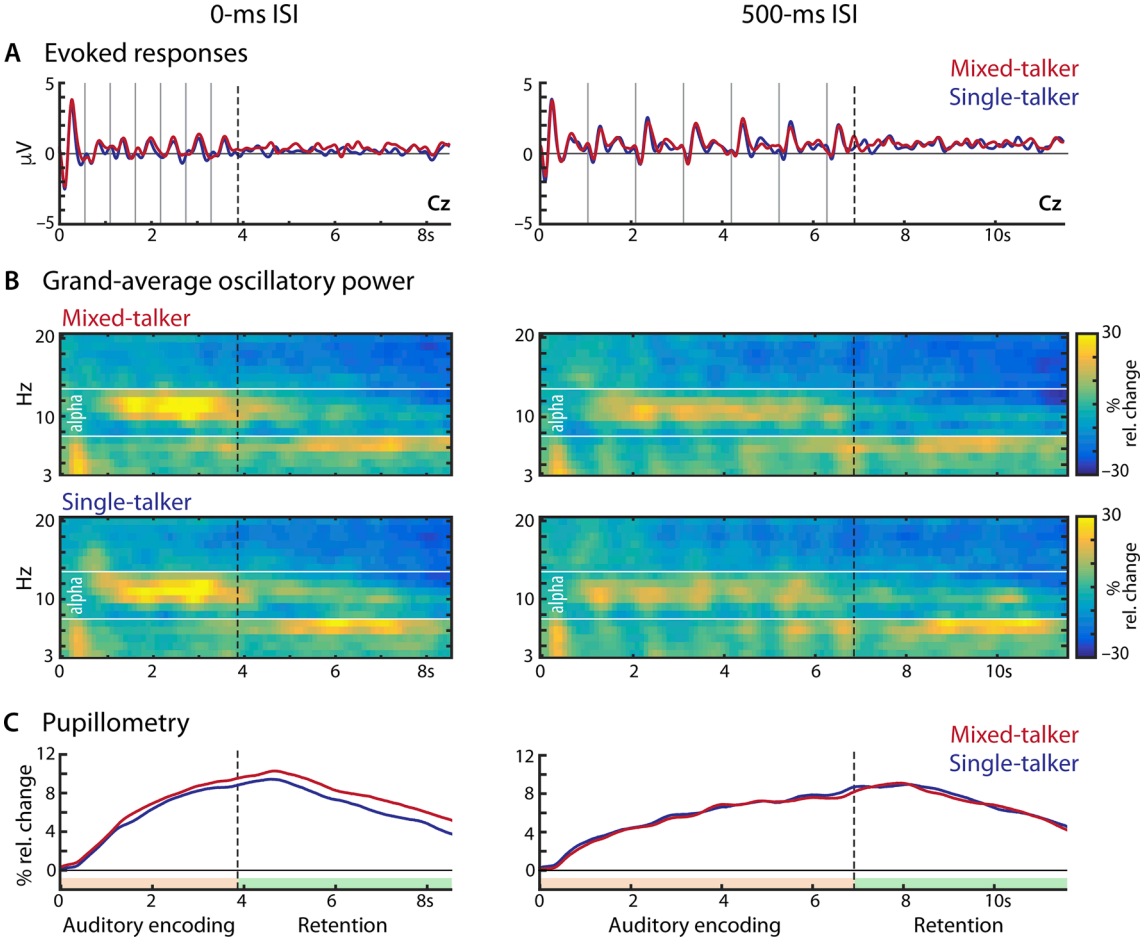

Fig. 3. Illustration of the mean time courses of the three neurophysiological responses during the delayed digit sequence recall task. **(A)** Grand average evoked response time courses across participants, time-locked to the onset of the trial extracted from one representative electrode (Cz). For illustration, the grand average time courses were low-pass filtered at 15 Hz. Thin vertical lines denote the onsets of spoken digits in the sequence. Dashed lines mark the offset of the last digit in the sequence (i.e., end of sequence encoding and beginning of retention) as denoted in panel C. Red and blue lines indicate average responses of the corresponding neurophysiological time courses during mixed- and single-talker condition trials, respectively. **(B)** Condition-specific grand average time–frequency representations across participants, and across all scalp channels. The color bar indicates relative oscillatory power change from baseline. Alpha frequency range (8–13 Hz) is demarcated by white horizontal lines. **(C)** Mean pupillometry response time courses across participants in each condition. Y-axis indicate relative pupil dilation change from the average pupil responses during 500 ms pre-trial baseline.

effect of talker discontinuity across all digit positions: 0-ms ISI: 0.52%; 500-ms ISI: –0.18%].

We also examined how these factors affected pupil dilation during memory retention. We examined the effects of *talker discontinuity*, *temporal discontinuity*, and their interaction on the average pupillary responses during the 5-s memory retention phase. A linear mixed-effects model did not yield any significant main or interaction effects [*talker*: $F_{1,21.53} = 0.57$, $p = 0.46$; *temporal discontinuity*: $F_{1,19.77} = 0.12$, $p = 0.73$; *talker* × *temporal discontinuity:* $F_{1,2851} = 1.78$, $p = 0.18$]. Finally, correlations among the talker discontinuity-related differences in the various dependent measures (behavior, ERP, neural oscillatory power, and pupillometry) are enumerated in Table S1.

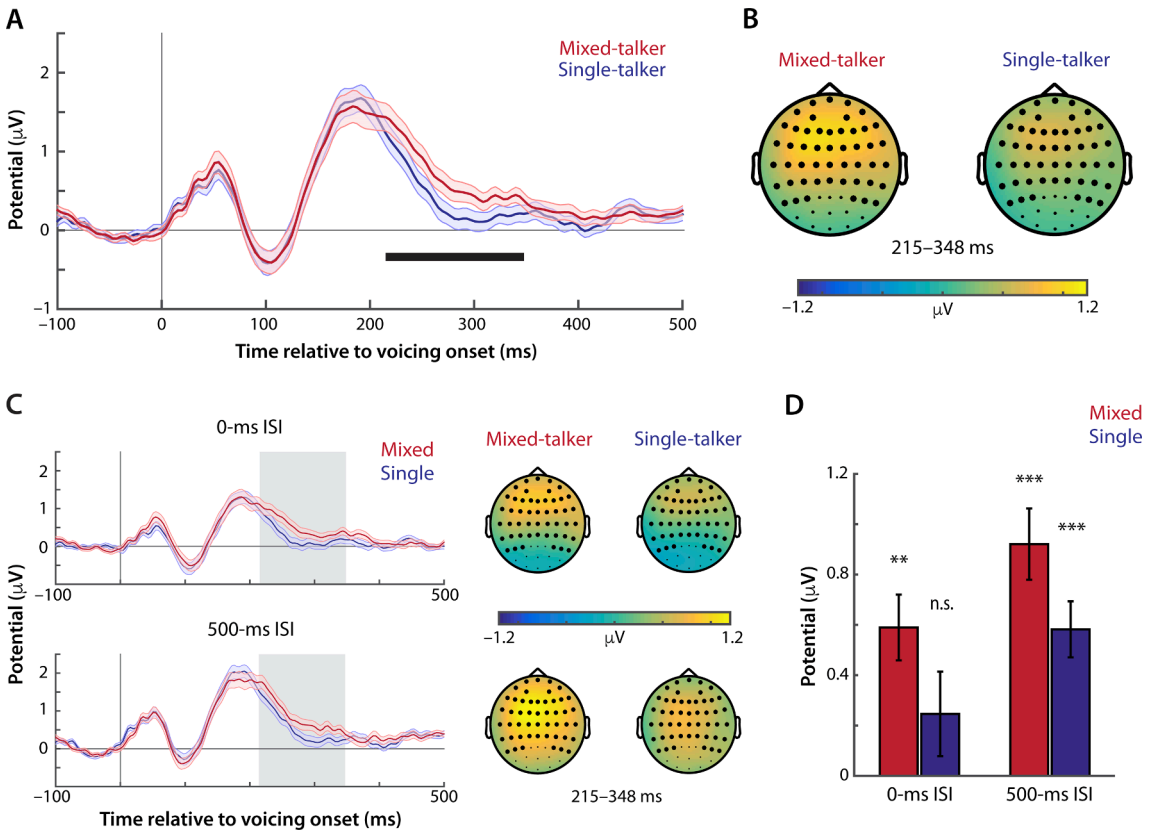

Fig. 4. Grand average event-related potentials (ERPs) across participants time-locked to the onsets of voicing of spoken digits in the sequence (excluding the first digit of each sequence). **(A)** Evoked responses in the mixed-talker vs. single-talker trials, averaged across 49 electrodes belonging to the significant cluster. The black horizontal line denotes the time period in which the cluster-level statistic showed a significant main effect of the talker conditions (i.e., significant difference between the mixed- vs. single-talker conditions collapsed across the two ISI conditions). Error bars indicate ± 1 SEM. **(B)** Topographical maps of average evoked response amplitudes within the time–electrode cluster exhibiting a significant mixed-talker vs. single-talker contrast. Highlighted scalp channels belong to the significant cluster. **(C)** Grand average time courses (left) and topographical maps (right) of the ERPs in the 2 *talker* × 2 *temporal discontinuity* conditions. The grey boxes in the time courses denote the same time points that belong to the cluster shown in (A). **(D)** A bar plot of the condition-specific average evoked response amplitudes over electrodes and times of the cluster shown in (A). Comparisons to zero: ** $p < 0.001$, *** $p \ll 0.0001$.

4. Discussion

Although several psycholinguistic models have been put forward to explain how the differences in acoustic–phonetic mappings across talkers place additional processing demands on listeners during speech perception (Nearey, 1989; Johnson, 2005; Nusbaum and Magnuson, 1997; Kleinschmidt and Jaeger, 2015), previous research has not explicitly examined whether the psychophysiological signatures of mixed-talker speech perception are consistent with the distinct cognitive mechanisms proposed by these frameworks (Belin and Zatorre 2003; Wong et al., 2004; Chandrasekaran et al., 2011; Perrachione et al., 2016). In the current study, we examined whether the psychophysiological correlates of encoding and retention of mixed-talker speech were consistent with the cognitive processes posited by one of these frameworks: stimulus-driven auditory attention and streaming (Bregman, 1990; Shinn-Cunningham, 2008; Winkler et al. 2009).

We found that talker discontinuity interfered with listeners' serial recall of speech from working memory. This behavioral interference from talker discontinuity was accompanied by neurophysiological responses signaling attentional reorientation and interruption of an ongoing auditory streaming process. Encountering discontinuities in incoming speech elicited evoked responses similar to the P3a/Pd ERP component, the magnitude of which increased additively with increased discontinuity from either switches in talkers, temporal separation, or both. Mixed-talkers' speech led to increased phasic pupil responses, particularly when listeners had to process speech rapidly. Furthermore, compared to single-talker speech, mixed-talker speech produced less alpha oscillatory power during working memory maintenance, but not during initial speech encoding.

Broadly, these results are in line with the neural mechanisms predicted by an auditory attention and streaming framework of talker-specific speech processing (Bressler et al., 2014; Choi and Perrachione, 2019; Kapadia and Perrachione, 2020). Specifically, talker discontinuity interrupts listeners' attentional focus to an ongoing auditory stream formed by the attended talker; obligatory exogenously driven attentional shifts to novel speech sources encumber the auditory system, resulting in the processing interference classically associated with mixed-talker speech perception.

4.1. Talker discontinuity interferes with speech working memory recall

We found that listeners were less accurate at recalling speech sequences spoken by mixed talkers than by a single consistent talker. This performance interference by talker discontinuity is in line with decades of results demonstrating a consistent and robust interference effect of mixed talkers' speech on immediate recognition of speech (e.g., Choi et al., 2018; Mullennix and Pisoni, 1990; Nusbaum and Morin, 1992). Behavioral measures alone cannot adjudicate whether less accurate recall of mixed-talker speech is driven solely by inaccurate encoding of mixed-talker speech or also by an influence of talker discontinuity on working memory maintenance (Lim et al., 2019). However, our EEG results, especially our alpha oscillatory power (discussed below), suggest that talker discontinuity continues to impose processing costs even while listeners maintain speech information in working memory

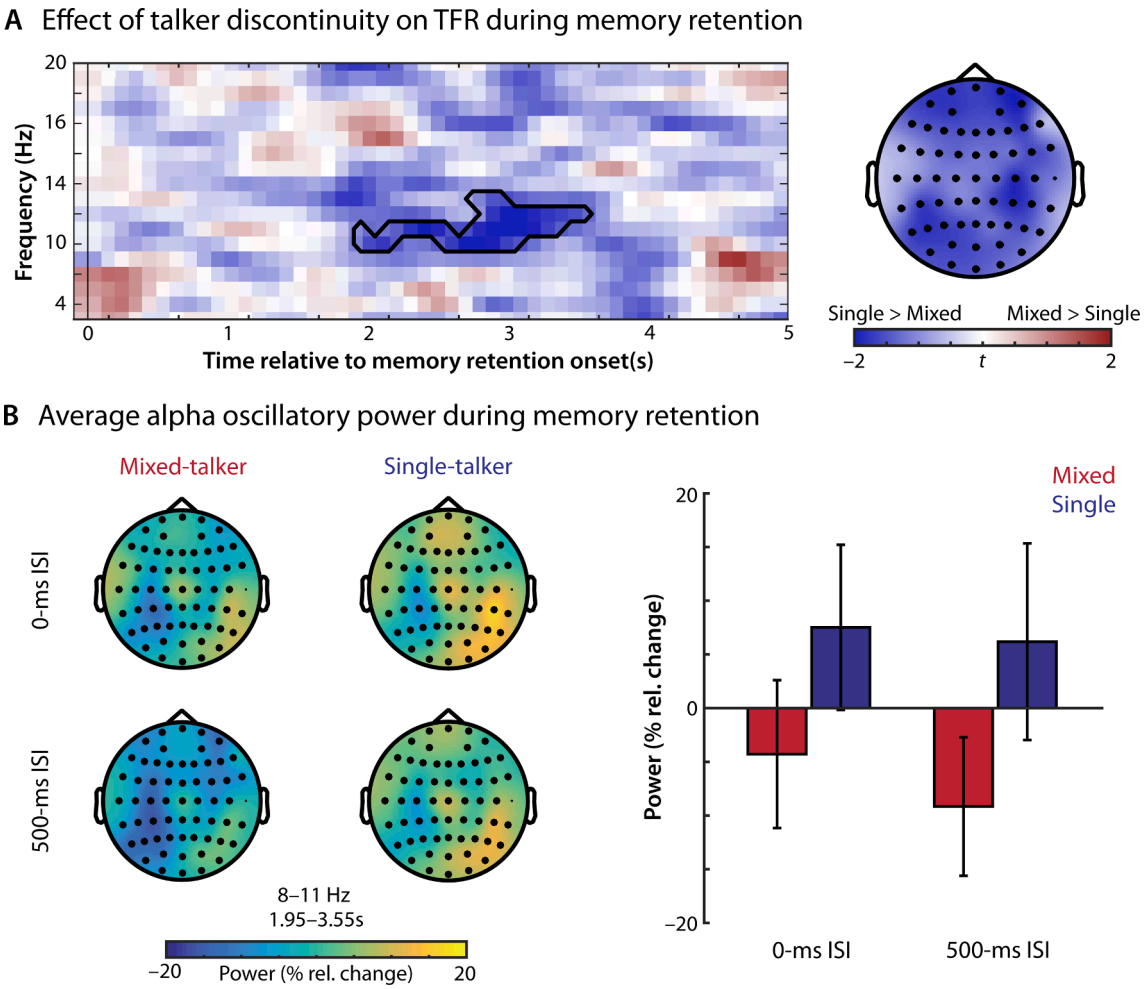

Fig. 5. Effect of talker discontinuity on neural oscillatory power during the 5-s memory retention phase. **(A)** Illustrations of the time–frequency representation and topographical map of the cluster exhibiting a significant main effect of the talker condition (i.e., mixed- vs. single-talker contrast aggregated across the two ISIs). The colors indicate the distributions of *t*-statistics of the mixed-talker vs. single-talker contrast. The highlighted electrodes on the topography belong to the corresponding cluster. **(B)** Average oscillatory power in the 2 *talker* × 2 *temporal discontinuity* conditions over the single-talker > mixed-talker cluster shown in (A). Left, topographical illustrations of the condition-specific averages of oscillatory power over the frequencies and times of the cluster. Right, a bar plot of condition-specific oscillatory power average over the electrodes, frequencies, and times of the cluster. Error bars indicate ± 1 SEM.

Table 2
Mixed-effects linear modeling results on the pupil dilation responses to spoken digits during the sequence encoding phase.

Fixed factors	*df1*	*df2*	*F*	*p*
Talker discontinuity	1	21.8	0.19	0.67
Temporal discontinuity	1	21.3	2.61	0.12
Digit position	1	21.8	58.76	≪ 0.0001
Talker × Temporal discontinuity	1	20,312	5.45	0.020
Talker discontinuity × Digit position	1	20,305	0.11	0.74
Temporal discontinuity × Digit position	1	20,310	9.88	0.0017
Talker × Temporal discontinuity × Digit position	1	20,300	0.11	0.75

Note: Type III Analysis of Variance with Satterthwaite's method

(Fig. 5). Specifically, the fact that there is a significant difference in alpha power when maintaining digits suggests that there are differences in working memory demands for single- vs. mixed-talker stimuli, even after the digits have been encoded and abstracted to memory.

Overall, talker discontinuity was equally detrimental to performance accuracy across the two temporal discontinuity conditions (Fig. 2). This pattern of recall accuracy replicates what we found in our prior study utilizing the identical experimental paradigm (Lim et al. 2019). However, it is of note that using the accuracy measure as a dependent variable can limit the sensitivity in capturing how fine-grained speech processing dynamics in single- vs. mixed-talker settings can differ with temporal continuity in speech. Prior work on immediate speech identification has shown that talker discontinuity has a much greater effect on word identification speed than on accuracy (Kapadia & Perrachione, 2020). Consistent with this, our previous study found that talker discontinuity had a larger effect for temporally continuous speech when considering response time and speech processing efficiency (which simultaneously accounts for the speed and accuracy of memory recall), but not for recall accuracy considered alone (Lim et al., 2019). As noted above, we could not reliably assess participants' memory recall speed in the current study, where we simultaneously measured EEG and pupil dilation. Specifically, in order to obtain artifact-free physiological measures, participants were asked to withhold responses until after the retention period. Moreover, because participants were told to refrain from blinking during the encoding and retention periods, many actually delayed responding until after blinking.

4.2. *Talker discontinuity affects neural dynamics of automatic attentional reorientation during speech processing*

Across the two temporal discontinuity conditions we found that, compared to a single talker's speech, listening to mixed talkers' speech

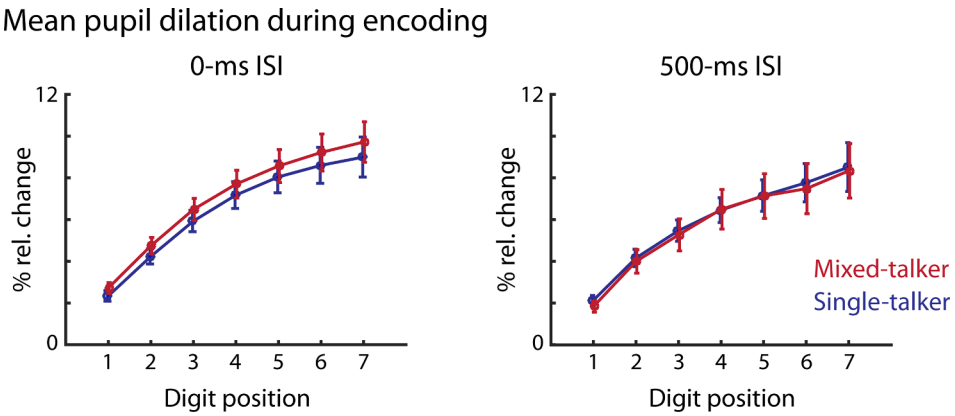

Fig. 6. Mean pupil dilations in the 2 *talker* × 2 *temporal discontinuity* conditions during encoding of each spoken digit in the sequences. The error bars indicate ± 1 SEM across participants.

consistently evoked greater fronto-central positive deflections arising about 200–350 ms after the onset of spoken digits (Fig. 4). A relevant family of ERP components, which aligns well with the evoked fronto-central positivity in the mixed- vs. single-talker speech, is the P3a/distractor positivity (Pd) components. These components are reliably known to appear around 250 ms after encountering a novel and potentially distracting stimulus, as demonstrated in both auditory and visual search tasks (Comerchero and Polich, 1999; Polich, 2007; Sawaki and Luck, 2010). Task-irrelevant novel sounds and/or distractors have been shown to disrupt ongoing task performance following the distractors; the P3a/Pd responses elicited by distractors reflect this involuntary attentional reorientation to a novel stimulus (or a change) in the perceptual environment (Polich, 2007; Escera et al., 1998; Sawaki and Luck, 2010). More specifically, a recent study demonstrated that during sequential encoding of sounds, Pd was elicited when listeners' auditory attention became distracted by a change in the task-irrelevant feature of a subsequent target sound, and that this Pd component manifested as a strong fronto-central positivity around 200–300 ms after the onset of the target sound (Stewart et al., 2017). This task and the Pd component closely resemble our finding; the increased positivity we found in the fronto-central scalp during encoding of mixed talkers' speech suggests that a discontinuity in a task-irrelevant feature (here, talker voice) of a task-relevant stimulus (the digits in the stream) is, on its own, enough to cause a performance-disrupting reorienting response. That is, changes in the task-irrelevant perceptual feature of voice disrupt the encoding of the incoming speech due to an involuntary reorienting to the novel speech source.

We found that both talker discontinuity and temporal discontinuity increased the neural signature of reorienting (the magnitude of the P3a/Pd), but that there was no significant interaction between these two main effects (Fig. 4**D**); instead, the effects of talker discontinuity and temporal proximity appear to be additive. It is notable that even when listeners were encoding speech spoken by one consistent talker, temporal discontinuity alone was sufficient to produce a robust P3a/Pd response. Temporal continuity of sounds plays an important role in the emergence of auditory streaming (van Noorden, 1975) and temporal gaps (e.g., a 500-ms ISI) between sounds can break down automatic and obligatory buildup of streaming. For speech, this breakdown with large ISIs can occur even if speech tokens are spoken by one consistent talker (Best et al., 2008; Bressler et al., 2014; Lim et al., 2019), and introducing temporal gaps in speech reduces the facilitatory effect of talker continuity (Choi and Perrachione, 2019). Our evidence suggests that any discontinuities in the acoustic stream, whether due to temporal or featural aspects of incoming sounds, trigger involuntary reorienting of attentional focus at the cortical level. Our finding is consistent with a domain-general stimulus-driven attention and auditory streaming account of talker-specific speech processing.

However, based on the evoked response waveforms of the mixed- vs. single-talker conditions, we can also consider an alternative interpretation of the observed significant effect as a change in a variant of N2. The N2 component has been implicated in the detection of stimulus novelty, attentional allocation, and cognitive control (e.g., Nieuwenhuis et al., 2004; Folstein and Van Petten, 2008). Especially in the domain of speech perception, recent findings observed enhanced N2 amplitudes for discriminating a pair of perceptually similar sounds that requires greater attention and cognitive control (Shao and Zhang, 2019; Zhang and Shao, 2018). Thus, the significant effect that we found might suggest that N2 responses are reduced in mixed- vs. single-talker speech because talker discontinuity disrupts attention and cognitive control. However, one caveat to this interpretation is that the scalp topography of the observed effect exhibits a pattern opposite that of the typical N2. While the typical auditory N2 responses arise as stronger negativity in the fronto-central sites (Nieuwenhuis et al., 2004; Folstein and Van Petten, 2008; Shao and Zhang, 2019; Zhang and Shao, 2018), we observed stronger positivity in the fronto-central scalp distributions (Fig. 4). This discrepancy might be due to differences in experimental task (i.e., N2 response observed in auditory oddball paradigms; Zhang and Shao, 2018); nevertheless, based on the opposite pattern of scalp topography distributions, we believe that the fronto-central positivity effect that we found more likely reflects a change in P3a/Pd, rather than in N2. While the most parsimonious interpretation of the observed fronto-central positivity based on its temporal and topographical properties is a change in P3a/Pd, the evoked response waveforms within the significant ERP cluster does not clearly differentiate N2 vs. P3a/Pd responses. To fully address this question, future studies of how talker discontinuity in speech affects N2 and P3a/P3 responses should directly manipulate stimulus dimensions and task variables known to preferentially drive N2 or P3a/Pd responses.

In addition to the neural evoked responses, we found that the phasic pupil dilation response during speech encoding was affected by talker discontinuity. We found increased pupil dilation responses when listeners encoded speech spoken by mixed talkers compared to one consistent talker, especially when listeners had to rapidly process incoming speech (i.e., for stimuli with the 0-ms ISI; Fig. 6). This observation is in line with recent findings that phasic pupil dilation is sensitive to detecting abrupt, bottom-up changes in the auditory environment that unfold on rapid timescales, but not for gradual and predictable changes in the environment (Zhao et al., 2019). Thus, our evidence provides further support for the idea that exogenous attentional reorientation is a critical cognitive mechanism engaged in processing speech in mixed-talker contexts. Furthermore, given the robust link between pupil dilation and activity in the LC-NE system (Aston-Jones and Cohen, 2005; Joshi et al., 2016; Costa and Rudebeck, 2016), our results suggest that rapid processing of mixed-talkers' speech is mediated by the pupil-

linked LC-NE response.

Our experiment provides the first psychophysiological evidence that a specific, domain-general neural system for attention and arousal influences cognitive operations during processing of talker variability in speech. One of the functional roles of the pupil-linked LC-NE system is to maintain and update the perceptual model of the surrounding environment (Bouret and Sara, 2005; Dayan and Yu, 2006; Sara and Bouret, 2012). The pupil dilation response serves as an index for NE release within the LC (Aston-Jones and Cohen, 2005; Samuels and Szabadi, 2008; Gilzenrat et al., 2010). NE is the neuromodulator known to serve as an interruption signal that halts ongoing top-down processes in the brain; this in turn, allows the system to automatically and exogenously switch attention to abrupt changes in sensory surroundings (Dayan and Yu, 2006; Sara and Bouret, 2012; Bouret and Sara, 2005). Consistent with this account, pupil dilation responses not only reflect the degree to which salient and surprising sound events draw attention involuntarily (Liao et al., 2016a; Huang and Elhilali, 2017), but also specifically track rapid and unexpected changes in the auditory environmental structure even when the changes are behaviorally irrelevant (Zhao et al., 2019). Our results suggest a novel, but consistent perspective on parsing speech with talker discontinuity (and hence talker variability): mixed talkers' speech appears to involuntarily reorient listeners' attention to the acoustic features of an unexpected, novel speech source (i.e., talker), and this process is, in part, accomplished by the pupil-linked LC-NE system.

Using the two distinct neurophysiological measures of attentional reorientation obtained through simultaneous EEG and pupillometry recording, we demonstrated that talker discontinuity in speech evokes signals that are consistent with stimulus-driven disruptions to listeners' attentional focus. Interestingly, while we found a robust evoked component (i.e., P3a/Pd) for parsing mixed-talker speech irrespective of the temporal continuity of speech, the phasic pupil dilation response differences in mixed- vs. single-talker speech were significantly larger for processing the temporally continuous (0-ms ISI) than the discontinuous (500-ms ISI) speech. This difference between these two measures is not surprising given that they reflect distinct neural mechanisms, and that they unfold over different timescales. Our findings using these two measures have notable implications for understanding how neural systems are impacted by talker discontinuity in speech. Encountering any talker discontinuity disrupts auditory streaming by switching listeners' attentional focus to the features of the novel source of speech; that is, although listeners have enough time (with longer gaps; e.g., 500-ms ISI) to encode and switch attention to each speech token, attentional reorientation to the newly encountered talker is inevitable. However, only when such discontinuity occurs very close in time (e.g., 0-ms ISI), does the demand to quickly switch attention seem to additionally trigger the LC-NE system in order to prioritize tracking of the changes in bottom-up sensory information.

Following the typical approaches used in prior studies of talker variability (e.g., Choi et al., 2018; Mullennix and Pisoni, 1990; Nusbaum and Morin, 1992; Nusbaum and Magnuson, 1997; Shao and Zhang, 2019; Zhang et al., 2013), the current study created quite extreme listening situations, in which listeners must constantly accommodate talker switches (thus, variability) on every spoken word. While these specific task demands have limited ecological parallels (e.g., a waiter taking the drink order of each person at a table with many diners), similar demands are in fact ubiquitous in everyday auditory scenes where listeners encounter frequent talker discontinuities during speech processing (e.g., following a conversation that switches from one talker to another in a meeting, at a party, or on television). Thus, our findings can provide insight about listeners' speech processing efficiency in the context of naturalistic listening conditions, and can guide future studies to test generalizability of this finding in more naturalistic listening environments.

4.3. Consideration of other potential explanations for the cost of processing mixed talkers' speech

Among previous psycholinguistic models that have been proposed to explain how talker variability is accommodated during speech processing, each posits a different cognitive mechanism to account for the additional processing cost of mixed-talker speech. These mechanistic explanations include uncertainty in accessing long-term, talker-specific memory representations (Kleinschmidt and Jaeger, 2015), increased computational complexity in resolving acoustic–phonetic ambiguity (Nearey, 1989; Johnson, 2005), and higher demand for working memory resources to maintain multiple, potential acoustic–phonetic interpretations of upcoming speech (Nusbaum and Magnuson, 1997). However, in this study we did not see clear evidence of the sorts of neurophysiological signatures one would expect to be associated with the cognitive mechanisms posited by these models in either the ERP or neural oscillatory changes measured in response to talker variability.

The classical model describing how listeners resolve acoustic–phonetic ambiguities introduced by cross-talker differences is called *talker normalization* (e.g., Johnson, 2005; Nusbaum and Magnuson, 1997; Pisoni, 1997). This framework posits that listeners actively utilize intrinsic information (i.e., acoustic patterns) of speech signal (Ainsworth, 1975; Nearey 1989; Syrdal and Gopal, 1986) and extrinsic information from preceding speech context (e.g., Johnson 2005; Sjerps et al., 2013) to configure a talker-specific mapping between acoustics and phonemic representations. Thus, processing mixed-talker speech would lead to higher computational demands in order to continuously adjust and "normalize" the phonemic representations of each successive talker based on their unique acoustic features (i.e., speech formants; Sussman, 1986). According to this explanation, the computational cost of processing differences in mixed- vs. single-talker's speech should likely manifest in early auditory processing.

One neural signature likely relevant to talker normalization is the magnitude of the N1 component of the auditory ERP, arising around 100 ms after a sound onset. Extensive research has demonstrated that the auditory N1 response, associated with basic auditory perception (Hillyard and Picton, 1978; Näätänen and Picton, 1987; Pantev et al., 1996; Hyde, 1997; Martin et al., 2008), is modulated by top-down attention (Hillyard et al., 1973, 1998; Choi et al., 2013, 2014; Woldorff et al., 1993), and reflects the extent of neural adaptation in the auditory cortex (Maess et al, 2007; Pantev et al., 1988). As the N1 response magnitude is sensitive to changes in the acoustic features (repeating vs. non-repeating sound events; Todorovic et al., 2011; Herrmann et al., 2015), we would expect that the computational demands of forming talker-specific acoustic–phonetic representations in mixed- vs. single-talker's speech would be reflected in differences in N1 magnitude. However, we did not find any such differences between the two talker conditions (Fig. 4).

Although we did not observe any difference in the N1 magnitude between the talker conditions, it is worth noting some considerations regarding the N1 response, as the existing evidence is somewhat mixed. Previous EEG studies in the context of talker discontinuity (hence, variability) in speech have observed larger N1 magnitudes when parsing mixed-talker compared to single-talker speech (e.g., Kaganovich et al., 2006; Uddin et al., 2020; Zhang et al., 2013), or when listeners encountered an unexpected change in the talker of an attended speech stream (Mehraei et al., 2018). In contrast, other work has reported the opposite pattern: processing mixed talker speech either reduced (Shao and Zhang, 2019) or did not affect N1 magnitude (Zhang et al., 2013). One potential source of this discrepancy might arise from the amount of acoustic variability in the stimulus set, as the early evoked neural responses up to ~200 ms post-stimulus are contingent on the characteristics of the stimulus (Näätänen and Picton, 1987; Digeser et al., 2009; Khalighinejad et al., 2017; Tremblay et al., 2003). Given the many factors that affect the N1, future studies are necessary to disambiguate the unique contributions of acoustic variability across and within talkers

to markers of early auditory processing like the N1—as well as to speech processing efficiency.

Our findings on how talker discontinuity influences neural alpha oscillatory power do not seem consistent with the cognitive mechanism proposed by the other prominent model accounting for resolving talker differences in speech, that processing talker variability in speech is handled by an *active control process* (Nusbaum and Nusbaum, 1997). This model posits that listeners pre-allocate cognitive resources in order to flexibly resolve potential acoustic-phonemic ambiguity in speech signals, and to maintain robust speech perception accuracy in the face of variation (Nusbaum and Magnuson, 1997; Magnuson and Nusbaum, 2007; Heald et al., 2014). Thus, when listeners encounter mixed talkers' speech, there is greater demand on working memory as listeners must simultaneously entertain multiple interpretations of incoming speech simultaneously in working memory. From this account, processing mixed talkers' speech should presumably enhance alpha oscillatory power during speech encoding, as increasing working memory demand leads to parametric increases in alpha power (Jensen et al., 2002; Tuladhar et al., 2007; Obleser et al., 2012). However, we did not find any alpha power differences between the single and mixed talker conditions when listeners encoded the speech. Instead, we found the opposite pattern—a decrease in alpha power for mixed- compared to single-talker speech—during speech memory retention (Fig. 5). This finding is inconsistent with increased working memory load as a mechanism for accommodating talker variability as suggested by the active control framework, but is consistent with the attentional enhancement of working memory maintenance (Lim et al., 2015, 2018).

One potential interpretation of our finding that alpha power during retention is lower for mixed- vs. single-talker speech sequences is that talker discontinuity disrupts attention directed to working memory. Attention enables effective encoding and maintenance of relevant information in working memory (Awh and Jonides, 2001; Serences and Kastner, 2014; Gazzaley and Nobre, 2012), whereas attentional disruption can impair working memory maintenance. Also, attention directed to working memory items improves memory recall of stored items (Oberauer and Hein, 2012; Lim et al., 2018; 2021), and the degree of attentional benefit on memory recall can be related to alpha oscillatory power enhancement (Lim et al., 2015). As auditory-based attention depends on the ability to form a coherent auditory object (Shinn-Cunningham, 2008), it is possible that emergence of a single auditory stream via talker- and temporal- continuity might also facilitate efficient storage of the resulting coherent object in memory. In contrast, talker and/or temporal discontinuities that break down streaming may impair efficient storage and allocation of attention to working memory objects.

Although our present results are not consistent with the predictions of the active control framework (Nusbaum & Magnuson, 1997), it is worth noting that this framework may not be mutually exclusive with the auditory streaming account. Recent behavioral research demonstrates that both auditory streaming and active control processes appear to work in parallel to support processing of talker variability in speech (Kapadia & Perrachione, 2020; Choi, Kuo, & Perrachione, 2020). For instance, when listeners must parse speech in the presence of background noise or when they have an ongoing expectation of uncertainty about the upcoming talker, they seem to pre-allocate cognitive resources to cope with the ongoing listening challenge. This enhanced cognitive load can tie up available resources, which can then impede the attentional benefit that listeners otherwise gain when processing speech that is continuous in talker (Kapadia & Perrachione, 2020).

Finally, the *ideal adapter framework* (Kleinschmidt & Jaeger, 2015) formalizes the longstanding episodic approach to understanding processing variability in speech perception (e.g., Goldinger, 1996, 1998). These approaches assert that listeners maintain long-term memory representations of talkers' speech, against which the incoming speech signal is compared to support recognition. According to this framework, the additional processing cost of mixed-talker speech is associated with the greater number of possible competing interpretations (or "models") of a given speech token. By anticipating the speech of a particular talker (or class of talker) in advance, listeners can reduce the decision space to a subset of these (e.g., talker-specific) models, leading to more efficient speech processing. However, the current formulation of this framework does not address how its model-selection operations might be implemented in specific neurobiological processes that are distinct from the input or processing predictions made by talker normalization or the active control hypotheses, as discussed above.

Using scalp EEG and pupillometry, the current study did not find evidence clearly consistent with predictions made by other psycholinguistic accounts of accommodating talker variability in speech, such as computational complexity in acoustic–phonetic mappings (Nearey, 1989; Johnson 2005), allocation of cognitive resources (Nusbaum and Magnuson, 1997), accessing long-term memory (Kleinschmidt and Jaeger, 2015). However, it is important to note that the current study also does not completely rule out the mechanisms posited by these accounts. As mentioned above, the mechanisms are likely not mutually exclusive with each other, as they can work in parallel (e.g., Kapadia and Perrachione, 2020). Other neurophysiological signals and methods may be more appropriate or sensitive to the mechanisms invoked by other accounts (e.g., talker normalization using intracranial recordings; Sjerps et al., 2019). In order to better understand the neurobiologically plausible processes engaged in accommodating talker variability, accounts of processing talker variability in speech must be made more explicit about the specific neurophysiological mechanisms that would support the cognitive operations they propose. In turn, future studies will be necessary to test those predictions, using the neuroscientific methods most appropriate to capturing such mechanisms.

5. Conclusions

The present work shows that talker discontinuity in speech interferes with both immediate processing of, and subsequent working memory for, speech. Differences in ERP and pupil dilation responses suggest that the behavioral costs associated with processing variable, mixed-talker speech are the result of added auditory processing demands incurred by automatic attentional reorientation to the new source of speech upon talker discontinuity. Neural alpha oscillatory power results suggest that the interference effect of talker variability in speech is also present when listeners maintain speech information in working memory after speech encoding. Collectively, our results demonstrate that talker changes evoke an involuntary reorientation of attention via domain-general processes, which interact with the pupil-linked LC-NE system in determining processing efficiency of mixed-talker vs. single-talker speech.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by NIH grants (R03DC14045 to TKP, R01DC009477 to BGSC, and R01DC004545 to Gerald Kidd), a Brain and Behavior Research Foundation NARSAD Young Investigator grant to TKP, and the Office of Naval Research grant (N00014-20-1-2709) to BGSC. SJL was supported by grants from the NIH (T32DC013017) and the NSF (BCS 1840674). We are grateful to the two anonymous reviewers, and to Anna Kasdan and Dr. Andre Cravo for providing the open review on the preprint of this work via the SfN Review Mentoring Program.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.

org/10.1016/j.bandl.2021.104996.

References

Ainsworth, W. (1975). Intrinsic and extrinsic factors in vowel judgments. In G. Fant & M. Tatham (Eds.), Auditory analysis and perception of speech (pp. 103–113). London.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience, 28*(1), 403–450. https://doi.org/10.1146/annurev.neuro.28.061604.135709.

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *Trends in Cognitive Sciences, 5*(3), 119–126.

Bala, A. D. S., & Takahashi, T. T. (2000). Pupillary dilation response as an indicator of auditory discrimination in the barn owl. *Journal of Comparative Physiology, 186*, 425–434.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision, 10*(4), 433–436. https://doi.org/10.1163/156856897X00357.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In *Handbook of psychophysiology* (2nd ed., pp. 142–162). New York, NY, US: Cambridge University Press.

Belin, P., & Zatorre, R. J. (2003). Adaptation to speaker's voice in right anterior temporal lobe. *Neuroreport, 14*(16), 2105–2109. https://doi.org/10.1097/01.wnr.0000091689.94870.85.

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.

Best, V., Swaminathan, J., Kopčo, N., Roverud, E., & Shinn-Cunningham, B. (2018). A "Buildup" of Speech Intelligibility in Listeners With Normal Hearing and Hearing Loss, 233121651880751 *Trends in Hearing, 22*. https://doi.org/10.1177/2331216518807519.

Bouret, S., & Sara, S. J. (2005). Network reset: A simplified overarching theory of locus coeruleus noradrenaline function. *Trends in Neurosciences, 28*(11), 574–582. https://doi.org/10.1016/j.tins.2005.09.002.

Bregman, A. S. (1990). *Auditory Scene Analysis*. Cambridge, MA: MIT Press.

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.

Carter, Y. D., Lim, S.-J., & Perrachione, T. K. (2019). In *Talker continuity facilitates speech processing independent of listeners' expectations* (pp. 1620–1624). Canberra, Australia: Australasian Speech Science and Technology Association Inc.

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, I., Wang, L.e., Bharadwaj, H., & Shinn-Cunningham, B. (2014). Individual differences in attentional modulation of cortical responses correlate with selective attention performance. *Hearing Research, 314*(C), 10–19. https://doi.org/10.1016/j.heares.2014.04.008.

Choi, I., Rajaram, S., Varghese, L. A., & Shinn-Cunningham, B. G. (2013). Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography. *Frontiers in Human Neuroscience, 7*, 115. https://doi.org/10.3389/fnhum.2013.00115.

Choi, J. Y., Kou, R. S., & Perrachione, T. K. (2020). Parametrically varying speech adapter length suggests two mechanisms for talker adaptation. *Journal of the Acoustical Society of America, 148*(4). https://doi.org/10.1121/1.5146957, 2505–2505.

Choi, J. Y., & Perrachione, T. K. (2019). Time and information in perceptual adaptation to speech. *Cognition, 192*, Article 103982. https://doi.org/10.1016/j.cognition.2019.05.019.

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.

Comerchero, M. D., & Polich, J. (1998). P3a, perceptual distinctiveness, and stimulus modality. *Cognitive Brain Research, 7*(1), 41–48. https://doi.org/10.1016/S0926-6410(98)00009-3.

Comerchero, M. D., & Polich, J. (1999). P3a and P3b from typical auditory and visual stimuli. *Clinical Neurophysiology, 110*, 24–30.

Costa, V. D., & Rudebeck, P. H. (2016). More than Meets the Eye: The Relationship between Pupil Size and Locus Coeruleus Activity. *Neuron, 89*(1), 8–10. https://doi.org/10.1016/j.neuron.2015.12.031.

Dayan, P., & Yu, A. J. (2006). Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network: Computation in Neural Systems, 17*(4), 335–350. https://doi.org/10.1080/09548980601004024.

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods, 134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009.

Digeser, F. M., Wohlberedt, T., & Hoppe, U. (2009). Contribution of Spectrotemporal Features on Auditory Event-Related Potentials Elicited by Consonant-Vowel Syllables. *Ear and Hearing, 30*(6).

Donchin, E., Spencer, K. M., & Dien, J. (1997). The Varieties of Deviant Experience: ERP Manifestations of Deviance Processors. In G. J. M. van Boxtel & K. B. M. Bocker (Eds.), Brain and Behavior Past, Present, and Future (pp. 67–91).

Escera, C., Alho, K., Winkler, I., & Naatanen, R. (1998). Neural Mechanisms of Involuntary Attention to Acoustic Novelty and Change. *Journal of Cognitive Neuroscience, 10*(5), 590–604.

Folstein, J. R., & Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: A review. *Psychophysiology, 45*, 152–170. https://doi.org/10.1111/j.1469-8986.2007.00628.x.

Foxe, J. J., & Snyder, A. C. (2011). The role of alpha-band brain oscillations as a sensory suppression mechanism during selective attention. *Frontiers in Psychology, 2*. https://doi.org/10.3389/fpsyg.2011.00154.

Gaeta, H., Friedman, D., & Hunt, G. (2003). Stimulus characteristics and task category dissociate the anterior and posterior aspects of the novelty P3. *Psychophysiology, 40*, 198–208.

Gazzaley, A., & Nobre, A. C. (2012). Top-down modulation: Bridging selective attention and working memory. *Trends in Cognitive Sciences, 16*(2), 129–135. https://doi.org/10.1016/j.tics.2011.11.014.

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective & Behavioral Neuroscience, 10*(2), 252–269. https://doi.org/10.3758/CABN.10.2.252.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1166–1183. https://doi.org/10.1037//0278-7393.22.5.1166.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279. https://doi.org/10.1037//0033-295X.105.2.251.

Goldstein, A., Spencer, K. M., & Donchin, E. (2002). The influence of stimulus deviance and novelty on the P300 and Novelty P3. *Psychophysiology, 39*, 781–790.

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: Neural models of stimulus-specific effects. *Trends in Cognitive Sciences, 10*(1), 14–23. https://doi.org/10.1016/j.tics.2005.11.006.

Heald, S., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience, 8*, 1–15. https://doi.org/10.3389/fnsys.2014.00035.

Heitz, R. P., Schrock, J. C., Payne, T. W., & Engle, R. W. (2008). Effects of incentive on working memory capacity: Behavioral and pupillometric data. *Psychophysiology, 45*, 119–129. https://doi.org/10.1111/j.1469-8986.2007.00605.x.

Herrmann, B., Henry, M. J., Fromboluti, E. K., McAuley, J. D., & Obleser, J. (2015). Statistical context shapes stimulus-specific adaptation in human auditory cortex. *Journal of Neurophysiology, 113*(7), 2582–2591. https://doi.org/10.1152/jn.00634.2014.

Hickey, C., Di Lollo, V., & McDonald, J. J. (2009). Electrophysiological indices of target and distractor processing in visual search. *Journal of Cognitive Neuroscience, 21*(4), 760–775. https://doi.org/10.1162/jocn.2009.21039.

Hilimire, M. R., Mounts, J. R. W., Parks, N. A., & Corballis, P. M. (2011). Dynamics of target and distractor processing in visual search: Evidence from event-related brain potentials. *Neuroscience Letters, 495*(3), 196–200. https://doi.org/10.1016/j.neulet.2011.03.064.

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.

Hillyard, S. A., & Picton, T. W. (1978). ON and OFF components in the auditory evoked potential. *Perception & Psychophysics, 24*(5), 391–398.

Hillyard, S. A., Hink, R. F., Schwent, V. L., & Picton, T. W. (1973). Electrical Signs of Selective Attention in the Human Brain. *Science, 182*(4108), 177–180. https://doi.org/10.1126/science.182.4108.177.

Hillyard, S. A., Vogel, E. K., & Luck, S. J. (1998). Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 353*(1373), 1257–1270. https://doi.org/10.1098/rstb.1998.0281.

Hoeks, B., & Levelt, W. J. M. (1993). Pupillary dilation as a measure of attention: A quantitative system analysis. *Behavior Research Methods, Instruments, & Computers, 25*(1), 16–26.

Hong, L., Walz, J. M., & Sajda, P. (2014). Your eyes give you away: Prestimulus changes in pupil diameter correlate with poststimulus task-related EEG dynamics. *PloS one, 9*(3), Article e91321. https://doi.org/10.1371/journal.pone.0091321.

Huang, N., & Elhilali, M. (2017). Auditory salience using natural soundscapes. *The Journal of the Acoustical Society of America, 141*(3), 2163. https://doi.org/10.1121/1.4979055.

Hyde, M. (1997). The N1 Response and Its Applications. *Audiology and Neurotology, 2*(5), 281–307.

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.

Jensen, O., & Mazaheri, A. (2010). Shaping Functional Architecture by Oscillatory Alpha Activity: Gating by Inhibition. *Frontiers in Human Neuroscience, 4*, 186. https://doi.org/10.3389/fnhum.2010.00186.

Jensen, O., Gelfand, J., Kounios, J., & Lisman, J. E. (2002). Oscillations in the Alpha Band (9–12 Hz) Increase with Memory Load during Retention in a Short-term Memory Task. *Cerebral Cortex, 12*, 877–882.

Johnson, D. A. (1971). Pupillary responses during a short-term memory task: Cognitive processing, arousal, or both? *Journal of Experimental Psychology, 90*(2), 311–318. https://doi.org/10.1037/h0031562.

Johnson, E., Miller Singley, A., Peckham, A., Johnson, S., & Bunge, S. (2014). Task-evoked pupillometry provides a window into the development of short-term memory capacity. *Frontiers in Psychology, 5*, 218. https://doi.org/10.3389/fpsyg.2014.00218.

Johnson, K. (2005). Speaker Normalization in Speech Perception. In D. B. Pisoni, & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 363–389). Malden, MA: John Wiley & Sons, Ltd.. https://doi.org/10.1002/9780470757024.ch15

Joshi, S., Li, Y., Kalwani, R. M., & Gold, J. I. (2016). Relationships between Pupil Diameter and Neuronal Activity in the Locus Coeruleus, Colliculi, and Cingulate Cortex. *Neuron, 89*(1), 221–234. https://doi.org/10.1016/j.neuron.2015.11.028.

Kaganovich, N., Francis, A. L., & Melara, R. D. (2006). Electrophysiological evidence for early interaction between talker and linguistic information during speech perception. *Brain Research, 1114*(1), 161–172. https://doi.org/10.1016/j.brainres.2006.07.049.

Kahneman, D., & Beatty, J. (1966). Pupil diameter and load on memory. *Science, 154*(3756), 1583–1585.

Kamp, S. M., & Donchin, E. (2015). ERP and pupil responses to deviance in an oddball paradigm. *Psychophysiology, 52*(4), 460–471. https://doi.org/10.1111/psyp.12378.

Kapadia, A. M., & Perrachione, T. K. (2020). Selecting among competing models of talker adaptation_ Attention, cognition, and memory in speech processing efficiency. *Cognition, 204*, Article 104393. https://doi.org/10.1016/j.cognition.2020.104393.

Khalighinejad, B., Cruzatto da Silva, G., & Mesgarani, N. (2017). Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience, 37*(8), 2176–2185. https://doi.org/10.1523/JNEUROSCI.2383-16.2017.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in psychtoolbox-3. *Perception, 36*(14), 1–16.

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), 148–203. https://doi.org/10.1037/a0038695.

Klimesch, W., Sauseng, P., & Hanslmayr, S. (2007). EEG alpha oscillations: The inhibition–timing hypothesis. *Brain Research Reviews, 53*(1), 63–88. https://doi.org/10.1016/j.brainresrev.2006.06.003.

Liao, H.-I., Kidani, S., Yoneya, M., Kashino, M., & Furukawa, S. (2016). Correspondences among pupillary dilation response, subjective salience of sounds, and loudness. *Psychonomic Bulletin & Review, 23*(2), 412–425. https://doi.org/10.3758/s13423-015-0898-0.

Liao, H.-I., Yoneya, M., Kidani, S., Kashino, M., & Furukawa, S. (2016). Human Pupillary Dilation Response to Deviant Auditory Stimuli: Effects of Stimulus Properties and Voluntary Attention. *Frontiers in Neuroscience, 10*(89), 403. https://doi.org/10.3389/fnins.2016.00043.

Lim, S.-J., Shinn-Cunningham, B. G., & Perrachione, T. K. (2019). Effects of talker continuity and speech rate on auditory working memory. *Attention, Perception, & Psychophysics, 138*(2), 571–611. https://doi.org/10.3758/s13414-019-01684-w.

Lim, S.-J., Thiel, C., Sehm, B., Deserno, L., Lepsien, J., & Obleser, J. (2021). Distributed networks for auditory memory contribute differentially to recall precision. *bioRxiv, 5*, d202–d236. https://doi.org/10.1101/2021.01.18.427143.

Lim, S.-J., Wöstmann, M., & Obleser, J. (2015). Selective Attention to Auditory Memory Neurally Enhances Perceptual Precision. *The Journal of Neuroscience : the Official Journal of the Society for 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.

Maess, B., Jacobsen, T., Schröger, E., & Friederici, A. D. (2007). Localizing pre-attentive auditory memory-based comparison: Magnetic mismatch negativity to pitch change. *NeuroImage, 37*(2), 561–571. https://doi.org/10.1016/j.neuroimage.2007.05.040.

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.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024.

Martin, B. A., Tremblay, K. L., & Korczak, P. (2008). Speech Evoked Potentials: From the Laboratory to the Clinic. *Ear and Hearing, 29*(3).

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.

Mehraei, G., Shinn-Cunningham, B., & Dau, T. (2018). Influence of talker discontinuity on cortical dynamics of auditory spatial attention. *NeuroImage, 179*, 548–556. https://doi.org/10.1016/j.neuroimage.2018.06.067.

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.

Murphy, P. R., Robertson, I. H., Balsters, J. H., & O'connell, R. G. (2011). Pupillometry and P3 index the locus coeruleus-noradrenergic arousal function in humans. *Psychophysiology, 48*(11), 1532–1543. https://doi.org/10.1111/j.1469-8986.2011.01226.x.

Näätänen, R., & Picton, T. (1987). The N1 Wave of the Human Electric and Magnetic Response to Sound: A Review and an Analysis of the Component Structure. *Psychophysiology, 24*(4), 375–425. https://doi.org/10.1111/j.1469-8986.1987.tb00311.x.

Nearey, T. M. (1989). 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.

Nieuwenhuis, S., Yeung, N., & Cohen, J. D. (2004). Stimulus modality, perceptual overlap, and the go/no-go N2. *Psychophysiology., 41*(1), 157–160. https://doi.org/10.1046/j.1469-8986.2003.00128.x.

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). San Diego: 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.

Obleser, J., Wöstmann, M., Hellbernd, N., Wilsch, A., & Maess, B. (2012). Adverse listening conditions and memory load drive a common α oscillatory network. *The Journal of Neuroscience : the Official Journal of the Society for Neuroscience, 32*(36), 12376–12383. https://doi.org/10.1523/JNEUROSCI.4908-11.2012.

Oberauer, K., & Hein, L. (2012). Attention to Information in Working Memory. *Current Directions in Psychological Science, 21*(3), 164–169. https://doi.org/10.1177/0963721412444727.

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience, 2011*(1), 156869–156879. https://doi.org/10.1155/2011/156869.

Pantev, C., M, H., Lehnertz, K., Lütkenhöner, B., Anogianakis, G., & Wittkowski, W. (1988). Tonotopic organization of the human auditory cortex revealed by transient auditory evoked magnetic fields. Electroencephalography and Clinical Neurophysiology, 69, 160–170.

Pantev, C., Roberts, L. E., Elbert, T., Ross, B., & Wienbruch, C. (1996). Tonotopic organization of the sources of human auditory steady-state responses. *Hearing Research, 101*, 62–74.

Peavler, W. S. (1974). Pupil Size, Information Overload, and Performance Differences. *Psychophysiology, 11*(5), 559–566. https://doi.org/10.1111/j.1469-8986.1974.tb01114.x.

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.

Pierrehumbert, J. B. (2003). Phonetic Diversity, Statistical Learning, and Acquisition of Phonology. *Language and Speech, 46*, 115–154.

Pisoni, D. B. (1997). Some thoughts on "normalization" in speech perception. In K. Johnson, & J. W. Mullennix (Eds.), *Talker variability and speech processing* (pp. 9–32). San Diego: Academic Press.

Polich, J. (2007). Updating P300: An integrative theory of P3a and P3b. *Clinical Neurophysiology, 118*(10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019.

Samuels, E. R., & Szabadi, E. (2008). Functional Neuroanatomy of the Noradrenergic Locus Coeruleus: Its Roles in the Regulation of Arousal and Autonomic Function Part II: Physiological and Pharmacological Manipulations and Pathological Alterations of Locus Coeruleus Activity in Humans. *Current Neuropharmacology, 6*, 254–285.

Sara, S. J. (2009). The locus coeruleus and noradrenergic modulation of cognition. *Nature Reviews. Neuroscience, 10*(3), 211–223. https://doi.org/10.1038/nrn2573.

Sara, S. J., & Bouret, S. (2012). Orienting and Reorienting: The Locus Coeruleus Mediates Cognition through Arousal. *Neuron, 76*(1), 130–141. https://doi.org/10.1016/j.neuron.2012.09.011.

Sawaki, R., & Luck, S. J. (2010). Capture versus suppression of attention by salient singletons: Electrophysiological evidence for an automatic attend-to-me signal. *Attention, Perception, & Psychophysics, 72*(6), 1455–1470. https://doi.org/10.3758/APP.72.6.1455.

Scott, T. L., & Perrachione, T. K. (2019). Common cortical architectures for phonological working memory identified in individual brains. *NeuroImage, 202*, Article 116096. https://doi.org/10.1016/j.neuroimage.2019.116096.

Serences, J. T., & Kastner, S. (2014). A multi-level account of selective attention. In A. C. Nobre, & S. Kastner (Eds.), *The Oxford Handbook of Attention* (pp. 76–104). Oxford University Press.

Shao, J., & Zhang, C. (2019). Talker normalization in typical Cantonese-speaking listeners and congenital amusics: Evidence from event-related potentials. *NeuroImage: Clinical, 23*, Article 101814. https://doi.org/10.1016/j.nicl.2019.101814.

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.

Sjerps, M. J., McQueen, J. M., & Mitterer, H. (2013). Evidence for precategorical extrinsic vowel normalization. *Attention, Perception, & Psychophysics, 75*(3), 576–587. https://doi.org/10.3758/s13414-012-0408-7.

Sjerps, M. J., Fox, N. P., Johnson, K., & Chang, E. F. (2019). Speaker-normalized sound representations in the human auditory cortex. *Nature Communications, 10*, 2465. https://doi.org/10.1038/s41467-019-10365-z.

Stewart, H. J., Amitay, S., & Alain, C. (2017). Neural correlates of distraction and conflict resolution for nonverbal auditory events. *Scientific Reports, 7*(1), 1–11. https://doi.org/10.1038/s41598-017-00811-7.

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.

Sussman, H. M. (1986). A neuronal model of vowel normalization and representation. *Brain and Language, 28*(1), 12–23. https://doi.org/10.1016/0093-934X(86)90087-8.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *The Journal of the Acoustical Society of America, 79*(4), 1086–1100. https://doi.org/10.1121/1.393381.

Thut, G., Nietzel, A., Brandt, S. A., & Pascual-Leone, A. (2006). α-Band Electroencephalographic Activity over Occipital Cortex Indexes Visuospatial

Attention Bias and Predicts Visual Target Detection. *Journal of Neuroscience, 26*(37), 9494–9502. https://doi.org/10.1523/JNEUROSCI.0875-06.2006.

Todorovic, A., van Ede, F., Maris, E., & de Lange, F. P. (2011). Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: An MEG study. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 31*(25), 9118–9123. https://doi.org/10.1523/JNEUROSCI.1425-11.2011.

Tremblay, K. L., Friesen, L., Martin, B. A., & Wright, R. (2003). Test-Retest Reliability of Cortical Evoked Potentials Using Naturally Produced Speech Sounds. *Ear and Hearing, 24*(3), 225–232. https://doi.org/10.1097/01.AUD.0000069229.84883.03.

Tuladhar, A. M., Huurne, N. T., Schoffelen, J.-M., Maris, E., Oostenveld, R., & Jensen, O. (2007). Parieto-occipital sources account for the increase in alpha activity with working memory load. *Human Brain Mapping, 28*(8), 785–792. https://doi.org/10.1002/hbm.20306.

Uddin, S., Reis, K. S., Heald, S. L. M., Van Hedger, S. C., & Nusbaum, H. C. (2020). Cortical mechanisms of talker normalization in fluent sentences. *Brain and Language, 201*, Article 104722. https://doi.org/10.1016/j.bandl.2019.104722.

Unsworth, N., & Robison, M. K. (2015). Individual differences in the allocation of attention to items in working memory: Evidence from pupillometry. *Psychonomic Bulletin & Review, 22*(3), 757–765. https://doi.org/10.3758/s13423-014-0747-6.

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. http://doi.org/10.6100/IR152538.

Verney, S. P., Granholm, E., & Marshall, S. P. (2004). Pupillary responses on the visual backward masking task reflect general cognitive ability. *International Journal of Psychophysiology, 52*(1), 23–36. https://doi.org/10.1016/j.ijpsycho.2003.12.003.

Wark, B., Lundstrom, B. N., & Fairhall, A. (2007). Sensory adaptation. *Current Opinion in Neurobiology, 17*(4), 423–429. https://doi.org/10.1016/j.conb.2007.07.001.

Winkler, I., Denham, S. L., & Nelken, I. (2009). Modeling the auditory scene: Predictive regularity representations and perceptual objects. *Trends in Cognitive Sciences, 13*(12), 532–540. https://doi.org/10.1016/j.tics.2009.09.003.

Winn, M. B., Edwards, J. R., & Litovsky, R. Y. (2015). The Impact of Auditory Spectral Resolution on Listening Effort Revealed by Pupil Dilation. *Ear and Hearing, 36*, e153–e165. https://doi.org/10.1097/AUD.0000000000000145>.

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started. *Trends in Hearing, 22*, 1–32. https://doi.org/10.1177/2331216518800869.

Woldorff, M. G., Gallen, C. C., Hampson, S. A., Hillyard, S. A., Pantev, C., Sobel, D., et al. (1993). Modulation of early sensory processing in human auditory cortex during auditory selective attention. *Proceedings of the National Academy of Sciences, 90*(18), 8722–8726. https://doi.org/10.1073/pnas.90.18.8722.

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

Wöstmann, M., Herrmann, B., Maess, B., & Obleser, J. (2016). Spatiotemporal dynamics of auditory attention synchronize with speech. *Proceedings of the National Academy of Sciences of the United States of America, 113*(14), 3873–3878. https://doi.org/10.1073/pnas.1523357113.

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.

Wöstmann, M., Schröger, E., & Obleser, J. (2015). Acoustic Detail Guides Attention Allocation in a Selective Listening Task. *Journal of Cognitive Neuroscience, 27*(5), 988–1000. https://doi.org/10.1162/jocn_a_00761.

Zhang, C., Peng, G., & Wang, W. S. Y. (2013). Achieving constancy in spoken word identification: Time course of talker normalization. *Brain and Language, 126*(2), 193–202. https://doi.org/10.1016/j.bandl.2013.05.010.

Zhang, C., & Shao, J. (2018). Normal pre-attentive and impaired attentive processing of lexical tones in Cantonese-speaking congenital amusics. *Scientific reports, 8*(1), 8420. https://doi.org/10.1038/s41598-018-26368-7.

Zhao, S., Chait, M., Dick, F., Dayan, P., Furukawa, S., & Liao, H.-I. (2019). Pupil-linked phasic arousal evoked by violation but not emergence of regularity within rapid sound sequences. *Nature Communications, 10*, 4030.