Julie I. Cohen and Sandra Gordon-Salant: JASA Express Letters [http://dx.doi.org/10.1121/1.4983399] Published Online 17 May 2017

The effect of visual distraction on auditory-visual speech perception by younger and older listeners

Julie I. Cohen[a] **and Sandra Gordon-Salant**
Department of Hearing and Speech Sciences, University of Maryland, College Park, Maryland 20742, USA
jcohen6@umd.edu, sgsalant@umd.edu

Abstract: Visual distractions are present in real-world listening environments, such as conversing in a crowded restaurant. This study examined the impact of visual distractors on younger and older adults' ability to understand auditory-visual (AV) speech in noise. AV speech stimuli were presented with one competing talker and with three different types of visual distractors. SNR_{50} thresholds for both listener groups were affected by visual distraction; the poorest performance for both groups was the AV + Video condition, and differences across groups were noted for some conditions. These findings suggest that older adults may be more susceptible to irrelevant auditory and visual competition in a real-world environment.

[DDO]
Date Received: December 22, 2016 **Date Accepted:** April 27, 2017

1. Introduction

Speech communication typically occurs in dynamic environments where, in addition to the speech signal of interest, there is time-varying noise, competing speech, and reverberation. Also present in everyday communication environments is a variety of relevant and irrelevant visual information. Some of this visual information provides enhanced speech understanding, as when visual information from the talker's face and body language complements the spoken auditory information. However, some of this visual information may be in the form of irrelevant and distracting visual input (referred to in this paper as visual distractors), as with a television program playing in the background, a person speaking in another conversation nearby, or a person walking within the visual field of view. While much is known about the benefit of visual information from speechreading as a supplement to auditory speech input, relatively little is known about the impact of irrelevant visual distraction on speech understanding in noise.

Speech perception is now widely accepted to be a multimodal process involving interactions between the auditory and visual input, especially in typical face-to-face communication situations. These multi-modal interactions have been studied extensively to determine the benefit afforded by visual cues when combined with auditory speech information, especially for listeners operating in noise and/or with hearing loss (Grant and Seitz, 1998). The amount of auditory-visual (AV) benefit typically increases as auditory-alone speech recognition deteriorates (Thorn and Thorn, 1989; Walden *et al.*, 1993; Tye-Murray *et al.*, 2007).

One issue not addressed in previous studies of AV speech perception is whether or not the presence of visual distractors has a negative impact on recognition performance. It could be hypothesized that visual distraction diverts the listener's attention from the primary speech perception task, resulting in a decline in speech perception performance. Recent evidence suggests that divided attention tasks are particularly difficult for face-matching (Palermo and Rhodes, 2002). These data suggest that human faces, other than that of the speaker, are especially difficult to ignore.

A related issue is the effect of listener age on the impact of visual distractors on performance. Previous studies examining the benefit of visual input from the speaker's face in an AV stimulus have shown that speech perception performance improves compared to auditory (A)-only input in both older and younger adults (Middelweerd and Plomp, 1987; Walden *et al.*, 1993; Sommers *et al.*, 2005; Jesse and Janse, 2012), although the magnitude of benefit may not be as great for older compared to younger listeners possibly due to age-related changes in auditory and visual-only perception

[a]Author to whom correspondence should be addressed.

(Tye-Murray *et al.*, 2010; Tye-Murray *et al.*, 2016). The presence of a competing or irrelevant visual signal could reduce or negate this benefit, especially for older adults. There is an age-related decline in divided attention and normal inhibitory processes (Hasher and Zacks, 1988), suggesting that older adults may be less able than younger adults to suppress a visual distractor and, as a result, will experience greater difficulty in AV speech perception performance in the presence of visual distraction compared to younger listeners.

The purpose of this study was to determine if the presence of visual distractors affects AV speech perception in older and younger normal hearing adults, and to determine if older adults experience greater detrimental effects than younger listeners on AV performance. The experiment evaluated three different types of visual distractors that are encountered in daily life: a talking face other than that of the primary speaker, text, and a video unrelated to the speech recognition task. It was hypothesized that performance for both younger and older listeners would decline as the visual distractor becomes more dynamic and salient, with the poorest performance observed for the competing video distractor, and best performance observed for the text distractor. It was also predicted that the older adults would perform more poorly than younger adults across all conditions.

2. Methods

2.1 Participants

Fifteen young adults (18–29 years, mean: 22.4 years), and 14 older adults (60–80 years, mean: 69.0 years), with normal hearing consistent with pure-tone thresholds of ≤ 25 dB hearing level (HL) from 250 to 4000 Hz were recruited for this study (Fig. 1). Further requirements for study inclusion were monosyllabic word recognition scores in quiet- $\geq 80\%$ (Northwestern University Auditory Test No. 6), normal tympanometry, and present acoustic reflex thresholds. Participants were screened for normal or corrected-normal vision, with a minimum visual acuity of 20/40 in both eyes using the Snellen chart. All participants were native speakers of English, and completed at least a high school level of education.

2.2 Stimuli

Two classes of stimuli were created: those with and without the presence of a visual distractor. There were three conditions with visual distractors, in which AV sentence stimuli were presented with the addition of (1) a second talking face (AV + Face), (2) a frozen caption (AV + Text), and (3) a short video clip (AV + Video). There were two conditions without visual distractors, one in which an auditory-only (A-only) stimulus was presented, and the other in which an AV stimulus was presented without a competing visual distractor (AV-only). The AV stimuli were selected from the TVM (Theo-Victor-Michael) sentence corpus (Helfer and Freyman, 2009). The original TVM corpus is composed of 1080 unique sentences, with 360 unique sentences for

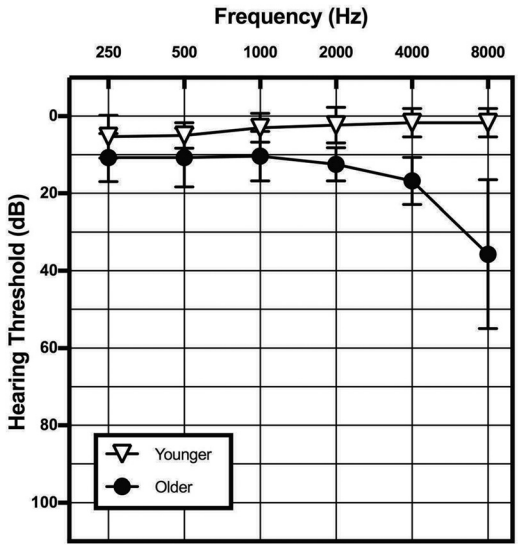

Fig. 1. Average pure-tone hearing threshold levels (dB HL) of the test ear for the younger and older listener groups. Error bars represent ± 1 standard deviation of the mean.

each call name (Theo, Victor, Michael). These stimuli were recorded originally by three native English male talkers. For the current study, 220 unique sentences spoken by two of these three talkers were selected (i.e., 100–120 unique sentences per talker, 20 of which were used for a practice condition as described below). The sentences followed the format of "*Call name* discussed the ____ and the ___ today," where the call name varied and the blanks represented the target words (nouns of one or two syllables). The same target talker was used for all experimental conditions and a second male talker was used as a competing talker. Thus, the competing sentences were never the same as the target sentences. The second male talker who recorded the original TVM sentences was used as the competing talker for all conditions with the exception of the AV + Video condition, for which 20 original competing sentences were recorded. The sentence structure for these new sentences modeled that of the TVM stimuli and used the format "*Call name* [verb] the [noun] [prepositional phrase] [noun]"; each sentence described the action that occurred during the competing video and was used as a "voice-over." For example, in one video a car drove in to a parking space and the recorded sentence was "Will parked the car in the lot today." These unique stimuli were recorded by a male native speaker of English onto a PC at a 44.1 kHz sampling rate using a Shure SM48 Vocal Dynamic Microphone, a Shure FP42 preamplifier, and Creative Sound Blaster Audigy soundcard.

The AV conditions, both with and without a visual distractor, were generated using the Adobe Premiere Pro (APP, Version 5) video editing software. The target talker visual for all videos, which showed a close-up of the head and shoulders of the talker, appeared in a box on the left side of the screen and was fixed in size across all conditions. For the AV-only condition, the visual of the target talker was present with no additional image on the screen. Three types of distractors were created for the distraction conditions. For the AV + Face condition, a competing talking face matching that of the competing auditory stimulus appeared on the right half of the screen. The AV + Text condition was composed of the target talker video and a frozen line of text that was centered on the bottom portion of the screen. The competing text corresponded to the sentence spoken by the competing talker (i.e., a closed caption). Last, for the AV + Video condition, a competing video appeared on the right side of the screen. The videos were recorded using a Flip Video UltraHD camera and edited in APP. Each video depicted a person doing a simple action (e.g., watering a plant or parking a car). As previously mentioned, the sentence spoken by the competing talker described this action.

The audio channels for the target and competing stimuli for all conditions were edited using Cool Edit Pro (version 2.0) to equate the root-mean-square level across all sentences, and to align the target and competing stimuli onset times. The generated auditory and video channels were then combined in APP; one list of 20 sentences was generated for each of the five test conditions (A-only, AV-only, AV + Face, AV + Text, and AV + Video) for a total experimental corpus of 100 unique sentences. All stimuli and distractors (when present) for each condition were burned to a DVD.

2.3 Procedures

The study was performed in a double-walled sound-attenuating booth, with participants seated 1.5 m away from the television screen. The visual stimuli were presented through a DVD player (Pioneer DV-490 V) and sent to a 25 in. Hannspree LCD television (HSG1074) located inside the test booth. The target and competing stimuli were routed through an audiometer (Interacoustics AC40) and presented monaurally via an insert earphone (Etymotic ER3A) to the better hearing ear, or to the right ear if hearing sensitivity was symmetrical across ears. Stimuli were presented monaurally to reduce the effects of possible interaural asymmetries or potential binaural interference, which may occur in some older adults with binaural stimulus presentation (Jerger *et al.*, 1993). The competing sentences were presented at a fixed level of 65 dBA and the target signal levels were adjusted adaptively to determine 50% correct sentence performance (SNR_{50}) similar to the procedure described for the Hearing in Noise Test (HINT) (Nilsson *et al.*, 1994). The first sentence in the list was presented at 0 dB SNR, and the target presentation level was increased in 4 dB steps until the listener responded correctly. A correct response was defined as the repetition of *both* nouns verbatim. The signal level was adjusted in 4 dB steps for the first four trials and then by 2 dB steps for the remaining sentences in the list. The SNR_{50} was calculated as the average presentation level of the 5th through the level at which the 21st sentence would

be presented. In all cases, listeners converged on their SNR_{50} by the 13th trial, with the remaining 7 trials confirming the reliability of the SNR_{50} estimate.

Prior to completing the experimental conditions, all participants completed a practice list that included examples of each of the five test conditions. The practice list was comprised of four sample stimuli from each condition presented at a fixed +10 dB SNR. None of the practice target or competing sentences were used in the experimental conditions. The experimental conditions were presented in a randomized order for each participant. The total listening time for each participant was approximately 1 h.

3. Results

3.1 Analyses

The approach to data analysis was to first compare performance of the two listener groups in the A-only and AV-only (non-distractor) conditions, using analysis of variance (ANOVA), to verify that all listeners derived the expected benefit of visual cues. Subsequently, a repeated measures ANOVA was conducted to determine whether there was an effect of visual distractor on performance (compared to the baseline AV-only condition), and whether older adults performed differently than younger adults when a visual distractor was present. In this analysis, there were four levels of the within-subjects "distractor condition" factor: AV-only (baseline), AV + Face, AV + Text, and AV + Video; listener group served as the between-subjects factor.

An analysis of covariance (ANCOVA) was conducted to control for possible differences in auditory-only speech recognition performance between the younger and older listener groups. In this analysis, the four levels of the within-subjects "distractor condition" factor and the between-subjects factor (listener group) were the same as those used in the repeated measures ANOVA; the A-only condition served as the covariate. Finally, a step-wise multiple linear regression analysis was conducted to determine which predictor variable, age or hearing sensitivity, contributed more to the variance in speech recognition performance in the various AV distractor conditions.

3.2 Auditory vs auditory-visual ability

SNR_{50} thresholds for the younger and older adults in the two conditions without visual distraction, A-only and AV-only, are shown in Fig. 2. It is apparent that the AV-only thresholds were significantly better (i.e., lower SNR) than in the A-only condition, particularly for younger adults. A repeated measures ANOVA showed a significant main effect of condition [$F(1,27) = 29.010$, $p < 0.01$, $\eta_p^2 = 0.518$], a significant main effect of group [$F(1,27) = 22.822$, $p < 0.001$, $\eta_p^2 = 0.458$], and a significant condition × group interaction [$F(1,27) = 5.962$, $p < 0.05$, $\eta_p^2 = 0.181$]. *Post hoc* analysis revealed that both younger and older adults had lower SNR_{50} thresholds in the AV-only condition, but older adults showed a smaller improvement than younger adults.

3.3 Effect of auditory-visual distraction

Mean SNR_{50} scores for the younger and older listeners in the four AV distractor conditions are illustrated in Fig. 3 (note that the AV-only data were also shown in Fig. 2

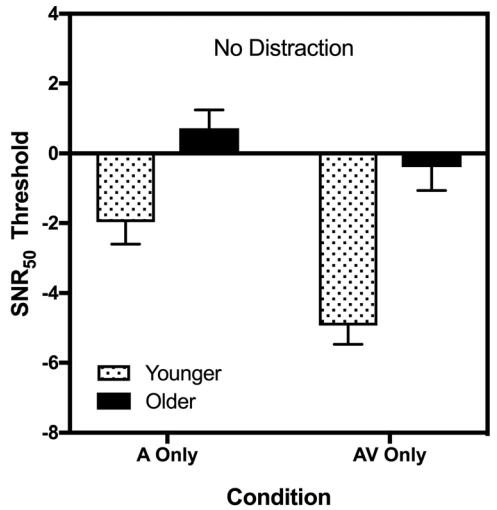

Fig. 2. Mean SNR_{50} thresholds for the younger and older listening groups in the A-only and AV-only conditions. Error bars represent ±1 standard error of the mean.

and represent baseline AV performance). SNR scores appear to be better for the younger listeners across all conditions. Additionally, the AV + Video distractor appears to have the greatest impact on the listeners' performance. Repeated measures ANOVA was conducted with one within-subjects variable, AV distractor condition (4 levels: AV-only, AV + Face, AV + Text, AV + Video) and one between-subjects variable, listener group. The results revealed a significant main effect of AV distractor condition [$F(1, 63.962) = 99.762$, $p < 0.001$, $\eta_p^2 = 0.787$], group [$F(1,27) = 12.930$, $p < 0.01$, $\eta_p^2 = 0.324$], and their interaction [$F(1, 63.962) = 3.073$, $p < 0.05$, $\eta_p^2 = 0.101$] (Geiser-Greenhouse correction used for degrees of freedom). *Post hoc* analysis of the AV distractor condition × group interaction revealed that both younger and older groups performed worse in the AV + Video condition than in the other distractor conditions ($p < 0.05$). Pairwise comparisons between younger and older groups for each condition indicated that the younger group performed significantly better than the older group in the AV-only and AV + Face conditions ($p < 0.01$). This suggests that different types of distraction impact younger and older listeners differently.

An ANCOVA was conducted to determine the impact of visual distraction across the four AV distractor conditions (AV + Face, AV + Text, AV + Video, and AV-only as the baseline no-distractor condition) for younger and older listeners, while controlling for their SNR_{50} thresholds on the A-only condition. The results of the ANCOVA revealed a significant main effect of distractor condition [$F(2.323, 60.403) = 104.554$, $p < 0.001$, $\eta_p^2 = 0.801$ (Greenhouse-Geisser correction)] and an interaction between distractor condition and group [$F(1, 60.403) = 5.315$, $p < 0.01$, $\eta_p^2 = 0.170$ (Greenhouse-Geisser correction)]. The main effect of group was not statistically significant [$F(1, 26) = 2.031$, $p > 0.05$, $\eta_p^2 = 0.072$]. Simple main effects analyses were conducted to examine the effect of distractor condition separately for each listener group, and the effect of group for each distractor condition. The effect of distractor condition was consistent for each group: AV speech perception was significantly poorer for the video distractor condition than all other conditions (AV-only, AV + Text, AV + Face; $p < 0.001$). Pairwise comparisons between groups for each distractor condition revealed a significant group difference for the AV-only condition ($p < 0.01$). That is, when thresholds measured in the A-only condition were accounted for, the age groups only differed on the AV-only (no distraction) condition.

A stepwise multiple linear regression analysis was conducted separately for each AV speech recognition measure with visual distraction. The purpose of this analysis was to determine if SNR_{50} thresholds in the AV distraction conditions could be predicted from participant age and hearing sensitivity. The predictor variable for hearing sensitivity was a high-frequency pure tone average (HFPTA), calculated as the average of thresholds for 1k, 2k, and 4k Hz. As seen in Table 1, the predictor of age was retrieved as the only significant variable in each condition, accounting for 13.5%–46.13% of the variance in thresholds. The variable HFPTA was not retrieved in any of the analyses, reinforcing that minor differences in hearing sensitivity between the two age groups did not contribute significantly to differences in AV speech recognition ability between them.

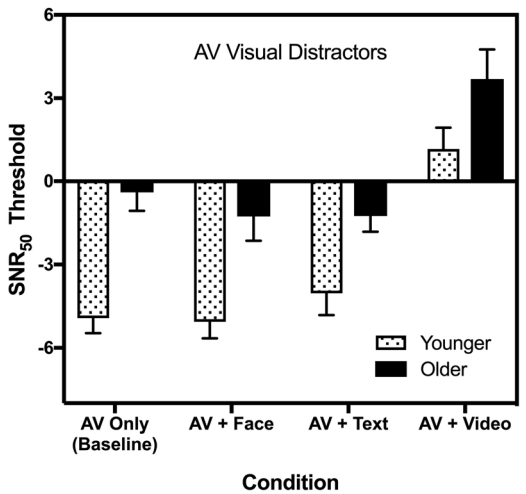

Fig. 3. Mean SNR_{50} thresholds for the younger and older listening groups across the AV distractor conditions, including the AV-only baseline measure. Error bars represent ± 1 standard error of the mean.

4. Discussion

The main purpose of this experiment was to examine the effect of visual distraction on AV speech perception ability by younger and older adults. The results generally showed that thresholds can increase with visual distraction, but that the type of visual distraction can have a differential effect on listener ability. That is, neither a competing text nor a competing face had a significant effect on SNR_{50} thresholds, relative to the AV-only (no distraction) condition, whereas the competing video had a significant effect. The face and text distractors could be considered low-level distractors as they did not involve considerable movement on the screen. The competing face only had subtle movements of the mouth, and the competing text appeared and then disappeared at the end of the stimulus. In contrast, the videos were more dynamic than the other two distractors because each depicted a different action. The finding of poorest ability by both groups in the video distraction condition suggests that a competing video results in a greater amount of distraction (higher SNR) than the other two distractors.

Performance on the tasks without visual distraction, A-only and AV-only, confirmed that both age groups received benefit from an AV stimulus. This finding is consistent with previous reports that both younger and older adults benefit from the addition of a visual cue (Cienkowski and Carney, 2002). The current findings also show that the older adults scored more poorly than the younger adults on most tasks, including those with no distraction. However, when "baseline" A-only thresholds were used as a covariate, there were no differences between groups in the different AV distraction conditions. This suggests that the older group was not more adversely affected by visual distraction than the younger group. It was expected that older adults would have greater difficultly than younger adults in the highly distracting environments due to age-related changes in inhibition and attention (Hasher *et al.*, 1991; Tun *et al.*, 2009) that are especially notable on tasks involving divided attention (Mattys and Scharenborg, 2014). It is possible that the older adults tested in this study did not differ from the younger adults in cognitive abilities of selective attention and inhibition, as these cognitive abilities were not measured specifically. Additionally, it is possible that younger adults are more likely than older adults to multi-task and switch attention between the target and competing video, whereas the older adults may be more likely to focus exclusively on the target to optimize performance. These two contrasting listening and watching strategies may have minimized age-related differences on the impact of the highly distracting competing video.

One of the major findings of this study was that the video distraction condition resulted in significantly poorer ability than the other AV distractor conditions. The finding should be viewed as tentative, however, because the AV + Video condition used a different competing talker than the other conditions. A new talker was required to record the sentences that accompanied the distracting videos created for this experiment. As noted earlier, new sentences describing the competing videos were generated that closely resembled the TVM structure and sentence duration; however, differences in the grammatical structure did exist. Additionally, regional dialect was somewhat different between the competing video talker and the original TVM talkers. Finally, the voice pitch of the competing video talker was higher in F0 than the original competing talker. This difference in voice pitch between the target talker and the competing talker of the video condition may have increased the masking release of the competing video talker relative to that achieved with the other competing male talker used in all other conditions (Bregman, 1990; Darwin *et al.*, 2003). Thus, the detrimental effect of a competing video may be even greater in everyday situations when the voice pitch of the competing talker is more similar to that of the target talker.

This study sheds some light on the impact of listening in a real-world environment where the auditory scene is composed of both competing auditory speech and visual distractors. In an attempt to quantify this effect in a laboratory setting, AV target stimuli were presented on a television in the presence of different visual distractors.

Table 1. Results of the stepwise multiple linear regression analysis for each AV visual distractor variable with predictors of age and HFPTA.

	Predictor variable	R^2	p value
AV + Face	Age	0.328	0.001
AV + Text	Age	0.219	0.010
AV + Video	Age	0.135	0.049

However, the AV target stimuli and distracting visual stimuli appeared in separate locations on the television screen. These distinct locations on the screen may have allowed the listener to completely ignore the low distraction conditions such as the competing face and text. In a real-world environment, dynamic visual distraction may be in the same visual frame (i.e., behind the speaker or partially in front of the speaker), and thus could cause a greater impact on performance.

Results of this study suggest that both younger and older adults are impacted by competing visual distraction, and that AV speech perception ability across younger and older adults varies with distractor type. Performance was poorest for both groups when listening in the presence of a competing video distractor, but few differences were observed across the other distractors compared to a baseline (i.e., no visual distraction) condition. It appears that younger and older adults may be susceptible to relatively dynamic distractions, as captured by the competing videos in the current experiment.

Acknowledgments

The authors thank Ken W. Grant for advice and assistance with this project, and Karen Helfer for providing the TVM stimuli. This research was conducted in a laboratory facility that was supported by Grant No. R01AG009191 from the National Institute on Aging, NIH.

References and links

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT Press, Cambridge, MA).

Cienkowski, K. M., and Carney, A. E. (**2002**). "Auditory-visual speech perception and aging," Ear Hear. **23**, 439–449.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913–2922.

Grant, K. W., and Seitz, P. F. (**1998**). "Measures of auditory-visual integration in nonsense syllables and sentences," J. Acoust. Soc. Am. **104**, 2438–2450.

Hasher, L., Stoltzfus, E. R., Zacks, R. T., and Rypma, B. (**1991**). "Age and inhibition," J. Exp. Psychol. Learn. Mem. Cogn. **17**, 163–169.

Hasher, L., and Zacks, R. T. (**1988**). "Working memory, comprehension, and aging: A review and a new view," in *Psychology of Learning and Motivation* (Elsevier, Amsterdam), pp. 193–225.

Helfer, K. S., and Freyman, R. L. (**2009**). "Lexical and indexical cues in masking by competing speech," J. Acoust. Soc. Am. **125**, 447–456.

Jerger, J., Silman, S., Lew, H. L., and Chmiel, R. (**1993**). "Case studies in binaural interference: Converging evidence from behavioral and electrophysiologic measures," J. Am. Acad. Audiol. **4**, 122–131.

Jesse, A., and Janse, E. (**2012**). "Audiovisual benefit for recognition of speech presented with single-talker noise in older listeners," Lang. Cogn. Process. **27**, 1167–1191.

Mattys, S. L., and Scharenborg, O. (**2014**). "Phoneme categorization and discrimination in younger and older adults: A comparative analysis of perceptual, lexical, and attentional factors," Psychol. Aging **29**, 150–162.

Middelweerd, M. J., and Plomp, R. (**1987**). "The effect of speechreading on the speech-reception threshold of sentences in noise," J. Acoust. Soc. Am. **82**, 2145–2147.

Nilsson, M., Soli, S. D., and Sullivan, J. A. (**1994**). "Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise," J. Acoust. Soc. Am. **95**, 1085–1099.

Palermo, R., and Rhodes, G. (**2002**). "The influence of divided attention on holistic face perception," Cognition **82**, 225–257.

Sommers, M. S., Tye-Murray, N., and Spehar, B. (**2005**). "Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults," Ear Hear. **26**, 263–275.

Thorn, F., and Thorn, S. (**1989**). "Speechreading with reduced vision: A problem of aging," J. Opt. Soc. Am. A **6**, 491–499.

Tun, P. A., McCoy, S., and Wingfield, A. (**2009**). "Aging, hearing acuity, and the attentional costs of effortful listening," Psychol. Aging **24**, 761–766.

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., and Hale, S. (**2010**). "Aging, audiovisual integration, and the principle of inverse effectiveness," Ear Hear **31**, 636–644.

Tye-Murray, N., Sommers, M. S., and Spehar, B. (**2007**). "Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing," Ear Hear. **28**, 656–668.

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., and Sommers, M. (**2016**). "Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration," Psychol. Aging **31**, 380–389.

Walden, B. E., Busacco, D. A., and Montgomery, A. A. (**1993**). "Benefit from visual cues in auditory-visual speech recognition by middle-aged and elderly persons," J. Speech Hear. Res. **36**, 431–436.