Text Captioning Buffers Against the Effects of Background Noise and Hearing Loss on Memory for Speech

Brennan R. Payne,[1,2,4] Jack W Silcox,[1] Hannah A. Crandell,[1] Amanda Lash,[3]
Sarah Hargus Ferguson,[4] and Monika Lohani[1,5]

Objective: Everyday speech understanding frequently occurs in perceptually demanding environments, for example, due to background noise and normal age-related hearing loss. The resulting degraded speech signals increase listening effort, which gives rise to negative downstream effects on subsequent memory and comprehension, even when speech is intelligible. In two experiments, we explored whether the presentation of realistic assistive text captioned speech offsets the negative effects of background noise and hearing impairment on multiple measures of speech memory.

Design: In Experiment 1, young normal-hearing adults (*N* = 48) listened to sentences for immediate recall and delayed recognition memory. Speech was presented in quiet or in two levels of background noise. Sentences were either presented as speech only or as text captioned speech. Thus, the experiment followed a 2 (caption vs no caption) × 3 (no noise, +7 dB signal-to-noise ratio, +3 dB signal-to-noise ratio) within-subjects design. In Experiment 2, a group of older adults (age range: 61 to 80, *N* = 31), with varying levels of hearing acuity completed the same experimental task as in Experiment 1. For both experiments, immediate recall, recognition memory accuracy, and recognition memory confidence were analyzed via general(ized) linear mixed-effects models. In addition, we examined individual differences as a function of hearing acuity in Experiment 2.

Results: In Experiment 1, we found that the presentation of realistic text-captioned speech in young normal-hearing listeners showed improved immediate recall and delayed recognition memory accuracy and confidence compared with speech alone. Moreover, text captions attenuated the negative effects of background noise on all speech memory outcomes. In Experiment 2, we replicated the same pattern of results in a sample of older adults with varying levels of hearing acuity. Moreover, we showed that the negative effects of hearing loss on speech memory in older adulthood were attenuated by the presentation of text captions.

Conclusions: Collectively, these findings strongly suggest that the simultaneous presentation of text can offset the negative effects of effortful listening on speech memory. Critically, captioning benefits extended from immediate word recall to long-term sentence recognition memory, a benefit that was observed not only for older adults with hearing loss but also young normal-hearing listeners. These findings suggest that the text captioning benefit to memory is robust and has potentially wide applications for supporting speech listening in acoustically challenging environments.

Key words: Memory, Text Captioning, Noise, Speech, Aging

(Ear & Hearing XXX;XX;00–00)

INTRODUCTION

Age-related changes in sensory and cognitive functioning can have a profoundly negative effect on comprehension and memory for language. One of the most striking examples of this is the negative effect of age-related sensorineural hearing loss (i.e., presbyacusis) on speech understanding. By some estimates, sensorineural hearing loss is the third-most-prevalent chronic medical condition in older adults (after arthritis and hypertension), afflicting approximately 50% of adults over 65 and over 80% of adults over the age of 70 (Cruickshanks et al. 1998). At the same time, typical audiometric tests do not explain the full range of difficulties that adults with hearing loss report in speech listening. Even when adults can successfully perceive speech, the additional cognitive effort required to extract meaning from degraded sensory input has a substantially negative impact on comprehension and subsequent memory for speech (Wingfield et al. 2005; Peelle 2018; Pichora-Fuller et al. 2016). This increased listening effort is an oft-cited hidden effect of hearing loss and is crucially important in understanding the challenges listeners face in high-level speech understanding.

The visual presentation of captioned speech is a promising route that may reduce the cognitive burden of auditory perceptual decoding in the face of age-related hearing loss. For example, text captioning has been shown to improve perceived speech intelligibility, word perception, and the accuracy of word recognition in sentences (Gordon-Salant & Callahan 2009; Krull & Humes 2016). In the current study, we report the results of two experiments examining the effects of realistic text captioning on multiple measures of auditory sentence memory under listening conditions intended to impose different degrees of listening effort. In Experiment 1, we introduce a novel cumulative moving window paradigm to study the effects of realistic text captioning on immediate speech recall and delayed recognition memory in young adults with normal hearing. In Experiment 2, we extend this paradigm to a sample of healthy older adults with a wide range of hearing acuity.

Speech Memory, Listening Effort, and Aging

Everyday listening commonly occurs in environments that present challenges to speech perception. Even when listeners can successfully identify perceptually degraded speech, the additional effort allocated to decoding the noisy signal interferes with downstream cognitive processes such as successful memory encoding (Rabbitt 1968; Rabbitt 1991; Wingfield et al. 2005). This general finding has been observed across a large number of experiments, dating back to the 1960s. Rabbitt (1968) first presented evidence that sensory processing draws on higher-order and domain-general cognitive resources. He presented young normal-hearing adults with lists of spoken digits. The first block of each list contained unprocessed speech, while the second block was presented in background noise. Critically, memory for early list items—which themselves were not degraded—was poorer when items in the later part of the

[1] Department of Psychology, University of Utah; [2] Interdepartmental Program in Neuroscience, University of Utah; [3] Department of Psychological Sciences, Whittier College; [4] Deparment of Communication Sciences and Disorders, University of Utah; and [5] Department of Educational Psychology, University of Utah.

0196/0202/XXX/XXXX-00/0 • Ear & Hearing • Copyright © 2021 Wolters Kluwer Health, Inc. All rights reserved • Printed in the U.S.A.

<zdoi; 10.1097/AUD.0000000000001079>

list were degraded, suggesting that understanding speech in noise interfered with cognitive processes required for memory encoding. Similar effortfulness effects have been replicated in word lists (Cousins et al. 2014; Piquado et al. 2010) and in running memory for speech (McCoy et al. 2005). Indeed, the consequences of effortful listening extend beyond memory impairments, impacting high-level sentence and discourse comprehension as well (Rabbitt 1991; Surprenant 1999; Pichora-Fuller 2003; McCoy et al. 2005; Cousins et al. 2014).

For example, sentence comprehension interacts with perceptual demands in hearing-impaired adults, particularly when the sentences are more linguistically complex. Wingfield et al. (2006) presented younger adults with normal hearing and older adults with and without hearing impairment sentences that varied in syntactic complexity (i.e., subject- versus object-relative clauses), and perceptual intelligibility (via time-compressed speech). Older adults with hearing impairment showed a differential disruption in comprehension accuracy for the more demanding object-relative sentences when they were presented at fast speech rates, despite the fact that all participants could still accurately perceive the speeded speech. Similarly, Stine-Morrow and colleagues (Gao et al. 2011; Stine et al. 1986; Stine & Hindman 1994; Stine & Wingfield 1990; Stine-Morrow et al. 2008) have demonstrated across multiple experiments that older adults show differentially worse memory for propositionally-dense speech and text compared with younger adults, with effects that are exaggerated under increased perceptual demand. Related findings have been reported in auditory discourse comprehension, where adults with hearing loss show reduced memory for detailed semantic features in short narrative discourses, despite having high intelligibility (Piquado et al. 2012). These studies present compelling evidence that acoustic challenge interacts with linguistic cognitive demand, pointing toward the involvement of domain-general cognitive resources in understanding degraded speech.

Assistive Text Captioning

An oft-cited finding in the speech processing literature is that speech reception thresholds and speech comprehension in background noise are improved when adults are able to view the articulatory expressions of the speaker (e.g., audiovisual speech or "speechreading"). These findings suggest that integrating cues from visual and auditory sensory channels can provide a benefit to intelligibility when speech is degraded, with an effect equivalent to an improvement in the speech-to-noise ratio of as much as 15 dB (Sumby & Pollack 1954; Sommers & Phelps 2016; Grant 2002). This is critical, as studies have shown that a 1-dB improvement in signal-to-noise ratio (SNR) can correspond to upward of a 10% increase in listener intelligibility performance (Grant & Braida 1991). Are there more robust visual cues that could be used to supplement speech, besides the speaker's facial movements? Orthographic cues from simultaneously presented text, for example, would provide a secondary direct channel for word recognition.

To date, however, there is limited literature directly examining the simultaneous processing of text and speech. Functional neuroimaging studies show evidence for heteromodal activation of auditory cortex to presentation of speech sounds and of corresponding letters representing those same speech sounds in literate adults, suggesting that the processing of shared orthographic and acoustic representations relies on similar underlying neural mechanisms (van Atteveldt et al. 2004). Moreover, early studies have shown evidence that speech perception can be biased by presenting simultaneous text cues. Frost and Katz (1989) presented individual bi-syllabic words in noise with accompanying visual word presentation. Critically, the visually presented words either matched or did not match the speech, and speech sounds were either presented in background noise, or were noise bursts only (no speech was presented). Using signal detection methods, Frost and Katz found a strong bias effect, such that visual input appeared to make amplitude-modulated noise sound more like speech, but it did not improve the detectability of the speech. At the same time, however, reaction times to correct detections were reliably shorter when text and speech matched, suggesting some evidence of a benefit to perception in audiovisual word recognition. Later work by Grant and Seitz (2000) showed that presenting matched orthographic text with speech in noise improved auditory word detection and provided masking release in sentences. More recently, studies have shown that text cues can be used to recalibrate the perception of ambiguous phonetic information for individual speech sounds. Keetels et al. (2016), for example, showed that when participants were presented with ambiguous speech sounds halfway between /b/ and /d/ that were combined with the presentation of the letter *b* or *d*, participants were more likely to categorize the speech sounds in accordance with the visually presented letter. These results suggest that listeners can adjust their phonetic boundaries during speech perception in accordance with the disambiguating orthographic information.

Beyond word recognition, a small number of studies have begun exploring the added benefits of text captioning to the perceived clarity of degraded speech. Gordon-Salant &Callahan (2009) have shown that real-time closed captioning of speech in television improves intelligibility in adults with hearing impairment. Additionally, this study compared the benefits of hearing aids to speech recognition with and without captioning. Strikingly, although identification of words in sentences was improved with the hearing aid, the effects of captioning were much larger. In fact, hearing aids provided no appreciable benefit to word recognition when text captioning was available.

Similar results have been found in studies examining the benefits of text captioning for improving word recognition in varying levels of background auditory noise (Zekveld et al. 2008; Krull & Humes 2016; Wild et al. 2012). Some studies, however, have provided less robust evidence for benefits of text captioning to perceived intelligibility (Sohoglu et al. 2014; Stine & Wingfield 1990). For example, Sohoglu et al. (2014) examined the effects of written text on listeners' ratings of clarity of noise-vocoded speech in young normal-hearing adults. They showed that perceived clarity was highest when text was presented before rather than after speech onset. In their study, intelligibility benefits of text were lost after a short stimulus-onset asynchrony (i.e., if text was presented even 200 ms after speech onset). However, it is important to note that the spectral vocoding used in their study resulted in a massive reduction in intelligibility, making the task much more difficult than normal listening situations.

As reviewed above, a number of previous studies have shown that text cues may modulate the perception of degraded or ambiguous speech. Nevertheless, it is less clear what downstream benefits this improved clarity has on adults' subsequent speech memory. Two studies to date are most relevant to addressing this question. First, Grossman and Rajan (2017) showed that the simultaneous presentation of congruent (but not incongruent) text benefitted immediate keyword recall of noise-degraded speech in young adults with normal hearing. Second, Krull and Humes (2016) tested whether the presentation of partially accurate visual text from an automatic speech recognizer could be used to successfully supplement word recognition in noise among young normal-hearing adults and older adults with normal hearing or moderate hearing loss. They found that combining degraded speech with partially correct text increased the percent of speech keywords accurately repeated by both young and older adults, relative to a condition with either auditory-only or text-only performance. To date, however, no studies have systematically examined whether these text captioning benefits extend to long-term memory. Therefore, it is unclear if the caption benefit is fleeting or extends into long-term speech memory. Finally, there is some evidence to suggest that real-time text captioning may provide benefits in real-world scenarios, such as for improving informed consent; Spehar et al (2016) showed that patients retained more information during simulated informed consent when consent was done with real-time assistive text captioning.

As seen in the above review, the extant literature provides some evidence supporting the idea that text captioning may offset the demands of perceptual processing and reduce the effects of effortful listening. Importantly, the text presentation methods used in most prior studies of speech memory have not been ecologically valid and have not approximated real-world text captioning systems as currently implemented in most technology (e.g., captioned television or assistive hearing telephones). In Krull and Humes (2016), for example, text was presented word-by-word via rapid serial visual presentation, thus precluding the natural ability to control reading rate and re-read, which is critical for effective comprehension (see e.g., Schotter & Payne, 2019; Payne & Silcox 2019). Grossman and Rajan (2017) in contrast, used a whole-sentence presentation in which the entire sentence appeared on the screen at the onset of the speech. This method allows readers to preview text prior to its onset in speech, which is not possible in real-world assistive text captioning, and potentially process the text and speech asynchronously. Finally, prior studies have largely used short, simple, and non-naturalistic speech stimuli. Given that the effects of hearing loss and perceptual demand are more apparent when the cognitive demand of the language is more taxing (see Wingfield & Stine-Morrow 2000; Payne & Silcox 2019 for reviews), it is important to examine the assistive effects of text captioning for more demanding, naturalistic speech.

THE CURRENT STUDY

We conducted two experiments aimed at directly testing the effects of realistic assistive text captioning on immediate recall and delayed recognition memory and confidence for real-world, propositionally-dense speech. In Experiment 1, we examined the effects of background noise and text captioning on memory for speech in young normal-hearing listeners. Consistent with prior work, we predicted an effortfulness effect: i.e., that increasing background noise, even modestly, would result in poorer downstream speech memory. However, we predicted that these effortfulness effects would be attenuated when speech was presented with text captioning. In Experiment 2, we replicated Experiment 1 in a cohort of older adults with a wide range of hearing acuity. Here, we expected that the effects of background noise would be larger overall and exaggerated among older adults with poorer hearing acuity. However, if older adults can take advantage of simultaneously presented text, we would expect text captioning to buffer against both the effects of poorer hearing and the increased effects of background noise on speech memory, resulting in a larger caption benefit in cases where hearing thresholds were elevated and when SNR was lower.

EXPERIMENT 1

The aim of the first experiment was to examine the impact of real-world assistive text captioning on immediate and delayed memory for informationally dense speech as a function of background noise. Toward this goal, we presented real-world, propositionally dense sentences (adapted from Stine-Morrow et al. 2008) in background noise at varying levels to manipulate perceptual challenge. Speech was either presented alone (speech-only) or with cumulative text captioning (text-captioned speech). This was accomplished via a novel multi-word cumulative text presentation paradigm, described below in more detail.

METHOD

Participants

Forty-eight young adults gave written informed consent. Participants' mean age was 22.89 years old (range: 18 to 34). The sample was 81% Caucasian, 12% (*N* = 6) Asian, and 2% (*N* = 1) African American. All participants completed high school; 78% had at least some college experience, and 4% had completed some post-graduate education. Participants reported normal or corrected-to-normal vision and reported no hearing loss, use of a hearing aid, or any history of ear or hearing-related disorders. From this sample, two participants were excluded from further analysis because they did not complete the entire protocol. All procedures were approved by the University of Utah Institutional Review Board.

Stimulus Materials and Design

Participants were presented with a total of 90 propositionally dense sentences drawn from prior work on sentence memory and aging (Stine-Morrow et al. 2001; Payne & Stine-Morrow 2017). Stimulus sentences were each 18 words long and dealt with diverse topics in science, nature, and history, with many adapted from popular magazines like *National Geographic*. Table 1 presents example sentences. Sentences were recorded by a single female native English speaker in a quiet room via a cardioid microphone. Audio was recorded and digitized at a sampling rate of 44.1 kHz. Individual sentences were excised and then trimmed to eliminate silent periods preceding and following each sentence. Sound files were then scaled to the same root-mean-square amplitude. The noise masker was a stationary speech-shaped noise with the long-term frequency spectrum of the speech.

Sentences were presented in two blocks of 45 sentences each, one without text captioning (auditory-only) and one with text captioning, with block order counterbalanced across subjects. Sentences were presented in quiet or at one of two different SNRs, +7 and +3 dB. These levels were chosen based on pilot testing to increase reported listening effort while maintaining high levels of subjective intelligibility. Despite the stimuli being rated as highly intelligible, we nevertheless expected the

TABLE 1. Example stimulus sentences

1. The Central Georgia Railroad used to be the most complete and elegant railroad complex in the United States.
2. Every morning housewives in Bali put some rice on small pieces of banana leaves to ward off spirits.
3. The innermost later of fur on a Husky which is as soft as good down keeps it warm.
4. In many species it is the females who shape evolution through their subtle exercise of choice in mating.
5. Most Turkish peddlers pool their meager funds to travel to big cities in tired old buses.
6. The white-backed night heron hides by day in reed beds and does not venture out until after twilight.
7. As a boy Norman Rockwell drew pictures of sailing ships copying them from packs of American Fleet cigarettes.

increased effort required for perceptual decoding in background noise to increase listening effort, yielding subsequent deficits in immediate recall and delayed recognition memory. The noise manipulation was randomized across items, resulting in a 2 (caption versus no caption) × 3 (no noise, +7 dB SNR, +3 dB SNR) within-subjects design.

Procedure and Apparatus

Participants were tested in a quiet sound-attenuated testing room. Speech stimuli were presented through the sound card of the stimulus presentation computer and routed to a MAICO MA-41 audiometer via the auxiliary channel. The audiometer routed the speech to the participant via RadioEar IP-30 insert air conduction earphones. Stimuli were presented at a constant level of 60 dB HL for all participants via audiometer presentation (i.e., presentation level was not based on individual thresholds). All hardware was calibrated to standard in our testing room by a National Association of Special Equipment Distributors certified technician. Stimuli were presented diotically.

For the text-captioned speech conditions, we developed a novel multi-word cumulative text segmentation technique to present captions. This method, a modification of the moving-window paradigm commonly used in reading research (see Haberlnadt 1994), was used because it closely approximates the real-time captioning of automated speech recognition systems like those used in captioned telephone conversations. To maintain control over the timing of stimulus delivery, we fixed the offset between speech and text and maintained a fixed number of words per presentation interval. Text was presented cumulatively on the screen in fixed-interval 3-word "chunks" with the previous text segments remaining on the screen until the end of the sentence. The presentation rate was yoked to the speech stimuli on a single-trial basis, with the constraint that the text could not precede the auditory signal, similar to real-word text captioning (i.e., the intermodal asynchrony with respect to auditory input could not be negative). We implemented this by coordinating the visual presentation onset of each 3-word text segment with the auditory offset of the third word of that segment. This was accomplished by creating a video file of each audiovisual stimulus and editing the text onsets and offsets manually via Adobe Premiere Pro video editing software. Therefore, compared with real-world automatic speech recognition (ASR) speech-to-text captioning systems, the intermodal asynchrony between the text and speech was relatively low and constant throughout the text-speech presentation, approximating an "ideal" text captioning system. Note further that text captions contained no captioning errors, which are not uncommon in ASR systems. This was done to examine the text captioning benefit under the most optimal conditions, as a benchmark for comparisons with future work in which we would systematically vary intermodal asynchrony and error type and rate.

On each trial, the audio or audiovisual sentence was first presented. To ensure that participants read and listened simultaneously for the audiovisual trials, the whole sentence disappeared from the screen 1000 ms after the offset of the final word of the sentence. Following the sentence offset, a fixation cross ('+') appeared on the screen for a 5000-ms retention interval. After 5000 ms, a visual recall visual cue ('???') was presented, at which point participants were instructed to say aloud as much of the sentence as they could remember. After participants finished recalling the sentence, they pressed a button on a response box to initiate the next trial. Recalled sentences were recorded for later transcription and scoring. Recall was scored for each sentence as the proportion of individual words correctly recalled (e.g., Potter & Lombardi 1998; Gilchrist et al. 2008), which for single sentences correlates very strongly with gist-based propositional scoring (typically above $r = 0.9$; Kintsch & van Dijk 1978; Stine-Morrow et al. 2008; Payne & Stine-Morrow 2017)*.

After the end of each block of 45 sentences, a delayed recognition memory task was administered (see Payne et al. 2016; Payne & Federmeier 2017). In each memory task, participants were presented with a set of 42 sentences. For each sentence, participants were asked to indicate whether or not they heard the sentence in the study (recognition accuracy), and then rate their confidence in their response (retrospective confidence) on a Likert scale from 1 to 5, where 1 indicated no confidence and 5 indicated complete confidence (e.g., Busey et al. 2000). In each task, half of the items were previously viewed/heard and half of the items were foil sentences that shared some features with previously presented sentences but were not identical. For example, if the (previously heard) target sentence was: "*Swordfish and marlins have muscles behind their eyes which adjust the temperatures of their brains in colder waters,"* the foil sentence was: "*Swordfish heat up their eyes to above ambient ocean temperature in order to improve their tracking of prey.*" Foil and target sentences were matched for sentence length and propositional density. Both foil and target sentences were true and did not contain any linguistic or world-knowledge violations. Koeritzer et al (2018) recently showed that the inclusion of foils reduces ceiling effects, resulting in better sensitivity in recognition memory for spoken sentences. Moreover, the inclusion of foils that are yoked to specific items allows for the measurement of accuracy rates across stimulus conditions (i.e., background noise; see Koeritzer et al. 2018 for more detail).

Statistical Analysis

(Generalized) linear mixed-effects models ([G]LMMs) were fit to the sentence recall and recognition memory data. Completely crossed random intercepts were defined for

*For the first 18 participants in Experiment 1, recall for one item was lost due to a programming error, leaving 89 intact recall samples for these participants.

participants and items, and a maximal random-effects structure for all within-subject effects was included in the models (Barr 2013; Barr et al. 2013; Bates et al. 2015; Matuschek et al. 2017) with a variance-components structure for the random slopes (excluding covariance terms between the random slopes to improve convergence; see Barr 2013). Statistical inference on the fixed effects was conducted via separate likelihood ratio tests for each fixed-effects parameter. For follow-up tests decomposing higher-order interactions, unstandardized contrast estimates (*est.*) are presented with 95% confidence intervals (CIs). Approximate degrees of freedom are calculated via a Satterthwaite approximation. For recognition memory accuracy, a binary outcome at the single-trial level, a logit link function was used and the model was fit based on maximum likelihood estimation through the Laplace approximation. Follow-up tests for the logit models present estimates on a log odds scale.

RESULTS AND DISCUSSION

Immediate Speech Recall

Figure 1 (top left panel) plots correct word recall proportion as a function of text captioning and noise. Error bars represent within-subject standard errors. A 2 (Caption: Caption versus No Caption) by 3 (Noise: quiet, +7 dB SNR, +3 dB SNR) linear mixed-effects model was fit to the recall data. This model revealed statistically significant main effects of Caption [χ^2 (1) = 22.08, $p < 0.0001$] and Noise [χ^2 (1) = 14.89, $p < 0.001$]. Importantly, we observed a significant Caption × Noise interaction [χ^2 (2) = 9.40, $p < 0.01$] as well. Follow-up contrasts revealed that, in the no-captioning condition, recall was worse for sentences presented at +3 dB SNR compared with both the +7 dB SNR [*est.* = −0.03, 95% CI (−0.05 to −0.01)] and quiet conditions [*est.* = −0.04, 95% CI (−0.06 to −0.02)]. There was no reliable difference in recall between the +7 dB and quiet conditions. In contrast, in the text-captioning condition, there was no significant difference between Noise conditions. These findings thus replicated the listening effort effect—that speech recall is hindered by increasing background noise. More importantly, the simultaneous delivery of assistive text captioning offsets the negative effects of signal degradation, resulting in improved speech recall.

Delayed Recognition Memory Accuracy • Figure 1 (top middle panel) plots sentence recognition accuracy as a function of text captioning and noise. Error bars represent within-subject standard errors. A 2 (Caption: Caption versus No Caption) by 3 (Noise: quiet, +7 dB SNR, +3 dB SNR) logit mixed-effects model was fit to the accuracy data. This model revealed a statistically significant main effect of Caption [χ^2 (1) = 20.28, $p < 0.001$] such that recognition memory was higher in the presence of text captions compared with trials without text captions. In addition, we observed a significant Caption × Noise interaction [χ^2 (1) = 16.68, $p < 0.001$]. Follow-up tests revealed that, in the no-captioning condition (listening only), recognition accuracy was worse for sentences presented at +3 dB SNR compared with both the +7 dB SNR [*est.* = −1.38, 95% CI (−2.18 to −0.58)] and quiet conditions [*est.* = −0.91, 95% CI (−1.52 to −0.30)]. There was no reliable difference in recognition memory between the +7 dB and quiet conditions. These findings indicate that increased noise reduced recognition memory accuracy when listening to speech alone. In the text captioning condition, there was no significant difference in recognition memory as a function of listening condition. Overall, these findings indicate that text captioning improves delayed sentence recognition memory and offsets the negative effects of background noise on speech memory, even in young normal-hearing adults.

Recognition Memory Confidence

Figure 1 (top right panel) plots sentence recognition confidence ratings as a function of text captioning and noise. Error bars represent within-subject standard errors. A 2 (Caption: Caption versus No Caption) by 3 (Noise: quiet, +7 dB SNR, +3 dB SNR) linear mixed-effects model was fit to the ratings data. This model revealed a statistically significant main effect of Caption [χ^2 (1) = 13.43, $p < 0.001$], such that responses to test items that were previously paired with text captions were more confidently rated than items presented without text captions. In addition, we observed a significant Caption × Noise interaction [χ^2 (2) = 6.85, $p < 0.05$], suggesting that text captioning moderated the effects of noise on recognition confidence. In the no-caption condition, subsequent recognition confidence was lower for trials presented at a +3 dB SNR compared with items presented in quiet [*est.* = −0.10, 95% CI (−0.19 to −0.002)]. However, with text captions, there were no effects of background noise on subsequent recognition confidence. Thus, consistent with the recognition accuracy data, text captioning appeared to improve subsequent recognition confidence and offset the negative effects of background noise.

Collectively, the findings from the first experiment are clear. First, effects of background noise were apparent in immediate recall, delayed recognition memory, and retrospective recognition confidence. Replicating past work in young normal-hearing listeners (see Peelle, 2018 for a recent review), even subtle increases in acoustic perceptual challenge during speech processing can result in negative downstream consequences for subsequent speech memory. More importantly, we showed that the presentation of simultaneous assistive text captions offset the negative effects of background noise on all three speech memory measures in the current study. Thus, even in young adults with normal hearing, who would not be expected to be very susceptible to the negative effects of background noise, we still observe a benefit of assistive text captioning. In Experiment 2, we examined whether text captions yielded beneficial effects on speech memory in older adults with and without hearing loss.

EXPERIMENT 2

Method

Participants • Thirty-one older adults gave written informed consent. Participants' mean age was 71 years old (range: 61 to 80). The sample was 100% Caucasian. Eighty-seven percent of participants reported a high school education and at least some college experiences, and 48% had completed at least some postgraduate education. One participant reported using a hearing aid (though not at the time of the study), and 23% self-reported some degree of hearing loss or hearing problem. One participant reported deafness in the left ear and thus was only tested in the right ear. All procedures were approved by the University of Utah Institutional Review Board.

Hearing Assessment and Vision Screening • Two hearing tests were conducted. The stimuli were presented through MA-41 audiometer via RadioEar IP-30 insert air conduction earphones. First, pure-tone thresholds were measured using the modified Hughson-Westlake at octave frequencies from 250 to 8000 Hz

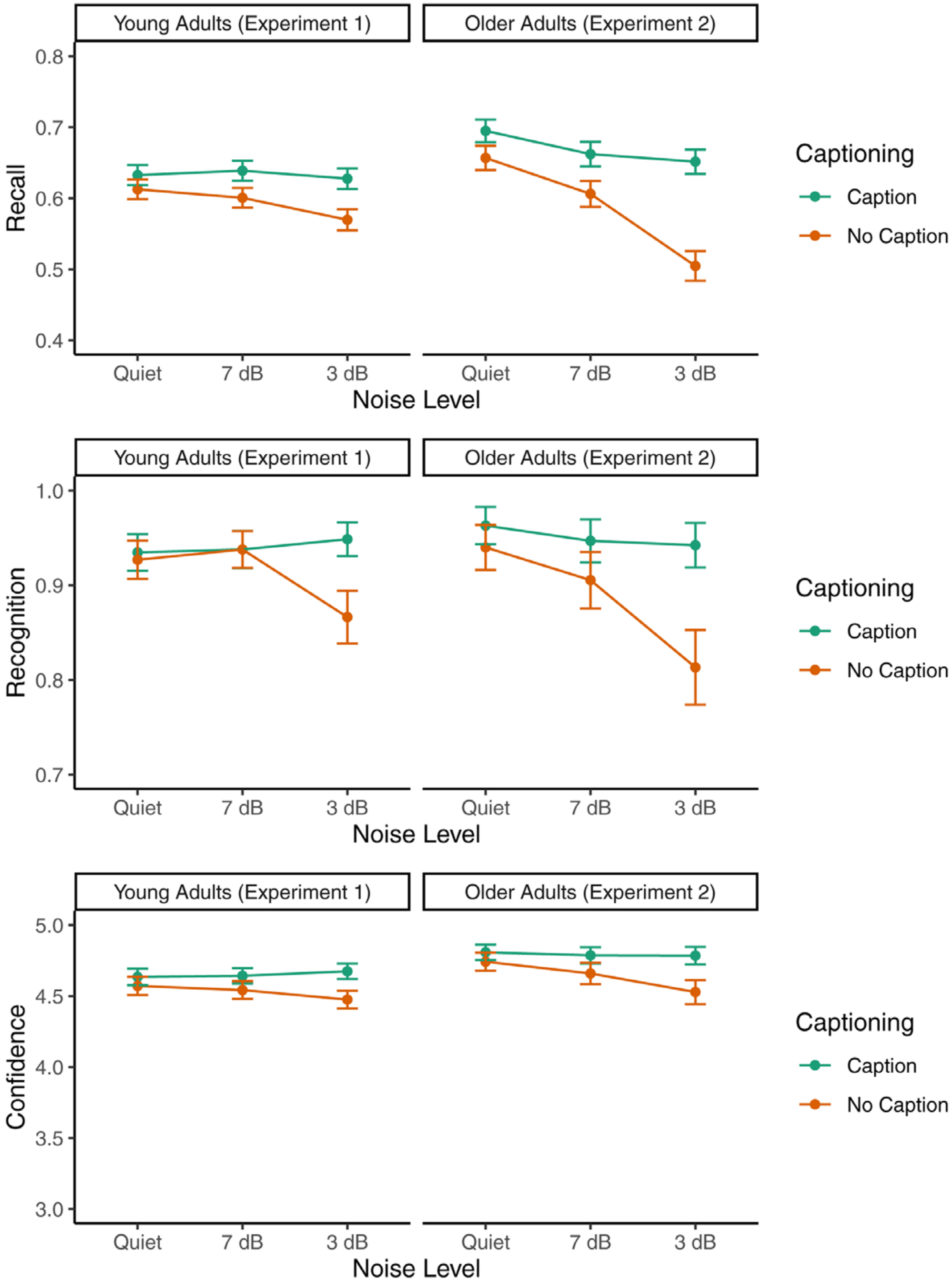

Fig. 1. Sentence memory as a function of background noise and text captioning in Experiments 1 (left panels) and 2 (right panels). Top panel: immediate recall accuracy; middle panels: delayed recognition memory; bottom panel: recognition memory confidence. Standard errors reflect within-subject 95% confidence intervals.

for each ear. Interoctaves 750, 1500, and 3000 Hz were included if a 20-dB or greater difference was found between the respective octave frequencies. In addition, we tested speech recognition thresholds (SRTs) using a recorded Central Institute for the Deaf (CID) W-1 spondee word list (Auditec, St. Louis, MO). Spondees were presented beginning at 30 dB HL and participants were asked to repeat the words they heard. The modified Hughson-Westlake method was then used to identify the participant's SRT in each ear. Near visual acuity was also tested for both the right and left eye using the Rosenbaum visual acuity test.

Cognitive Assessment • Participants completed a brief battery of cognitive assessments. Participants completed the Mini Mental State Exam (Folstein et al. 1975), a brief 30- item cognitive impairment assessment often used in clinical settings to assess general cognitive function and risk for Alzheimer's dementia. A score below 24 often indicates some risk for clinically relevant cognitive impairment. Participants also completed the Montreal Cognitive Assessment (MoCA; Nasreddine et al. 2005), which has been used as a clinical instrument to assess the presence of mild cognitive impairment (MCI). Moreover, the MoCA and has been used to predict amnestic MCI-related language comprehension difficulties (Payne & Stine-Morrow 2016). Although the appropriate cut-off score for cognitive impairment in the MoCA is variable across samples (e.g., Waldron-Perrine & Axelrod 2012), individuals scoring below 20 are generally considered to be at increased risk for MCI.

Participants also completed the phonemic fluency task as a measure of verbal fluency (Benton & Hamsher 1978). In this task, participants were given a letter cue (phonemic: e.g., "F") and asked to produce as many words that they could think of that begin with that letter in 60 s. A total score is calculated as the sum of accurate unique words correctly produced across the trials.

Verbal working memory capacity was measured via the short-form computerized version of the reading span task, as in Oswald et al. (2015). Participants were presented with a set of sentences of approximately 10 to 15 words in length and were asked to judge whether or not each sentence was sensible (approximately half were sensible). After each sentence, participants were presented with a letter for recall at the end of the set. Sets of sentences and letters ranged in size from 4 to 6, with three administrations for each set size. Participants' absolute scores are the number of trials in which the participant recalled all elements in the correct order without error. Higher scores reflect larger working memory capacity.

Finally, verbal ability was assessed via the extended range vocabulary tests from the ETS Kit of Factor Referenced Cognitive Tests (Ekstrom 1976). The test is timed such that participants must finish within 6 minutes. For each item, participants are asked to choose a correct synonym of a target word from a list of five possible words. The total score is the number of correct answers minus a fraction of incorrect answers. Higher scores reflect greater verbal ability.

Apparatus, Design, and Stimulus Materials • The apparatus, design, and stimulus materials were identical to Experiment 1, with the exception that stimuli were presented monaurally at 65 dB HL to the participant's better-hearing ear (based on average pure-tone thresholds at 1 to 4 kHz).

Audibility Control Task

An audibility control task was conducted to ensure that the speech-in-noise conditions were sufficiently audible to the participants at the target SNRs. Participants heard a total of six test sentences (5 to 7 words long, e.g., "Mary chose not to join the army"), three at +3 dB SNR and three at +7 dB SNR, all at the presentation level (65 dB HL) used in the primary experiment. The participants were asked to "shadow" the speech by repeating out loud each word as it was heard, therefore minimizing any memory component to the task. Word shadowing accuracy was 97%. Therefore, the memory analyses presented below cannot be explained simply by a lack of audibility, but rather by the increased cognitive effort for perceptual decoding as a function of SNR and hearing level.

Procedure • Participants first completed the cognitive test battery. They then completed hearing assessments and a vision screening. Participants then completed the audibility control task, followed by the primary experiment. The experimental paradigm and procedure were identical to Experiment 1.

RESULTS

Hearing Testing

All participants successfully completed both hearing evaluations. Figure 2 shows the average audiogram for the left and right ears from 250 to 8000 Hz. The audiogram shows the characteristic sloping configuration associated with typical age-related hearing loss. Using a better-ear PTA (1, 2, 4 kHz) of <25 dB HL as a criterion for normal hearing (Hall & Mueller 1997), $N = 12$ of the participants had clinically relevant hearing loss. Analyses reported below use a continuous measure of PTA. Average speech recognition thresholds were 25 dB HL

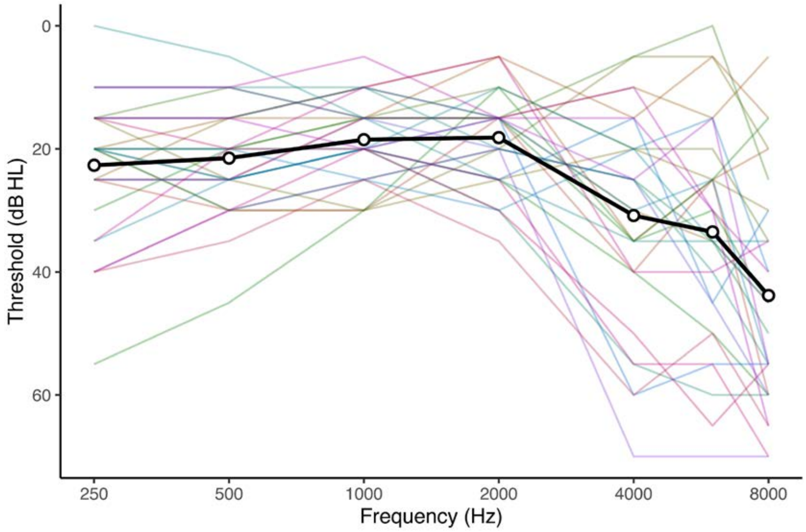

Fig. 2. Pure tone audiogram for better hearing ear for the older adult participants in Experiment 2.

[standard deviation (SD) = 7.54] 24.84 dB HL (SD = 8.21) for the left and right ears, respectively. SRTs were correlated with average 1 to 4 kHz PTAs (right ear: $r = 0.76$, left ear: $r = 0.57$).

Cognitive Test Battery

Table 2 provides means and standard deviations for the measures in the cognitive test battery. All participants in the current study scored in the normal range on the Mini Mental State Exam (range 25–30), and the MoCA (range 24–30), indicating that none of the participants showed clinically relevant cognitive impairment. Fluency and verbal ability scores are comparable to the performance of age and education matched samples (Tombaugh et al. 1999). Although normative data are not available for the verbal working memory assessment, Oswald et al. (2015) reported an average score of 23.25 ($SD = 4.78$) on the same computerized reading span assessment among $N = 172$ younger adults (mean age = 19.4). The older adult sample in the current study performed below the younger adults in Oswald et al. (2015) ($M = 17.45$, $t(201) = 5.87$, $p < 0.0001$), consistent with the well-reported age-related decline in verbal working memory (e.g., Bopp & Verhaeghen 2005).

Speech Recall

Figure 1 (bottom left panel) plots recall percentage as a function of text captioning and noise. A 2 (Caption: Caption versus No Caption) by 3 (Noise: quiet, +7 dB SNR, +3 dB SNR) linear mixed-effects model was fit to the recall data. This model revealed a statistically significant main effect of Caption [χ^2 (1) = 23.39, $p < 0.001$*]* such that sentence recall was higher among trials with text captions compared with trials without text captions. In addition, there was a main effect of noise [χ^2 (1) = 41.63, $p < 0.001$*]*, such that recall was lower for +3 dB than +7 dB [*est.* = −0.06, 95% CI (−0.07 to −0.04)] and quiet [*est.* = −0.10, 95% CI (−0.12 to −0.08)]. Moreover, sentence recall was worse for +7 dB than quiet [*est.* = −0.04, 95% CI (−0.06 to −0.02)].

Finally, we observed a significant Caption x Noise interaction [χ^2 (2) = 36.99, $p < 0.001$], indicating that text captioning moderated the effects of background noise on recall. Follow-up tests revealed that, in the no-captioning condition (listening only), recall was worse for sentences presented at +3 dB SNR compared with both the +7 dB [*est.* = −0.10, 95% CI (−0.12 to −0.08)] and quiet conditions [*est.* = −0.15, 95% CI (−0.18 to −0.12)]. Moreover, recall was worse at an SNR of +7 dB than in quiet [*est.* = −0.05, 95% CI (−0.08 to −0.02)]. In contrast, in the text captioning condition, effects of noise were reduced in magnitude, as can be seen in Figure 2. With text captions, there was no significant difference in recall between the +3 and +7 dB conditions [*est.* = −0.01, 95% CI (−0.03 to 0.01)]. Although recall was modestly worse in the +3 dB [*est.* = −0.04, 95% CI (−0.07 to −0.02)] and +7 dB [*est.* = −0.04, 95% CI (−0.06 to −0.01)] SNR conditions relative to quiet, effect sizes were substantially reduced in magnitude relative to the trials without text captioning. The effect size of the captioning benefit (difference between text captioned and non-text captioned block) was 0.15 [95% CI (0.11 to 0.18)] in the +3 dB noise condition, 0.05 [95% CI (0.02 to 0.09)] in the +7 dB condition, and .04 [95% CI (−0.001 to 0.08)] in the quiet condition. Overall then, these findings indicate that there were reliable improvements in immediate recall with text captioning among older adults with hearing impairment, and that the presence of text captioning helped to buffer against (but not completely eliminate) the negative impacts of increased background noise on speech memory.

TABLE 2. Descriptive statistics of cognitive test battery performance among participants in Exp 2

Measure	Mean	SD	Min	Max
1. MMSE	28.81	1.67	25	30
2. MoCA	27.68	1.72	24	30
3. Phonemic fluency	38.48	10.76	20	57
4. Reading span	17.45	6.44	3	26
5. Vocabulary	16.73	3.89	6	24

MMSE indicates Mini Mental State Exam; MoCA, Montreal Cognitive Assessment.

Recognition Memory

Figure 1 (bottom middle panel) plots delayed sentence recognition memory accuracy as a function of text captioning and noise. A 2 (Caption: Caption versus No Caption) by 3 (Noise: quiet, +7 dB SNR, +3 dB SNR) logit mixed-effects model was fit to the accuracy data. This model revealed a statistically significant main effect of Caption [χ^2 (1) = 28.45, $p < 0.0001$] such that, on average, the presence of adaptive text captioning improved sentence recognition accuracy compared with trials with speech only [*est.* = 0.81, 95% CI [0.44 to 1.17)]. In addition, we observed a significant Caption × Noise interaction [χ^2 (2) = 6.86, $p < 0.05$s, indicating that text captioning moderated the effects of noise on delayed recognition accuracy.

Follow-up tests revealed that, in the no-captioning condition (listening only), delayed recognition accuracy was worse for sentences presented at +3 dB SNR compared with both the +7 dB [*est.* = −1.21, 95% CI (−2.05 to −0.38)] and quiet conditions [*est.* = −1.67, 95% CI (−2.56 to −0.79)]. There was no reliable difference in recognition memory between the +7 dB and quiet conditions. These findings replicated those in Experiment 1 with younger adults in indicating that increased noise reduced recognition memory accuracy when listening to speech alone. In the text captioning condition, there were no significant differences in recognition memory as a function of acoustic clarity. As in Experiment 1 with younger adults, the overall findings indicated that text captioning improved delayed sentence recognition memory accuracy in older adults and, additionally, offset the negative effects of background noise on sentence recognition memory.

Recognition Memory Confidence

Figure 1 (bottom right panel) plots sentence recognition confidence ratings as a function of text captioning and noise. Error bars represent within-subject standard errors. A 2 (Caption: Caption versus No Caption) by 3 (Noise: quiet, +7 dB SNR, +3 dB SNR) linear mixed-effects model was fit to the ratings data. This model revealed a statistically significant main effect of Caption [χ^2 (1) = 27.29, $p < 0.0001$], such that responses to test items that were previously paired with text captions had higher recognition confidence ratings than items presented without text captions. In addition, we observed a significant Caption x Noise interaction [χ^2 (2) = 9.04, $p < 0.01$], suggesting that text captioning moderated the effects of noise on recognition confidence. In the no-caption condition, subsequent recognition confidence was lower for trials presented in noise at a +3 dB SNR compared with items presented at +7 dB SNR [*est.* = −0.13; 95% CI (−0.24 to −0.02)] and in quiet [*est.* = −.21, 95% CI (−0.32 to −0.11)]. However, with text captions, there were no effects of background noise on subsequent recognition confidence. Thus, consistent with the results from Experiment 1, text captioning appeared to improve subsequent recognition confidence and offset the negative effects of background noise in older adults.

Interactions With Hearing Level

Exploratory follow-up analyses were conducted to test whether individual differences in hearing level moderated the effects of text captioning and background noise on speech recall and recognition memory accuracy and confidence. For these analyses, hearing level was included as a covariate in the models fit above and treated as a moderator of the effects of text captioning, noise, and the Caption × Noise interaction (i.e., a 3-way interaction Hearing Level × Caption × Noise). This model is presented graphically in Figure 3 (left panel) and summarized in the top panel of Table 3. The Caption × Noise × Hearing Level interaction was significant, indicating that the text captioning attenuated the effects of hearing loss on speech recall in more acoustically challenging situations. As can be seen in Figure 3, in the no-text-captioning condition, there was a negative relationship between hearing threshold and speech recall, an effect that was stronger in conditions with higher background noise. This was confirmed by a significant Noise × Hearing Level interaction in the block with no text captions [χ^2 (2) = 10.00, $p < 0.01$]. However, in the text caption block, there were no reliable effects of noise [χ^2 (2) = 1.19, $p > 0.05$] or hearing level [χ^2 (1) = 0.19, $p > 0.05$] nor was there an interaction between noise and hearing level [χ^2 (2) = 1.71, $p > 0.05$]. Thus, text captioning appeared to lead to improvements in speech recall among older adults with lower hearing acuity, attenuating the effects of background noise and hearing level on recall.

Similar follow-up analyses were conducted for recognition memory accuracy and confidence. Results of these analyses are presented graphically in Figure 3 (middle and right panels for accuracy and confidence, respectively). Likelihood ratio tests are presented in the middle and bottom panels of Table 3. We observed a significant interaction between text captioning and hearing acuity on recognition accuracy. This effect was driven by a reliable negative effect of hearing level on recognition memory accuracy without text captioning [χ^2 (1) = 6.25, $p < 0.05$], whereas hearing level did not affect recognition accuracy in text-captioned speech [χ^2 (1) = 0.40, $p > 0.05$]. A similar pattern was observed for recognition confidence, where a significant Caption × Hearing Level interaction was observed, which was driven by a stronger relationship between hearing level and memory confidence in the block without text captions than in the block with text captions (see Fig. 3).

GENERAL DISCUSSION

Across two experiments, we examined the effects of realistic assistive text captioning on immediate and delayed memory for degraded speech. In our first experiment, we examined young adults with self-reported normal hearing, and in the second experiment, we examined older adults with a range of hearing acuity. To approximate real-world text captions, we developed a novel cumulative text-captioning method that presented captions simultaneously with speech. Importantly, under a listening effort model (Peelle 2018; Pichora-Fuller et al. 2016), even small increases in background noise should draw attentional resources away from memory encoding, even when speech is still intelligible (Piquado et al. 2012; McCoy et al. 2005). If text captions can be used as a compensatory tool to reduce the effects of effortful listening, then the negative effects of noise degradation on speech memory outcomes should be attenuated in text-captioned speech. Across both experiments, our findings were clear: the benefit provided by text captions improved not only immediate recall, but also long-term memory outcomes in both younger (Experiment 1) and older adults (Experiment 2) and attenuated both the effects of increased background noise (in both experiments) and hearing loss (in Experiment 2). In the

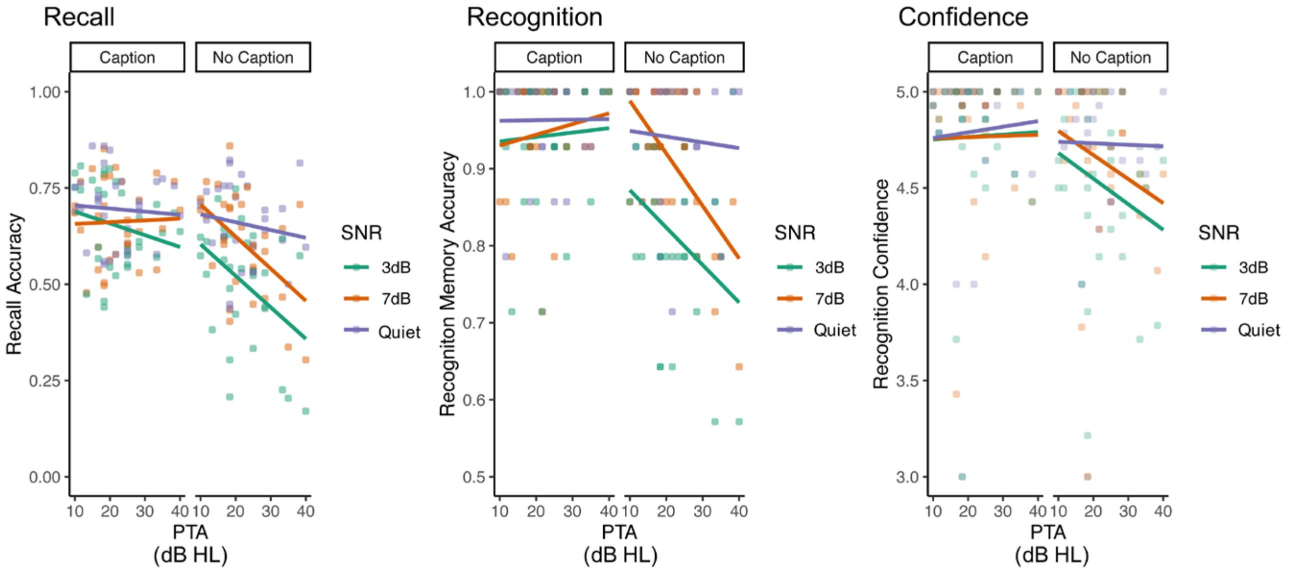

Fig. 3. Effects of hearing level (PTA), background noise, and text captioning on immediate recall accuracy (left panel), delayed recognition memory accuracy (middle panel), and delayed recognition confidence (right panel) in Experiment 2. SNR indicates signal-to-noise ratio; PTA, pure tone average.

TABLE 3. Likelihood ratio tests for hearing level, caption, and noise effects on immediate recall accuracy (top panel), delayed recognition memory (middle panel), and delayed recognition confidence (bottom panel)

Effect	*df*	χ^2	*p* value
Immediate recall accuracy			
Caption	1	10.82	<0.01
Noise	2	0.85	0.65
Hearing level	1	2.84	0.09
Caption × noise	2	3.37	0.19
Caption × hearing level	1	43.45	<0.0001
Noise × hearing level	2	15.34	<0.001
Caption × noise × hearing level	2	7.39	0.02
Recognition memory accuracy			
Caption	1	22.87	<0.0001
Noise	2	20.15	<0.0001
Hearing level	1	0.75	0.39
Caption × noise	2	6.7	0.04
Caption × hearing level	1	7.43	0.01
Noise × hearing level	2	0.5	0.78
Caption × noise × hearing level	2	3.43	0.18
Recognition memory confidence			
Caption	1	21.96	<0.0001
Noise	2	8.87	0.01
Hearing level	1	0.15	0.70
Caption × noise	2	5.6	0.06
Caption × hearing level	1	7.97	0.01
Noise × hearing level	2	5.04	0.08
Caption × noise × hearing level	2	1.82	0.40

following, we discuss specific findings in detail and discuss the implications of these findings for our understanding of listening effort and the potential real-world benefit of text captioned speech.

Listening Effort and Speech Memory

In both experiments, we observed that in the listening-only conditions, adding noise to speech negatively impacted immediate recall, delayed recognition memory accuracy, and recognition confidence relative to speech presented in quiet. Moreover, we found in Experiment 2 that hearing acuity moderated these effects, such that older adults with poorer hearing acuity were more negatively affected by the presence of background noise when listening to speech alone. These findings are consistent with a number of prior studies (Cousins et al. 2014; McCoy et al. 2005; Piquado et al. 2010; Rabbitt 1968; Rabbitt 1991; Wingfield et al. 2005), in showing that when listening is more perceptually effortful, it interferes with memory encoding. Some (e.g., Rabbitt 1968; Pichora-Fuller et al. 2016) have argued that listening effort increases demand on a limited reserve of cognitive resources. The listener's speech processing system prioritizes correct perception of the speech signal, and thus increasing effort leaves fewer resources available for higher-level processes such as memory encoding and retention. In Experiment 2, we presented participants with a short speech shadowing task using the same SNRs in the primary task and observed that participants performed near ceiling on this task, suggesting that they were able to correctly perceive the speech presented to them. Note, also, that the SNRs used were on par with those found in everyday listening scenarios (Wu et al. 2018). Thus, the effects found were present in normal SNRs experienced in everyday effortful listening. Collectively, our findings support effort-based theories that suggest that increased perceptual effort consumes resources that would otherwise be available for memory encoding and maintenance.

Importantly, this study extends prior research by showing robust effects of acoustic challenge across multiple memory outcomes in the same study. Although increases in listening effort have been reported for immediate recall (e.g., Cousins et al. 2014; McCoy et al. 2005), only very few studies have systematically examined how acoustic properties of the speech affect delayed recognition memory of spoken sentences. Van Engen et al. (2012) examined differences between clear and conversational speech on recognition memory. They observed improved recognition memory for clear speech relative to conversational speech presented in quiet, suggesting that the improvement in clarity with clear speech can benefit subsequent memory. Similarly, Gilbert et al. (2014) showed that subsequent recognition memory for clear speech in noise was improved relative to conversational speech in noise. Finally, Koeritzer et al. (2018) showed that sentence recognition memory was poorer for acoustically challenging sentences that contained lexically ambiguous words, suggesting that linguistic complexity can interact with perceptual difficulty in negatively influencing subsequent memory. Importantly, the current study is the first, to our knowledge, to simultaneously examine the effects of acoustic challenge on immediate recall and delayed recognition accuracy and confidence in the same study. Such a simultaneous investigation is important, as dissociations between recall and recognition and between recognition accuracy and confidence have shown that these measures, while highly overlapping, reflect partially dissociable memory systems (Balota & Neely 1980; Benjamin et al. 1998; Weidemann & Kahana 2016; Yonelinas & Jacoby 2012).

Text Captioning Benefit to Memory

Although we found that listening to degraded speech generally had a negative impact on memory in both older and younger listeners, simultaneously presenting text captions with speech dramatically reduced these negative effects for both groups. We found that when captions were presented, changes in background noise had little or no effect on a listener's speech memory. This is in line with previous work that has shown that text captions are beneficial to word intelligibility and immediate word recall (Zekveld et al. 2008; Krull & Humes 2016; Wild et al. 2012). However, our findings extend those found previously, in that we found broad benefits across multiple memory outcomes. Critically, the benefits of text captioning persist into delayed recognition memory outcomes, indicating that the benefits of text captioning extend beyond initial perception, impacting long-term verbal memory representations. Taken in line with prior research, our results suggest that text captions may not only benefit perceptual processes (e.g., improving intelligibility) but that the addition of the text input may reduce listening effort, leading to improved long-term retention.

It is also important to highlight that the caption benefit was broadly observed. Although younger listeners in Experiment 1 reported normal hearing acuity, we still observed a benefit from text captions across all memory performance measures. In our analysis of the effects of hearing acuity in older adults, we found that older listeners performed worse on all outcome measures as hearing acuity declined. But, when captions were present, hearing impairment effects were largely eliminated (see also

Gordon-Salant & Callahan 2009). Taken together, our results suggest that captions provide a broad benefit to literate adult listeners (even those without hearing loss), but that the benefit from captions is in fact greatest for older adults with hearing impairment. These findings parallel the literature on video captioning, which has shown that captioning improves comprehension and memory broadly across literate populations, from young children to older adults (see Gernsbacher 2015 for a review).

Another important dimension that the current study investigated was the use of more realistic text captioning. We used a novel multi-word cumulative text segmentation technique to simulate current real-time captioning systems with respect to the continuous and cumulative presentation of text. Prior work (e.g., Krull & Humes 2016; Wild et al. 2012) used single-word rapid serial visual presentation (RSVP) to present text, which allows the audio and visual information to be presented synchronously but is not used in real-world text captioning and would not allow for participants to read naturally, such as to re-read to process prior material (e.g., Schotter et al. 2014). In contrast, other studies (e.g., Zekveld et al. 2008; Grossman & Rajan 2017) presented whole sentences at the onset of the speech. This method allows for more naturalistic reading of the visual information, but the downside of this method is that it allows readers to read information potentially in advance of speech, which is unrealistic for real-world ASR applications, as well as the ability to process text and speech asynchronously. The advantage of our novel presentation method over those used previously is that it simulates real-world caption presentations, which are cumulative, such that text does not precede the speech and past text remains available.

One important caveat to the naturalistic presentation in the current study is that each 3-word text segment appeared on screen immediately after the corresponding speech segment had been heard. Thus, the captions appeared with lower intermodal asynchrony than what is presented in typical ASR systems. Real-world text captioning voice-to-text technology typically has a high degree of variability in the latency at which the captions are presented. Importantly however, as in the current study, the auditory input always leads the text, as speech-to-text transcription occurs in real-time. Several studies have found larger benefits to speech perception from text captions when the text precedes the speech rather than when it follows (Davis et al. 2005; Jones & Freyman 2012; Sohoglu et al. 2014). Likewise, when there is an asynchrony between visual facial cues and speech, it has been found that a successful audiovisual benefit occurs with a much larger asynchrony when the visual information leads the auditory information than vice versa (Grant & Greensberg 2001; Grant et al. 2004). Importantly, for text captioning benefits to have useful practical application, captioning benefits need to be observed under conditions where the auditory input precedes the text input, as this is a necessary precondition for real-time ASR systems. Indeed, in the current study, we found that text captions with the audio leading the text input yielded a benefit to immediate recall and delayed recognition memory. Given this, we encourage future work to systematically study the role of intermodal asynchrony on the benefits provided by text captions in the future.

It is important to note that, although our findings clearly showed evidence of captioning benefits across multiple speech memory outcomes, these experiments do not speak directly to the exact mechanism by which text captions yield improvements in speech memory. One possibility is that listeners are able to rapidly process and integrate text and speech information simultaneously; as information in the auditory channel becomes increasingly degraded, listeners benefit from the visual channel to aid in multi-modal word recognition (e.g., van Atteveldt et al. 2004; Frost and Katz 1989; Grant and Seitz 2000; Keetels et al. 2016). Notably, potential integration mechanisms in text-speech processing would likely vary considerably from those previously studied in the prior audiovisual integration literature on face-speech processing (e.g., Sommers & Phelps 2016) or the use of other coarse visual cues (e.g., Strand et al. 2020; Yuan et al. 2020). For example, facial cues provide information on place of articulation that can translate to fine phonetic detail, which can benefit lexical access (Erber 1969; Sumby & Pollack 1954; Grant & Walden 1996). In contrast, text provides a direct orthographic route to word recognition, leading to activation of lexical semantic features and corresponding phonological representations.

An alternative hypothesis is that listeners adopt a strategy whereby they attempt to suppress the speech input and allocate greater attention to the text captions, essentially down-weighting the auditory channel in favor of the visual channel. Note that there are a number of reasons why an extreme version of this hypothesis, in which listeners attempt to ignore the auditory channel is unlikely. First, while listeners can shift attention across multiple modalities, speech input cannot be completely inhibited (Conway et al. 2001; Strauß et al. 2014). Second, in the current study, although the effects of noise were attenuated with text captions, it is not the case that all negative effects of noise were completely eliminated (see Experiment 2 recall effects). Nevertheless, our current study cannot rule out a weaker version of this hypothesis, whereby listeners rely differentially on the visual channel over the auditory channel, potentially in a strategic manner as a function of increased noise-induced acoustic challenge or hearing loss.

One way to address these two accounts is to examine real-time cross-modal processing via temporally precise on-line methods, such as electroencephalography (EEG) or eye tracking, to directly measure how text and speech information is being utilized in real time, during the encoding of text and speech. For example, the reading literature has established clear empirical data and computational models on patterns of eye movement control and sensitivity to lexical features during silent reading (e.g., Rayner 2009). However, to our knowledge, very little work has examined eye movements during the reading of text captioned speech in noise, in part because no prior study has utilized realistic text captioning. Our multi-word text captioning paradigm could be adopted to examine how acoustic challenge modulates eye movements during text captioned speech reading, which would speak more directly to the mechanisms underlying the text captioning benefit.

CONCLUSION

The findings from the current study suggest that realistic text captions provide a clear benefit to both immediate and delayed memory for degraded speech. Findings from Experiment 1 and Experiment 2 converged in showing reliable improvements across multiple measures of speech memory with text captions. Notably, this study is the first to our knowledge to show that text captioning improves not only immediate recall but also has long-term benefits to delayed recognition memory. Although the captioning benefit was strongest in cases where acoustic challenge was high (i.e., high background noise and lower hearing acuity), benefits for subsequent memory were even apparent in young normal-hearing listeners, suggesting they may yield broad and lasting benefits across literate adult listeners.

ACKNOWLEDGMENTS

Portions of this work were presented at the Academy of Rehabilitative Audiology Institute and the Annual Meeting of the Psychonomic Society (2018). B.R.P. and M.L. developed the study concept. All authors contributed to the study design and measures. Testing and data collection were performed by J.W.S. and H.A.C. B.R.P. and J.W.S. performed data analysis and interpretation. B.R.P., J.W.S., and M.L., drafted the manuscript. All authors provided critical revision. This research was funded by a research and development grant from CaptionCall, LLC.

The authors have no conflicts of interest to disclose.

Address for correspondence: Brennan R. Payne, Department of Psychology, 380 S 1530 E BEH S, Salt Lake City, UT 84112, USA. E-mail: brennan.payne@utah.edu

Received May 8, 2020; accepted April 28, 2021

REFERENCES

Balota, D. A., & Neely, J. H. (1980). Test-expectancy and word-frequency effects in recall and recognition. *J Exp Psychol Learn Mem Cogn, 6*, 576–587.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Front Psychol, 4*, 328.

Barr, D. J., Levy, R., Scheepers, C., Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *J Mem Lang, 68*, 255–278.

Bates, D., Kliegl, R., Vasishth, S., Baayen, H. (2015). Parsimonious mixed models. *ArXiv150604967 Stat.* http://arxiv.org/abs/1506.04967.

Benjamin, A. S., Bjork, R. A., Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *J Exp Psychol Gen, 127*, 55–68.

Benton, A. L., & Hamsher, K. D. (1978). *Multilingual Aphasia Examination: Manual of Instructions*. AJA Associates.

Bopp, K. L., & Verhaeghen, P. (2005). Aging and verbal memory span: A meta-analysis. *J Gerontol B Psychol Sci Soc Sci, 60*, P223–P233.

Busey, T. A., Tunnicliff, J., Loftus, G. R., Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychon Bull Rev, 7*, 26–48.

Conway, A. R., Cowan, N., Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychon Bull Rev, 8*, 331–335.

Cousins, K. A., Dar, H., Wingfield, A., Miller, P. (2014). Acoustic masking disrupts time-dependent mechanisms of memory encoding in word-list recall. *Mem Cognit, 42*, 622–638.

Cruickshanks, K. J., Wiley, T. L., Tweed, T. S., Klein, B. E., Klein, R., Mares-Perlman, J. A., Nondahl, D. M. (1998). Prevalence of hearing loss in older adults in Beaver Dam, Wisconsin. The Epidemiology of Hearing Loss Study. *Am J Epidemiol, 148*, 879–886.

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., McGettigan, C. (2005). Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *J Exp Psychol Gen, 134*, 222–241.

Ekstrom, R. B. (1976). *Kit of Factor-referenced Cognitive Tests*. Educational Testing Service.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *J Speech Hear Res, 12*, 423–425.

Folstein, M. F., Folstein, S. E., McHugh, P. R. (1975). "Mini-mental state". A practical method for grading the cognitive state of patients for the clinician. *J Psychiatr Res, 12*, 189–198.

Frost, R., & Katz, L. (1989). Orthographic depth and the interaction of visual and auditory processing in word recognition. *Mem Cognit, 17*, 302–310.

Gao, X., Stine-Morrow, E. A., Noh, S. R., Eskew, R. T. Jr. (2011). Visual noise disrupts conceptual integration in reading. *Psychon Bull Rev, 18*, 83–88.

Gernsbacher, M. A. (2015). Video captions benefit everyone. *Policy Insights Behav Brain Sci, 2*, 195–202.

Gilbert, R. C., Chandrasekaran, B., Smiljanic, R. (2014). Recognition memory in noise for speech of varying intelligibility. *J Acoust Soc Am, 135*, 389–399.

Gilchrist, A. L., Cowan, N., Naveh-Benjamin, M. (2008). Working memory capacity for spoken sentences decreases with adult ageing: Recall of fewer but not smaller chunks in older adults. *Memory, 16*, 773–787.

Gordon-Salant, S., & Callahan, J. S. (2009). The benefits of hearing aids and closed captioning for television viewing by older adults with hearing loss. *Ear Hear, 30*, 458–465.

Grant, K. W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective. *J Acoust Soc Am, 112*, 30–33.

Grant, K. W., & Braida, L. D. (1991). Evaluating the articulation index for auditory-visual input. *J Acoust Soc Am, 89*, 2952–2960.

Grant, K. W., & Greenberg, S. (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. In Proceedings of the audio-visual speech processing workshop (AVSP-2001), Aalborg, Denmark, pp. 132–137.

Grant, K. W., Greenberg, S., Poeppel, D., Van Wassenhove, V. (2004). Effects of spectro-temporal asynchrony in auditory and auditory-visual speech processing. *Semin Hear, 25*, 241–255.

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *J Acoust Soc Am, 108*(3 Pt 1), 1197–1208.

Grant, K. W., & Walden, B. E. (1996). Evaluating the articulation index for auditory-visual consonant recognition. *J Acoust Soc Am, 100*(4 Pt 1), 2415–2424.

Grossman, I., & Rajan, R. (2017). The effect of simultaneous text on the recall of noise-degraded speech. *J Exp Psychol Hum Percept Perform, 43*, 986–1001.

Haberlnadt, K. (1994). Methods in reading research. In E. M. Fernández, & H. Smith Cairns (Eds.), *The Handbook of Psycholinguistics* (pp. 1–31). Wiley-Blackwell.

Hall III, J. W., & Mueller, H. G. (1997). *Audiologists' Desk Reference: Audiologic Management, Rehabilitation, and Terminology*. Vol. 2. United Nations Publications.

Jones, J. A., & Freyman, R. L. (2012). Effect of priming on energetic and informational masking in a same-different task. *Ear Hear, 33*, 124–133.

Keetels, M., Stekelenburg, J. J., Vroomen, J. (2016). A spatial gradient in phonetic recalibration by lipread speech. *J Phon, 56*, 124–130.

Kintsch, W., & Van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychol Rev, 85*, 363–394.

Koeritzer, M. A., Rogers, C. S., Van Engen, K. J., Peelle, J. E. (2018). The impact of age, background noise, semantic ambiguity, and hearing loss on recognition memory for spoken sentences. *J Speech Lang Hear Res, 61*, 740–751.

Krull, V., & Humes, L. E. (2016). Text as a supplement to speech in young and older adults. *Ear Hear, 37*, 164–176.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., Bates, D. (2017). Balancing type I error and power in linear mixed models. *J Mem Lang, 94*, 305–315.

McCoy, S. L., Tun, P. A., Cox, L. C., Colangelo, M., Stewart, R. A., Wingfield, A. (2005). Hearing loss and perceptual effort: Downstream effects on older adults' memory for speech. *Q J Exp Psychol A, 58*, 22–33.

Nasreddine, Z. S., Phillips, N. A., Bédirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *J Am Geriatr Soc, 53*, 695–699.

Oswald, F. L., McAbee, S. T., Redick, T. S., Hambrick, D. Z. (2015). The development of a short domain-general measure of working memory capacity. *Behav Res Methods, 47*, 1343–1355.

Payne, B. R., & Federmeier, K. D. (2017). Pace yourself: Intraindividual variability in context use revealed by self-paced event-related brain potentials. *J Cogn Neurosci, 29*, 837–854.

Payne, B. R., & Silcox, J. W. (2019). Aging, context processing, and comprehension. *Psychol Learn Motiv, 71*, 215–264.

Payne, B. R., & Stine-Morrow, E. A. (2016). Risk for mild cognitive impairment is associated with semantic integration deficits in sentence processing and memory. *J Gerontol B Psychol Sci Soc Sci, 71*, 243–253.

Payne, B. R., & Stine-Morrow, E. A. L. (2017). The effects of home-based cognitive training on verbal working memory and language comprehension in older adulthood. *Front Aging Neurosci, 9*, 256.

Payne, B. R., Stites, M. C., Federmeier, K. D. (2016). Out of the corner of my eye: Foveal semantic load modulates parafoveal processing in reading. *J Exp Psychol Hum Percept Perform, 42*, 1839–1857.

Peelle, J. E. (2018). Listening effort: How the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear Hear, 39*, 204–214.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W., Humes, L. E., Lemke, U., Lunner, T., Matthen, M., Mackersie, C. L., Naylor, G., Phillips, N. A., Richter, M., Rudner, M., Sommers, M. S., Tremblay, K. L., Wingfield, A. (2016). Hearing impairment and cognitive energy: the framework for understanding effortful listening (FUEL). *Ear Hear, 37*(Suppl 1), 5S–27S.

Pichora-Fuller, M. K. (2003). Cognitive aging and auditory information processing. *Int J Audiol, 42*(Suppl 2), 2S26–2S32.

Piquado, T., Benichov, J. I., Brownell, H., Wingfield, A. (2012). The hidden effect of hearing acuity on speech recall, and compensatory effects of self-paced listening. *Int J Audiol, 51*, 576–583.

Piquado, T., Cousins, K. A., Wingfield, A., Miller, P. (2010). Effects of degraded sensory input on memory for speech: behavioral data and a test of biologically constrained computational models. *Brain Res, 1365*, 48–65.

Potter, M. C., & Lombardi, L. (1998). Syntactic priming in immediate recall of sentences. *J Mem Lang, 38*, 265–282.

Rabbitt, P. M. (1968). Channel-capacity, intelligibility and immediate memory. *Q J Exp Psychol, 20*, 241–248.

Rabbitt, P. (1991). Mild hearing loss can cause apparent memory failures which increase with age and reduce with IQ. *Acta Otolaryngologica, 111*, 167–176.

Rayner, K. (2009). The 35th Sir Frederick Bartlett Lecture: Eye movements and attention in reading, scene perception, and visual search. *Q J Exp Psychol, 62*, 1457–1506.

Schotter, E. R., & Payne, B. R. (2019). Eye movements and comprehension are important to reading. *Trends Cogn Sci, 23*, 811–812.

Schotter, E. R., Tran, R., Rayner, K. (2014). Don't believe what you read (only once): comprehension is supported by regressions during reading. *Psychol Sci, 25*, 1218–1226.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., Davis, M. H. (2014). Top-down influences of written text on perceived clarity of degraded speech. *J Exp Psychol Hum Percept Perform, 40*, 186–199.

Sommers, M. S., & Phelps, D. (2016). Listening effort in younger and older adults: A comparison of auditory-only and auditory-visual presentations. *Ear Hear, 37* (Suppl 1), 62S–68S.

Spehar, B., Tye-Murray, N., Myerson, J., Murray, D. J. (2016). Real-time captioning for improving informed consent: Patient and physician benefits. *Reg Anesth Pain Med, 41*, 65–68.

Stine, E. A., & Hindman, J. (1994). Age differences in reading time allocation for propositionally dense sentences. *Aging Cogn, 1*, 2–16.

Stine, E. A., & Wingfield, A. (1990). How much do working memory deficits contribute to age differences in discourse memory? *Eur J Cogn Psychol, 2*, 289–304.

Stine-Morrow, E. A., Milinder, L., Pullara, O., Herman, B. (2001). Patterns of resource allocation are reliable among younger and older readers. *Psychol Aging, 16*, 69–84.

Stine, E. L., Wingfield, A., Poon, L. W. (1986). How much and how fast: rapid processing of spoken language in later adulthood. *Psychol Aging, 1*, 303–311.

Stine-Morrow, E. A., Soederberg Miller, L. M., Gagne, D. D., Hertzog, C. (2008). Self-regulated reading in adulthood. *Psychol Aging, 23*, 131–153.

Strand, J. F., Brown, V. A., Barbour, D. L. (2020). Talking points: a modulating circle increases listening effort without improving speech recognition in young adults. *Psychon Bull Rev, 27*, 536–543.

Strauß, A., Wöstmann, M., Obleser, J. (2014). Cortical alpha oscillations as a tool for auditory selective inhibition. *Front Hum Neurosci, 8*, 1–7.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J Acoust Soc Am, 26*, 212–215.

Surprenant, A. M. (1999). The effect of noise on memory for spoken syllables. *Int J Psychol, 34*, 328–333.

Tombaugh, T. N., Kozak, J., Rees, L. (1999). Normative data stratified by age and education for two measures of verbal fluency: FAS and animal naming. *Arch Clin Neuropsychol, 14*, 167–177.

van Atteveldt, N., Formisano, E., Goebel, R., Blomert, L. (2004). Integration of letters and speech sounds in the human brain. *Neuron, 43*, 271–282.

Van Engen, K. J., Chandrasekaran, B., Smiljanic, R. (2012). Effects of speech clarity on recognition memory for spoken sentences. *PLoS One, 7*, e43753.

Waldron-Perrine, B., & Axelrod, B. N. (2012). Determining an appropriate cutting score for indication of impairment on the Montreal Cognitive Assessment. *Int J Geriatr Psychiatry, 27*, 1189–1194.

Weidemann, C. T., & Kahana, M. J. (2016). Assessing recognition memory using confidence ratings and response times. *R Soc Open Sci, 3*, 150670.

Wild, C. J., Davis, M. H., Johnsrude, I. S. (2012). Human auditory cortex is sensitive to the perceived clarity of speech. *Neuroimage, 60*, 1490–1502.

Wingfield, A., & Stine-Morrow, E. A. (2000). Language and speech. In F. I. Craik, & T. A. Salthouse (Eds.), *The Handbook of Aging and Cognition* (pp. 359–416). Lawrence Erlbaum Associates Publishers.

Wingfield, A., McCoy, S. L., Peelle, J. E., Tun, P. A., Cox, L. C. (2006). Effects of adult aging and hearing loss on comprehension of rapid speech varying in syntactic complexity. *J Am Acad Audiol, 17*, 487–497.

Wingfield, A., Tun, P. A., McCoy, S. L. (2005). Hearing loss in older adulthood: What it is and how it interacts with cognitive performance. *Curr Dir Psychol Sci, 14*, 144–148.

Wu, Y. H., Stangl, E., Chipara, O., Hasan, S. S., Welhaven, A., Oleson, J. (2018). Characteristics of real-world signal to noise ratios and speech listening situations of older adults with mild to moderate hearing loss. *Ear Hear, 39*, 293–304.

Yonelinas, A. P., & Jacoby, L. L. (2012). The process-dissociation approach two decades later: convergence, boundary conditions, and new directions. *Mem Cognit, 40*, 663–680.

Yuan, Y., Wayland, R., Oh, Y. (2020). Visual analog of the acoustic amplitude envelope benefits speech perception in noise. *J Acoust Soc Am, 147*, EL246.

Zekveld, A. A., Kramer, S. E., Kessens, J. M., Vlaming, M. S., Houtgast, T. (2008). The benefit obtained from visually displayed text from an automatic speech recognizer during listening to speech presented in noise. *Ear Hear, 29*, 838–852.