Audiovisual Speech Recognition With a Cochlear Implant and Increased Perceptual and Cognitive Demands

Anna R. Tinnemore, Sandra Gordon-Salant, and Matthew J. Goupell

Abstract
Speech recognition in complex environments involves focusing on the most relevant speech signal while ignoring distractions. Difficulties can arise due to the incoming signal’s characteristics (e.g., accented pronunciation, background noise, distortion) or the listener’s characteristics (e.g., hearing loss, advancing age, cognitive abilities). Listeners who use cochlear implants (CIs) must overcome these difficulties while listening to an impoverished version of the signals available to listeners with normal hearing (NH). In the real world, listeners often attempt tasks concurrent with, but unrelated to, speech recognition. This study sought to reveal the effects of visual distraction and performing a simultaneous visual task on audiovisual speech recognition. Two groups, those with CIs and those with NH listening to vocoded speech, were presented videos of unaccented and accented talkers with and without visual distractions, and with a secondary task. It was hypothesized that, compared with those with NH, listeners with CIs would be less influenced by visual distraction or a secondary visual task because their prolonged reliance on visual cues to aid auditory perception improves the ability to suppress irrelevant information. Results showed that visual distractions alone did not significantly decrease speech recognition performance for either group, but adding a secondary task did. Speech recognition was significantly poorer for accented compared with unaccented speech, and this difference was greater for CI listeners. These results suggest that speech recognition performance is likely more dependent on incoming signal characteristics than a difference in adaptive strategies for managing distractions between those who listen with and without a CI.

Keywords
cochlear implants, speech perception, attention, visual perception, perceptual masking

Received 2 June 2020; Revised 24 August 2020; accepted 28 August 2020
In everyday environments, speech is rarely the only sound present. Many factors, such as background noise and other talkers, can make listening situations more difficult. Background sounds not only obscure the target speech with acoustic energy in overlapping temporal and frequency domains (energetic masking) but also can distract the listener due to the linguistic and semantic content in competing speech (informational masking; Brungart, 2001; Durlach et al., 2003). In such noisy situations, complementary audio and visual information (e.g., being able to see the talker’s face) can benefit listeners with and without hearing loss (e.g., Sommers et al., 2005; Tye-Murray et al., 2007; Walden et al., 1993) including CI listeners (e.g., Waddington et al., 2020; Yi et al., 2013; Zhou et al., 2019).

Real-world environments contain distracting visual information that may be detrimental to recognizing speech. For example, the actions of others in the room or images on a television may distract a listener from looking at the target talker. To best understand the target talker and facilitate successful communication, the listener must suppress irrelevant auditory and visual information and focus attention on the relevant cues. Many studies have found that auditory distractors affect performance on an auditory task (e.g., speech in noise; Tun et al., 2002), and visual distractors affect performance on a visual task (e.g., distracted driving; Ito et al., 2001). Fewer studies have investigated distraction effects across modalities. In the few cross-modal distraction studies available, auditory distractors such as tone bursts have been shown to affect performance on visual tasks such as visual temporal rate perception (e.g., Recanzone, 2003), and visual distractors such as a screen of moving dots (Rees et al., 2001) or a competing video (Cohen & Gordon-Salant, 2017) have been shown to affect performance on auditory speech recognition tasks.

Studies on the effect of cross-modal distraction on CI listeners are rare, and the conclusions are mixed. A neuroimaging study found that activity in the visual processing areas of the brain (occipital cortex) during a speech recognition task preimplantation was positively correlated with speech understanding abilities 6 months after implantation (Strelnikov et al., 2013). Another study found that CI users had an advantage over listeners with NH in unimodal visual selective attention (Dye et al., 2009), attributing that advantage to visual activation of auditory cortex. Recently, Wang et al. (2020) used eye tracking to demonstrate differences in visual attention strategies for audiovisual (AV) speech understanding in a familiar and unfamiliar language. Adult listeners with CIs and hearing aids looked significantly longer at the mouth area of the talker than NH adults, who focused on the eyes. When the talker was speaking in an unfamiliar language, NH listeners adjusted their strategy to focus more on the mouth than with the familiar talker but did not approach the proportion of time CI listeners looked at the mouth (unchanged between familiar and unfamiliar languages). Yet another study showed that visual distractors interfered with a speech recognition task for inexperienced CI users (Champoux et al., 2009), but experienced CI listeners (defined as >6 months postimplantation) maintained high levels of performance on that auditory task despite the presence of visual distractors. Champoux et al. proposed that parts of primary auditory cortex had been recruited by visual processes due to long periods of deafness in the inexperienced listeners, resulting in decreased performance on an auditory task in the presence of visual distractors. They also proposed that the auditory cortex resumed its responsiveness to auditory stimuli through cortical plasticity for the experienced CI listeners.

An alternative explanation, and the one probed in this study, is that more experienced CI users had developed strategies that allowed them to focus their attention on relevant auditory stimuli and suppress irrelevant visual stimuli. Inexperienced CI users might not yet have developed this ability to suppress irrelevant visual stimuli, resulting in greater detrimental effects of the visual distractors on the auditory task for these listeners. In the current study, NH listeners presented with vocoded speech were expected to perform in a comparable manner to these inexperienced CI users because they, too, have not had time to develop the ability to efficiently suppress visual distractors during a speech recognition task with a degraded signal.

Age is an important factor to consider, given that the age range is typically large in many studies with CI listeners and that there is a growing population of CI listeners older than 65 years. Age-related changes in auditory sensitivity (e.g., Pearson et al., 1995), speed of processing (e.g., Salthouse, 1991), inhibition (e.g., Hasher & Zacks, 1988), and selective attention (e.g., Kausler & Hakami, 1982) may influence the effect of distractors on a primary task. Ben-David et al. (2014) assessed young and older adults’ reaction times during a visual detection task in the presence of visual distractors. Older adults had slower reaction times when a distractor was present than when one was not. This slowed reaction time was interpreted to mean that older adults had more difficulty inhibiting distractors than younger adults. The older adults showed no differences from younger adults when detecting targets or recognizing specific objects, which was interpreted to mean that
cognitive rather than sensory processes are the source of the age differences seen on this task.

In real-world environments, one cannot simply ignore everything but the talker. One must often attend to background events to perform a task unrelated to speech recognition, such as crossing the street or packing a lunch. In the lab, dual-task paradigms challenge the cognitive resources of the participants by requiring them to perform two tasks at the same time (Pashler, 1994). First, the primary and secondary tasks (STs) are each performed alone, and then they are performed together. Typically, participants perform better on the primary task when it is the only task. A decrease in performance on the secondary task while simultaneously performing the primary task is attributed to fewer cognitive resources being available for the secondary task because the resources are being used to execute the primary task (e.g., the limited capacity model; Kahneman, 1973). When resources are limited, alternate strategies for maximizing performance are often revealed.

The purpose of this study was to determine the effect of visual distractions on the speech recognition of older adults when listening to a the spectrally degraded auditory signal, either with a CI or via vocoded speech. A dynamic background of passing crowds provided visual stimuli unrelated to the primary auditory task and, along with background babble, increased the perceptual complexity of the scene. The auditory task difficulty was manipulated by presenting target speech spoken by unaccented or foreign-accented talkers and by requiring listeners to divide their attention by focusing on visual information irrelevant to their primary auditory task in a secondary visual task. Accented speech was used because it distorts the typical rhythm and timing cues of the sentences (e.g., Gordon-Salant et al., 2010) and increases difficulty for listeners who have age-related reductions in auditory temporal processing (Pichora-Fuller, 2003). Spanish-accented speech was used because many of the pronunciation alterations in Spanish-accented speech are temporally based, such as relative onset of voicing and frication for voiced and voiceless fricatives, vowel duration, and silence duration preceding fricatives as a cue for affrication (see Gordon-Salant et al., 2010 for an analysis of these features). While it is possible to artificially manipulate many of these same features (e.g., Gordon-Salant et al., 2006), using naturally spoken accented speech is more ecologically valid.

The first hypothesis was that visual distractions would adversely affect all listeners who are challenged by receiving spectrally degraded speech, such as through a CI or a vocoder. Because listeners are processing a distorted signal, they may need to devote their finite cognitive resources to this processing task leaving insufficient resources for inhibiting distractors. The second hypothesis was that the distraction effect would be greater for all listeners when speech is presented with a foreign accent (i.e., when there are fewer temporal cues available to accurately perceive the target signals; e.g., Bent et al., 2008). As the demands of understanding foreign-accented speech tax listeners' cognitive resources, listeners may find it increasingly difficult to suppress irrelevant stimuli. The third hypothesis was that adding a secondary task, requiring listeners to pay attention to the background scene, would cause a greater decrease in performance for all listeners than simply having the background scene present, because adding a secondary task requires more cognitive resources than the primary task alone. The fourth and overarching hypothesis was that CI listeners would show less difference in their speech recognition between conditions with and without distractions than their NH peers listening to vocoded speech, because their long-standing reliance on visual cues to aid auditory perception of degraded signals has enabled CI listeners to differentially develop the ability to suppress irrelevant information. The fifth and final hypothesis was that older listeners would show a greater distraction effect than younger listeners because of age-related decreases in the ability to inhibit distractors.

**Experiment**

**Listeners**

Two groups of 20 listeners were recruited. One group was composed of CI users with at least 1 year of experience with their device(s). They were tested using their own processors with their personal, clinical settings. They listened with their CIs only—if they had a single CI, the nonimplanted ear had an earplug placed in it. The CI group had an average age of 57.2 years and ranged from 21 to 78 years. Additional demographic information can be found in Table 1.

The second group was composed of listeners who were matched in age (± 2 years) to the listeners in the CI group (average = 57.2 years old, range = 21–77 years). These listeners had near-NH in both ears (thresholds ≤25 dB HL at 0.5, 1, and 2 kHz and ≤35 dB HL at 4 kHz). They were matched in performance to their age-matched CI listener’s auditory-only speech recognition abilities by adaptively changing the number of vocoded channels in sentence stimuli (see details in the following Procedure section).

All listeners were native-English speakers with visual acuity (with or without correction) of 20/50 or better on a Snellen eye chart. They each passed a cognitive screening measure (the Montreal Cognitive Assessment; Nasreddine et al., 2005) with a minimum score of 22/30 to rule out any more than mild cognitive impairment (e.g., Dupuis et al., 2015; Goupell et al., 2017). The Institutional Review Board at the University of
Maryland approved the experimental protocol, and all participants provided written informed consent.

**Stimuli**

**Visual.** AV stimuli were created using Adobe Premiere Pro CC 2015. The videos of the target talkers were from the same corpus as those described in Waddington et al. (2020). They were professionally recorded at the National Foreign Language Center (University of Maryland) using green-screen technology and consisted of a close-up of the head and shoulders of a talker speaking sentences from the Institute of

| Listener number | Group | Age (years) | Baseline audio-only speech recognition (%) | Duration of deafness (years) | Device configuration |
|-----------------|-------|-------------|------------------------------------------|-----------------------------|----------------------|
| 1               | CI    | 70          | 72                                       | R: 1                        | Unilateral           |
| 2               | CI    | 73          | 66                                       | R: < 1, L: 22               | Bilateral            |
| 3               | CI    | 60          | 52                                       | R: 12, L: 4                 | Bilateral            |
| 4               | CI    | 61          | 42                                       | R: 11, L: 11                | Bilateral            |
| 5               | CI    | 54          | 94                                       | R: 3, L: 6                  | Bilateral            |
| 6               | CI    | 50          | 42                                       | R: 8, L: 4                  | Bilateral            |
| 7               | CI    | 50          | 56                                       | R: 20, L: 24                | Bilateral            |
| 8               | CI    | 76          | 72                                       | R: 19, L: 17                | Bilateral            |
| 9               | CI    | 39          | 70                                       | L: 8                        | Unilateral           |
| 10              | CI    | 21          | 82                                       | R: 2, L: 3                  | Bilateral            |
| 11              | CI    | 78          | 90                                       | R: 7, L: 3                  | Bilateral            |
| 12              | CI    | 43          | 40                                       | L: 37                       | Unilateral           |
| 13              | CI    | 55          | 70                                       | R: 7                        | Unilateral           |
| 14              | CI    | 56          | 46                                       | R: 4, L: 2                  | Bilateral            |
| 15              | CI    | 37          | 98                                       | R: 3, L: 2                  | Bilateral            |
| 16              | CI    | 67          | 86                                       | R: 8, L: 2                  | Bilateral            |
| 17              | CI    | 70          | 52                                       | R: 3, L: 1                  | Bilateral            |
| 18              | CI    | 72          | 60                                       | R: 7, L: 2                  | Bilateral            |
| 19              | CI    | 61          | 88                                       | R: < 1, L: 5                | Bilateral            |
| 20              | CI    | 50          | 68                                       | R: 1, L: 5                  | Bilateral            |

| Listener number | Group | Age (years) | Baseline audio-only speech recognition (%) | Vocoder channels used | Change in vocoded speech recognition (post–pre) (%) |
|-----------------|-------|-------------|------------------------------------------|----------------------|-----------------------------------------------|
| 1               | NH    | 72          | 78                                       | 6                    | 10                                            |
| 2               | NH    | 71          | 68                                       | 7                    | 12                                            |
| 3               | NH    | 60          | 66                                       | 4                    | 2                                             |
| 4               | NH    | 62          | 56                                       | 5                    | −10                                           |
| 5               | NH    | 55          | 92                                       | 10                   | 8                                             |
| 6               | NH    | 52          | 40                                       | 3                    | 8                                             |
| 7               | NH    | 49          | 48                                       | 4                    | −2                                            |
| 8               | NH    | 74          | 66                                       | 6                    | 18                                            |
| 9               | NH    | 40          | 70                                       | 3                    | −6                                            |
| 10              | NH    | 21          | 78                                       | 7                    | 8                                             |
| 11              | NH    | 77          | 94                                       | 8                    | −12                                           |
| 12              | NH    | 44          | 44                                       | 3                    | −8                                            |
| 13              | NH    | 55          | 76                                       | 4                    | −24                                           |
| 14              | NH    | 57          | 42                                       | 3                    | −10                                           |
| 15              | NH    | 37          | 90                                       | 8                    | −4                                            |
| 16              | NH    | 66          | 84                                       | 8                    | 2                                             |
| 17              | NH    | 68          | 42                                       | 3                    | −10                                           |
| 18              | NH    | 70          | 68                                       | 5                    | −4                                            |
| 19              | NH    | 62          | 76                                       | 7                    | 16                                            |
| 20              | NH    | 51          | 60                                       | 4                    | 0                                             |

*Note. CI = cochlear implant; NH = normal hearing.*

---

Table 1. Demographic Information for Paired Sets of Listeners in the Two Listener Groups Including Age and Baseline Speech Recognition, Duration of Deafness (CI Group), Device Configuration (CI Group), and Number of Vocoder Channels for Matched Performance (NH Group).
Electrical and Electronics Engineers (IEEE) corpus (IEEE, 1969). Both talkers were adult males without facial hair. The background video selected for this study showed a ground-level view along a busy outdoor walkway with people walking toward and away from the camera with no accompanying audio. This video was downloaded with permission from the creator (Exploring Alabama, 2017). Only selected segments of the video were played such that each clip contained one to five people entering the screen from the right side and walking away from the camera toward the middle of the top of the screen. A varying number of people were already in the scene or were walking toward the camera.

Videos were created to test six conditions (three stimulus/task conditions by two talkers—unaccented and accented). Each talker was featured in three videos. The videos were made so that each could be used to test any of the stimulus/task conditions. This entailed creating two versions of each video. One version showed the head and shoulders of the talker in front of a solid dark blue background. This version was used for the AV-only condition. Between each spoken sentence, there was a black screen shown for 6 seconds with a prompt to repeat the sentence aloud. The second version of each video had the same foreground videos of the talker and breaks between sentences, but each sentence was presented with the background video of a busy boardwalk scene. The target-talker and boardwalk-scene videos were combined using green-screen technology so that the talker appeared in front of the boardwalk scene. This version was used for the distraction and secondary task conditions.

A separate video was created for the secondary task-alone condition. One of the talkers spoke 20 sentences that were not used elsewhere in the protocol, and the background video clips each portrayed one to five people entering the scene and walking away from the camera. A practice video was created to allow the listener to hear and see stimuli representative of the six test conditions and to practice the secondary task. Two sentences that were not used in the rest of the protocol were repeated throughout the practice video by both talkers.

Audio. The two experimental talkers had marked differences in their pronunciation of English sentences. One talker was a male native speaker of American English, and the other was a male native speaker of Spanish from Peru. Both talkers were enrolled in graduate programs at the time of recording and were 31 and 39 years old, respectively. The talker from Peru started learning English at 6 years old, moved to the United States in 2006, and recorded the AV stimuli in 2015. He was rated as having a moderate accent when speaking English (average rating of 5.7 on a scale of 1 to 9, with 9 being very heavily accented; following Munro & Derwing, 1995) in a separate preliminary experiment with 19 NH native-English listeners younger than the age of 30 years (Waddington et al., 2020). A third talker (male, native speaker of American English, mid-30s) made recordings that were used for the performance-matching procedure and baseline testing.

In addition to the accent ratings, five young NH monolingual American English speakers gave ratings of the perceived intelligibility of each sentence spoken by the accented talker. These ratings were used to select only those IEEE sentences with an average rating between 75% and 95% intelligible. This reduced set of sentences was then randomized and divided into six lists of 20 sentences each (experimental stimuli), one list of 40 sentences (performance-matching stimuli), and four lists of 10 sentences (one used for baseline testing for CI listeners and verification testing for NH listeners, and three used for additional verification testing if needed and for posttesting of NH listeners). The final lists each had an average intelligibility rating of 87% when spoken by the accented talker.

The audio recordings presented with the videos to the CI group were low-pass filtered at 4 kHz with a third-order Butterworth filter using Adobe Audition CC 2015 and combined with six-talker babble. The stimuli were low-pass filtered at 4 kHz because many of the nominally NH listeners were older and had hearing loss at frequencies higher than 4 kHz. Filtering the stimuli allowed for a better comparison between groups because they had access to similar acoustic cues. The babble, described in an earlier study (Gordon-Salant et al., 2013), consisted of six male talkers, three with American English accents and three with Spanish accents. The level of the stimulus sentences was equalized in root-mean-square (RMS) intensity in MATLAB (Mathworks, Natick, MA). Segments of babble, matched in length to each stimulus, were combined with the stimuli at a +10 dB signal-to-noise ratio (SNR) in a single audio track for presentation via loudspeaker. The audio was then synchronized with the videos. The +10 dB SNR was selected based on pilot testing with CI listeners, as an SNR that avoided ceiling and floor effects.

For the NH control group, the filtered, level-equalized IEEE sentences with babble added at the specified SNR were vocoded using MATLAB. For an n-channel vocoder, the auditory speech signal was band-pass filtered using third-order forward-backward Butterworth filters into n logarithmically spaced bands (36 dB/octave) between 200 and 4000 Hz. The temporal speech envelope from each band was extracted with a Hilbert envelope cutoff of 400 Hz and used to modulate n sine carriers of the center frequency of each channel. The modulated sine carriers were then combined to create the final vocoded output. This method was used to create 31 vocoded versions of the audio stimuli.
representing 2 to 32 channels. These vocoded audio tracks were then synchronized to each of the 12 videos (two versions of six videos), the secondary task video, and the practice video. In total, 31 additional sets of the 14 videos were created.

Audio recordings from the third talker were used to create one list of 40 sentences and four lists of 10 sentences for purposes of baseline testing, performance matching, and verification testing. These five lists had no background noise, were low-pass filtered at 4 kHz, and were equated in RMS intensity to the other stimuli. The four lists of 10 sentences were processed in MATLAB to create 31 vocoded versions as described earlier, while the list of 40 sentences was divided into four subsets that were vocoded with 4, 8, 16, and 32 channels.

**Procedure**

Initially, all listeners completed screening measures of vision, cognitive awareness, and hearing thresholds to ensure they met the criteria for the study. Each listener also completed three other cognitive measures: the NIH Toolbox Flanker Inhibitory Control and Attention Test (Zelazo et al., 2014), the NIH Toolbox Pattern Comparison Processing Speed Test (Carlozzi et al., 2014), and the NIH Toolbox Dimension Change Card Sort (a measure of executive control; Zelazo et al., 2014).

Before the main experiment, CI listeners performed a baseline speech recognition task, and NH listeners performed a performance-matching task. For both tasks and the main experiment, listeners were seated in a double-walled sound-attenuating booth (IAC, NY) 1.5 meters from a 37-inch LCD television screen (Model: LN32D450G1D, Samsung Electronics America, Ridgefield Park, NJ). The screen was positioned directly in front of the seated listener and at eye level. A single loudspeaker (Model: HS 50 M, Yamaha Corporation of America, Buena Park, CA), also at 1.5 meters and directly in front of the listener, was used to present the audio tracks at 65 dB(A). The AV stimuli were played on an Apple Air II laptop. The video was routed to the presentation TV in the booth via an HDMI cable; the audio was routed through an audiometer (Model: AC40, Interacoustics, Eden Prairie, MN) and played via a single channel to the free-field loudspeaker. The overall speech level was calibrated before each test session by playing speech-shaped noise equal in average RMS intensity to the stimuli and observing the sound pressure level measured on a sound level meter (Brul & Kjaer 2250, Brul & Kjaer North America, Duluth, GA) using the substitution method (Beynon & Munro, 1993).

Each sentence in the IEEE corpus has five keywords. The number of keywords correct out of the number of keywords possible per list was used to compute the speech recognition performance of each listener.

**Performance Matching.** A performance-matching procedure was used to determine the number of vocoded channels needed for each individual NH listener to have equal auditory-only speech recognition performance for an unaccented talker with their age-matched peer who listened with a CI. All CI listeners were presented with the audio of one list of 10 sentences in quiet (spoken by the third talker as described earlier) and were asked to repeat the sentences aloud. Their performance on this task became their baseline speech recognition score for the purposes of matching performance.

NH listeners were presented with two practice sentences (spoken by the third talker) accompanied by their written transcripts in both the relatively easiest (32 channel) and relatively most difficult (4 channel) vocoded conditions to gain familiarity with the spectral degradation of the auditory signal. They were then asked to repeat each sentence from the list of 40 test sentences, which had been vocoded into 4, 8, 16, and 32 channels and presented in a randomly interleaved order. The number of correct keywords was computed into a percent correct for each level of vocoding, and a logistic regression with four parameters was fit to the data points (Psignifit, Wichmann & Hill, 2001). This psychometric function allowed the computation of the number of channels (rounded to the nearest integer channel) each NH listener needed to match the performance of their age-matched CI listener. When the number of channels was estimated, the listener heard a new list of 10 sentences vocoded with that number of channels to confirm that performance was matched. If the NH listener’s percent correct on the confirmation list was not within 15% of their age-matched peer, the number of channels was adjusted based on their performance (one channel fewer if performance was too high, one channel more if performance was too low), and a different confirmation list was administered. Half of the NH listeners (10/20) required at least one additional list. This confirmation process also provided practice for the NH listener with the vocoded speech that would be used in the main experiment.

**Listening Conditions and Tasks.** All listeners in both the CI and NH listener groups participated in seven conditions (three stimulus/task conditions by two talkers—unaccented and accented—and the visual task-only condition). The first stimulus/task condition, AV only, used the version of the video that showed the target talker with the solid background. The second stimulus/task condition, AV plus visual distractions (AV+D), used the version of the video that showed the target talker with the busy background scene. For these two
conditions, listeners were required to repeat the target sentence presented. In the AV+D condition, they were instructed to ignore the background. The third stimulus/task condition, AV plus visual secondary task (AV+ST), used the version of the video that showed the target talker with the busy background scene. Listeners were instructed to repeat the target sentence and report the number of people who entered the scene and walked away from the point of view of the camera. In other words, individuals already present on the screen when the scene started and those walking toward the camera were not counted. Listeners were instructed to prioritize the primary and secondary tasks equally. This created a scenario where decrements in performance on either task could be attributed to cognitive resources being directed to the other task. In the ST-alone condition, the listeners were asked to only count the number of people walking away using the same aforementioned criteria while ignoring the spoken sentences (spoken by the unaccented talker). In the ST-alone and AV+ST conditions, the listeners were scored on whether the reported number of people walking away was correct or incorrect. Response time was not measured because of the nature of the secondary task—counting rather than detection of a visual target. All listeners had to wait until the end of the video clip, repeat the sentence, and then provide the number of visual targets.

Before beginning the main experiment, the practice video was presented using one of two practice sentences for each of the six sentence-repetition conditions. Both talkers had a sentence with the solid color background, a sentence with the distracting video, and a sentence with the secondary task. If mistakes were made or questions were asked, listeners could repeat the practice video until both the sentence-repetition task and the secondary task were clearly understood. After the practice video, both the order of the conditions and assignment of sentence list to condition were counterbalanced across listeners. Within each video, the sentences were shown in a fixed order. Each NH age-matched listener completed the protocol in the same order and with the same videos as their matched CI listener.

After completing the main experiment, the NH listeners were presented with another list of 10 sentences spoken by the third talker in an auditory-only condition; these sentences were vocoded with the same number of channels that were used for each listener in the main experiment. The purpose of this final task was to verify that the NH listeners had plateaued in their ability to recognize vocoded speech and had not continued to improve throughout the main experiment. Only three of the NH listeners scored outside of the targeted ± 15% from their CI peer at the end of the experiment. Two scored 16% better than the target, and one scored 18% worse than the target (see Table 1).

Data Analysis

Data from the six speech recognition conditions were analyzed using a generalized linear mixed-effects regression analysis with a binomial distribution. This method was chosen because it can take full advantage of the power of trial-level data, allows for nested effects, and calculates random effects for individual factors providing insight into the variability so common in CI and NH vocoder research. The mixed-effects model was fit to the trial-by-trial data using R version 3.6.2 (R Core Team, 2019), the buildmer package version 1.5 (Voeten, 2020), and the lme4 package version 1.1.23 (Bates et al., 2015). Generalized linear regression models, such as the model reported here, predict the (log-odds) probability of a correct response. The dependent variable was the proportion of correct keywords out of the five possible keywords per sentence. The referent levels were the NH listener group, the AV-only condition, and the unaccented talker. This means, for example, that a statistic labeled only “CI group” represents the difference in performance between the CI group and the NH listener group in the referent conditions (AV-only with an unaccented talker) with all other factors held constant. Age, baseline speech recognition, and each of the three cognitive measures were standardized and centered with a mean of zero and a standard deviation of one. Initial models with all three cognitive measures failed to converge due to collinearity between these measures. Therefore, a composite cognitive score (the average of the age-corrected scores for all three cognitive tests which was then standardized and centered) was used to represent the effect of cognitive abilities on speech recognition performance.

The buildmer package (v. 1.5, Voeten, 2020) using the default options performs model testing by determining the order of the random and fixed effects in the model that explain the most variance and result in a maximal model that will converge (Barr, 2013). The effects are then systematically reduced with backward stepwise elimination based on likelihood ratio tests to arrive at the final model of best fit that can be supported by the observed data. The number of iterations was set to 200,000 to allow models to converge, and the bound optimization by quadratic approximation (bobyqa) optimizer was used.

The full model incorporates all fixed effects and interactions of interest, including experimental variables (talker, condition, group) and other predictors (age, baseline speech scores, cognitive measures) and a maximal random effects structure (Barr, 2013). After the best-fitting model was discovered using a cognitive composite score, the model was refit with each of the individual cognitive scores and compared with the model using the composite score. Only the model using the
single cognitive measure of executive control was signif-
icantly better than the model using the composite score, 
\[\chi^2(0, N = 4,781) = 0.94, \quad p < .001,\] and is therefore pre-

The resulting model structure is reported in Table 2, which contains coefficient estimates ("Estimate") as a measure of effect size and a Wald test of statistical signif-
icance. Effects that were not significant or were not part of significant interactions (likelihood ratio test, \(p > .05\)) were dropped from the model and are not reported in the summary table. The formula for the final model is as follows: The Proportion of Correct Keywords Per Sentence is predicted by (\(\cdot\)) fixed effects of Talker \((\text{Condition + Group})\) + Baseline Speech Recognition \(\times\) Condition + Age + Executive Control + random effects of (Condition \(\times\) Talker | Listener) + (Group + Age \(\times\) Condition + Baseline Speech Recognition + Executive Control | Sentence). The model was run on 4,781 observations from 40 listeners to 60 sentences. There are not 4,800 observations because one listener reported recognizing 19 of the sentences randomly scattered among the conditions from a previous exposure to IEEE stimuli. The sentences she reported as being famil-

| Fixed effects                                      | Estimate | SE  | z    | p    |
|---------------------------------------------------|----------|-----|------|------|
| (Intercept)                                       | 0.86     | 0.23| 3.66 | <.001***|
| TalkerAccented                                    | -1.75    | 0.23| -7.46| <.001***|
| ConditionAV+D                                     | 0.05     | 0.11| 0.43 | .69   |
| ConditionAV+ST                                    | -0.85    | 0.12| -7.11| <.001***|
| GroupCI                                           | 0.42     | 0.27| 1.54 | .123  |
| Baseline speech recognition                       | 0.56     | 0.12| 4.52 | <.001***|
| Age                                               | -0.08    | 0.12| -0.67| .502  |
| Executive control                                 | 0.35     | 0.13| 2.75 | .006**|
| TalkerAccented:ConditionAV+D                      | 0.06     | 0.16| 0.39 | .699  |
| TalkerAccented:ConditionAV+ST                      | 0.46     | 0.17| 2.76 | .006**|
| TalkerAccented:GroupCI                            | -0.46    | 0.17| -2.76| .006**|
| ConditionAV+D:Baseline speech recognition         | 0.05     | 0.06| 0.88 | .382  |
| ConditionAV+ST:Baseline speech recognition        | -0.16    | 0.08| -2.12| .034* |
The values in the “SD” column of the random effects section of Table 2 are the standard deviations of those effects from the group mean. For example, the model estimates that across listeners, the effect of talker accent is –1.75 (on a log-odds scale) or about 40%, but the size of this varies across listeners with a standard deviation of 0.57 (on the same scale). This means that if we sampled from a similar population of listeners, we would expect 95% of them to show effects of accent between –0.63 and –2.87 or between 14% and 58%.

Performance on the secondary task was also evaluated. The cost of the secondary task was calculated for each subject by comparing performance during the dual-task condition (AV+ST) to that in the secondary task-alone condition. These values were divided by performance in the secondary task-alone condition to normalize across listeners (as in Gosselin & Gagné, 2011). These cost values were then used as the dependent variable in a repeated-measures analysis of variance (RM-ANOVA) with group (CI vs. NH vocoder) as the between-subjects factor and talker (unaccented vs. accented) as the within-subjects factor. This strategy was chosen to allow for straightforward comparisons with previously published results.

Results

Speech Recognition Performance

The average percentage of correctly repeated words in the AV-only, AV+D, and AV+ST conditions for both unaccented and accented talkers and for the two listener groups are plotted in Figure 1. Performance was comparable across the two listener groups. Both listener groups showed higher speech recognition scores for unaccented compared with the accented speech, and performance appears to be poorer in the AV+ST condition compared with the other two conditions.

A generalized linear regression model was created for this analysis. The variables entered into the model and how to interpret the model output are described in the Data Analysis section. The fitted model shown in Table 2 indicates statistically significant effects of executive control (Dimension Change Card Sort) and interactions of talker accent, group, condition, and baseline speech recognition accuracy.

The effect of an accented talker was significantly greater for CI listeners than the NH vocoder listeners (est. = –0.46, z = –2.76, p = .006). As seen in Figure 1, CI listeners performed slightly better than their NH vocoder peers in the unaccented condition (78% compared with 70% correct; est. = 0.42, z = 1.54, p > .05), but worse in the accented condition. There were no significant interactions of talker accent and group with any of the levels of condition (all had p > .05), meaning that this greater effect of talker’s accent on the performance of the CI listener group was similar across all conditions.

Speech recognition accuracy decreased more in the AV+ST condition when listening to the unaccented talker than it did for the accented talker (est. = 0.46, z = 2.76, p = .006; Figure 1). This was similar for all listeners because there was no significant interaction with listener group (p > .05).

There was also an interaction of condition by baseline speech recognition scores. Listeners with above-average baseline speech recognition scores performed slightly worse on the AV+ST condition than listeners with below-average baseline speech recognition scores (est. = –0.16, z = –2.12, p = .03). The interactions with group were not significant (p > .05) meaning that this effect was similar for all listeners, and the absence of an interaction with talker accent (p > .05) means that the effect of accent on the AV+ST condition was independent from the effect of baseline speech recognition ability on the AV+ST condition.

Higher-than-average scores on the measure of executive control (Dimension Change Card Sort) were associated with significantly higher speech recognition performance. The model summary provides statistics for the NH listeners in the AV-only condition with the unaccented talker (est. = 0.35, z = 2.75, p = .006), and the lack of a significant interaction with the other factors.
means that higher-than-average executive control affected the performance of both groups and all conditions similarly \((p > .05)\) for all two-, three-, and four-way interactions between executive control, group, talker, and condition. Greater-than-average age did not significantly affect speech recognition performance \((est. = –0.08, z = –0.67, p > .05)\), but this factor was left in the model to directly address the fifth hypothesis.

In summary, this analysis shows significant effects of executive control and interactions of talker accent with group, talker accent with condition, and baseline speech recognition with condition. Listeners with higher scores on the measure of executive control had better speech recognition accuracy in all conditions than those with below-average scores. The effect of the accented talker was greater for the CI group than the NH vocoder group, although both showed decrements with the accented talker. Accuracy in the AV+ST condition declined more from the baseline AV-only condition when listeners heard the unaccented talker than when they heard the accented talker and did not decline as much for listeners with lower baseline speech recognition scores than for those with higher scores. Greater-than-average age did not significantly affect performance.

**Secondary Task Performance**

The average percent correct performance of each group for each condition is shown in Figure 2. The stimuli for the secondary task were designed so that each sentence-length background video had a well-defined, limited number of occurrences of the activity to be counted (between 1 and 5). Listeners reported the number of individuals walking away from the viewer during each video segment. Performance on the secondary task was scored as the accuracy of the numbers reported. The average accuracy on the secondary task was at ceiling in the secondary task-alone condition. Accuracy was not different across groups in this baseline condition (secondary task-alone; paired samples \(t\) test, \(p = .27\), meaning that both groups were able to perform the secondary task-alone condition with great accuracy. In the current study, listeners were instructed to prioritize the primary and secondary tasks equally. Lower accuracy on the secondary task during the dual-task conditions (AV+ST) than during the secondary task-only condition may suggest that the listener’s cognitive resources cannot fully cover the demands of both tasks at once. As seen in Figure 2, both groups had a decrease in performance on the secondary task when also performing the primary speech recognition task.

The cost of a dual task can be calculated as the difference in performance on the secondary task from the secondary task-only condition to the AV+ST conditions relative to performance in the secondary task-only condition \([100 \times (1–B/A)]\), as in Gosselin and Gagné (2011). For CI listeners, the average divided attention cost on performance on the secondary task in the unaccented and accented conditions was 18.5% \((SEM = 2.4\%)\) and 20.6% \((SEM = 3.5\%)\), respectively. For NH listeners who were presented vocoded speech, the divided attention cost on performance on the secondary task in the unaccented and accented conditions was 24.7% \((SEM = 3.7\%)\) and 21.2% \((SEM = 4.1\%)\), respectively.

An RM-ANOVA was conducted with cost as the dependent variable, group (CI vs. NH vocoder) as the between-subjects factor and talker (unaccented vs. accented) as the within-subjects factor. There were no significant differences overall between groups, \(F(1, 38) = .569, p > .05\), nor between unaccented or accented talkers, \(F(1, 38) = .96, p > .05\), and no interaction between group and talker, \(F(1, 38) = 1.75, p > .05\).

**Discussion**

The goal of this study was to evaluate the effects of visual distraction and an secondary task on a primary speech recognition task by adults who listened to spectrally degraded speech processed through a CI or a vocoder. The overarching hypothesis was that listeners who use CIs are less affected by visual distractions and secondary tasks than NH adults presented vocoded speech. The rationale was that experienced CI listeners have differentially developed the ability to suppress irrelevant information because of their prolonged reliance on visual cues to aid auditory perception with degraded signals. Demands on individual listeners’ limited perceptual resources were increased by (a) presenting the stimuli in a background of multiple talkers to increase energetic and informational masking, (b) presenting signals degraded by CI processing or vocoding to minimize

![Figure 2. ST Performance. Accuracy, in percent correct, on the ST is shown when performed in isolation or while also repeating sentences spoken by an unaccented or an accented talker. Error bars represent ±1 standard error. CI = cochlear implant; NH = normal hearing; ST = secondary task.](image-url)
spectral information, and (c) presenting accented speech to reduce the availability of clear temporal speech cues. Demands on listeners’ cognitive resources were increased by presenting the target talker embedded in a dynamic visual scene to increase extraneous visual information and adding a secondary visual tracking task to the primary auditory task to increase demands on attention.

**Effects of Visual Distraction**

The first hypothesis was that visual distractions would adversely affect listeners who receive spectrally degraded speech, such as through a CI sound processor or a vocoder, because the cognitive capacity required to understand the degraded speech might leave insufficient resources to successfully inhibit distractions. This would result in poorer performance in the AV+D condition than in the AV-only condition. There was no evidence that the presence of the visual distractions in the stimulus videos affected speech recognition for either group (Figure 1, Table 2). This lack of support for the hypothesis was unexpected. It might be explained by three factors: (a) the differences in listener characteristics from previous literature, (b) the more demanding nature of the current task in comparison with previous studies resulting in greater focus and less distraction on the part of the listeners, or (c) the failure of the stimulus videos to be distracting enough to affect speech recognition performance despite their relative complexity.

One previous study that reported detrimental effects of visual distractions on auditory speech recognition in CI listeners (Champoux et al., 2009) only found these effects in relatively new CI listeners, those who had their devices <6 months and were still acclimating to CI-processed sound. The more experienced CI listeners (>6 months post activation) in that study showed no effects of visual distraction. The current study only recruited CI listeners who had >1 year of experience with their device(s). Even though the NH listeners in the current study had a relatively short time to acclimate to vocoded speech, similar to the inexperienced listeners in Champoux et al.’s study, they did not have the prolonged periods of auditory deprivation prior to implantation to which Champoux et al. attributed the distraction effects they observed in their inexperienced listeners. According to this reasoning, the group of NH vocoder listeners in the current study would not be expected to show distraction effects because they did not experience auditory deprivation and subsequent cortical reorganization.

In another study that reported detrimental effects of visual distractions (Cohen & Gordon-Salant, 2017), younger (18–30) and older (65–80) NH listeners were tested on sentences that were not spectrally distorted. That study found poorer performance on speech recognition for both groups when a competing video was present (i.e., a distraction effect). The current study presented sentences in background noise with spectral distortion and an accented talker embedded in a busy scene. The differences in the stimuli and presentation method between the current study and Cohen and Gordon-Salant (2017) may have contributed to the discrepant results. Specifically, the more challenging IEEE sentences spoken in background noise by an accented talker when paired with videos of mundane activity may have caused listeners in the current study to focus attention on the difficult speech recognition task and become less distracted by the concurrent, but irrelevant, video.

This alternative explanation is supported by studies showing that as the primary task difficulty increases, low-level processing of irrelevant stimuli is reduced, as measured by event-related potentials (e.g., Harmony et al., 2000; Simon et al., 2016). The primary speech recognition task in the current study was designed to be difficult (speech presented in background babble heard through a CI or a vocoder) even with the unaccented talker. The absence of perfect speech recognition scores, together with listener anecdotal comments, seems to indicate that listeners found the task to be challenging, though many other factors can play a role in performance (e.g., motivation, fatigue). This task difficulty may have prevented listeners in the current study from showing the same distraction effects as reported in Cohen and Gordon-Salant (2017).

In addition, the current study had more complexity in both the auditory (challenging sentences, multitalker babble, accented talkers) and visual (talkers embedded in, rather than adjacent to, distracting videos) stimuli than the Cohen and Gordon-Salant (2017) study. The stimuli in the current study may have been too perceptually complex to evoke a measurable distraction effect. This explanation is consistent with the predictions of perceptual load theory (Lavie, 1995), which states that in conditions with many complex stimuli in the environment (i.e., high perceptual load), listeners exhaust their limited sensory processing capacity, resulting in reduced perception of irrelevant stimuli (Forster & Lavie, 2007). That is, perceptual load is thought to be related to complexity of the environment rather than to task difficulty (i.e., an increase in cognitive load). The level of stimulus complexity in the current study may have resulted in reduced distraction by irrelevant stimuli.

It is, however, also possible that more engaging visual distractors might have yielded different results in the AV+D condition of the current study. There were, for example, no exotic or unlikely scenes that might have captured a listener’s full attention. Cohen and Gordon-Salant (2017) found visual distraction in a simple sentence-repetition task with videos of everyday
occurrences, but unlike the videos in the current study that were largely the same scene with minor differences, their background video changed from trial to trial. Their participants never knew which background video they would see for each sentence. This unpredictability may have been more distracting than the content of the videos and may provide an explanation for why they saw a distraction effect and the current study did not.

The second hypothesis was that the effect of visual distractions would be greater when task difficulty was increased by presenting the target speech with a foreign accent. This would result in a larger difference between performance in the AV-only and AV+D conditions when the talker had an accent than when the talker was unaccented (i.e., an interaction between talker and AV+D condition). This interaction was not significant (Table 2), reflecting that the difference in listeners’ performance between the AV-only and AV+D conditions did not vary with talker accent (Figure 1). The performance of both groups was, however, significantly affected by the talker’s accent. Speech recognition performance declined an average of 36% for listeners in the CI group and an average of 25% for listeners in the NH vocoder group when listening to accented speech compared with unaccented speech.

The overall effect of accent was expected given the known difficulties of understanding foreign-accented speech previously reported for both CI listeners and NH listeners presented with vocoded speech (e.g., Ji et al., 2014; Waddington et al., 2020). The alterations in speech rhythm and altered pronunciations stemming from the influence of the talker’s native language (Munro & Derwing, 1995) make recognizing foreign-accented speech in quiet a challenging task. CI processing and vocoding, as well as the inclusion of background noise in this study, further challenged listeners by degrading the available auditory cues.

There was a significant interaction between talker accent and listener group, showing that the detrimental effect of a moderate accent was significantly greater for CI listeners than for NH listeners presented with vocoded stimuli. The lack of a significant three-way interaction with condition means that this Group × Talker Accent effect was similar across all conditions. While performance by CI listeners in the accented speech conditions was low, it is unlikely that their performance reflected a floor effect. Each condition offered a chance at 100 keywords, and CI listeners averaged at least 30 keywords correct in each condition. Given that the stimuli were open-set sentences, the probability of chance performance was essentially zero. The CI listeners’ average performance of 30% was therefore clearly above floor performance and likely not the main reason for this significant interaction.

The greater effect of accent on CI listeners than NH listeners is somewhat consistent with previous literature. For example, Ji et al. (2014) showed that CI listeners needed a significantly more favorable SNR than NH listeners with vocoded stimuli to understand whole sentences 50% of the time when the sentences were spoken by an accented talker. However, Ji et al. (2014) did not use AV stimuli or visual distractions, nor were the NH vocoder listeners matched in age with the CI listeners. Therefore, the current results add to the existing literature on the excessive difficulties faced by CI listeners in understanding foreign-accented speech beyond those attributed to the spectral degradation of the device alone (simulated here with vocoded speech). These results may simply reflect the inadequacy of vocoding to accurately replicate CI processing or the fact that listeners who use CIs have other changes in their auditory systems that further differentiate them from NH listeners. On the other hand, the fact that many of the changes in Spanish-accented English are based on temporal alterations and that both CIs and vocoders convey temporal information fairly well means that the persistent deficit in accented-speech recognition by CI listeners may not be entirely due to the differences between vocoding and CI processing.

**Effects of Secondary Task**

The third hypothesis was that the effect of adding a secondary task would be greater than the effect of visual distractions alone. This would be seen as a larger difference between performance in the AV-only and AV+ST conditions than between the AV-only and AV+D conditions. As can be seen in Figure 1, speech recognition declined significantly for both groups when a secondary visual task was used to increase task difficulty relative to the AV-only condition and the AV+D condition. This decrease in performance on the primary task is consistent with the hypothesis and previous literature (Finke et al., 2015; Pashler, 1994; Verhaeghen et al., 2003). The difference in performance between AV+D and AV+ST is likely a result of increasing cognitive demands by adding an secondary task, rather than increased visual distraction, because the same background video was used for both conditions. In the AV+ST condition, listeners had to allocate their attentional resources among the two tasks, either switching between tasks or dividing attention. There was also a significant interaction between the talker’s accent and the AV+ST condition (see Table 2), where the change in speech recognition between AV-only and AV+ST conditions was greater for the unaccented talker than for the accented talker. This could be interpreted to mean that listeners were unable to derive as much benefit from undivided attention (AV-only condition) with the
accented talker than they were with the unaccented talker, possibly due to the difficulties of understanding accented speech.

The fourth and overarching hypothesis was that CI listeners’ speech recognition performance would decrease less than that of their NH peers from the AV-only condition to the AV+D and AV+ST conditions. The rationale was that CI listeners had experience with the spectral distortion of the CI device, using visual cues to aid in speech perception, and suppressing irrelevant information, while NH listeners did not. Careful performance matching and the rapid adaptation of NH listeners to vocoded speech resulted in nearly identical speech recognition performance between both groups at baseline despite their differences in prior experience with spectrally degraded signals. The ability of NH listeners to rapidly adapt to vocoded speech has been shown previously (e.g., Davis et al., 2005) but was not expected to affect strategies for recognition of accented speech in noise and dual-task performance. Despite the performance-matching control methods, or perhaps because of them, this hypothesis was also not supported. There were no significant differences between groups or relevant interactions in the current study, and therefore, there was no evidence of strategies used by CI listeners that were not used by NH listeners presented with vocoded speech.

Previous studies showed no performance differences between CI listeners and NH listeners presented with vocoded speech on the integration of AV cues (Zhou et al., 2019) or on the benefit of visual speech cues (Waddington et al., 2020). The current study extends these previous findings by suggesting that even in conditions with visual distraction and divided attention, there may be no systematic differences between CI listeners and NH vocoder listeners in strategies for dual-task performance and suppression of irrelevant visual stimuli.

It is, however, possible that differences in strategies do exist between groups, or even within groups, but the methods employed were not sensitive enough to reveal those differences. For example, an eye tracker could potentially have revealed differing gaze patterns between listeners. A different secondary task that allowed response time measurements might also have revealed distraction effects not seen in the current study. These possibilities offer opportunities to reveal strategic differences in future studies.

The fifth and final hypothesis was that older listeners’ speech recognition performance would decrease more than that of the younger listeners with the accented talker and in both the AV+D and AV+ST conditions due to the known age-related deficits in processing speed, inhibition, and executive control. This hypothesis was also not supported because age was not a significant predictor of speech recognition performance. The absence of an age effect in this study is surprising, given that CI processing and vocoding greatly reduce spectral information while mostly preserving temporal information about the incoming sound. An age-related deficit was hypothesized because the stimuli in the current study were designed to challenge temporal processing ability and previous studies have found deficits in auditory temporal processing among older adults (e.g., Fitzgibbons & Gordon-Salant, 1996; Gordon-Salant & Fitzgibbons, 1999; Goupell et al., 2017). It is possible that 20 pairs of listeners with a median age of 58 years spread across a range of approximately 60 years do not provide enough statistical power to observe the predicted age effect, but it is more likely that any potential age effect is obscured by the large variation in baseline speech recognition abilities. There was also no significant interaction between age and condition. It might have been expected that older adults would show more behavioral distraction effects than younger adults given the perceptual load theory (Lavie, 1995) and the studies that have shown greater event-related potential amplitudes for unattended stimuli in older adults than in younger adults (e.g., Alain & Woods, 1999). However, age did not seem to affect performance in the current study.

**Effects of Executive Control and Baseline Speech Recognition**

Tests of three cognitive domains were analyzed for their ability to contribute to the model of speech understanding performance in this task. Inhibition and processing speed are known to decrease with age (e.g., Hasher & Zacks, 1988; Salthouse, 1991; Wingfield, 1996), and executive control is thought to be central to performing well on a dual task but is not expected to change with age (e.g., Verhaeghen & Cerella, 2002). The cognitive test that best predicted speech recognition performance in the current study was the measure of executive control (Dimension Change Card Sort). The measures of inhibition (Flanker), processing speed (Pattern Comparison), and the composite score (described in the Data Analysis section) did not contribute as much to the model in comparison. Above-average performance on this executive control task resulted in better performance in the unaccented AV-only condition for the NH listeners (Table 2). The absence of any significant interactions means that a similar relationship between higher scores on the measure of executive control and better speech recognition performance can be extended to CI listeners, performance for all listeners with the accented talker, and performance for all listeners in the other AV conditions. This finding implies that greater ability to control the focus of attention is beneficial for listeners of all ages when performing a dual task, consistent with Verhaeghen & Cerella (2002).
The model also revealed a negative interaction between baseline audio-only speech recognition scores in quiet and speech recognition performance in noise when a secondary task was required (in the AV+ST condition). Listeners with above-average baseline speech recognition in quiet showed greater decreases in performance on the primary speech recognition task in the AV+ST condition compared with the AV-only condition than those with lower baseline scores. This interaction may reflect the need of those with lower audio-only speech recognition scores in quiet to focus their attention more on the visual cues and primary task (speech in noise) in the AV+ST condition than those with better speech recognition in quiet. These results provide additional evidence that it is the limitations of the device, rather than a difference in cognitive abilities or adaptive strategies, that result in poorer speech recognition in real-world environments for CI listeners than their NH peers. Matching age and speech recognition performance between groups removed several of the confounds typical in CI and acoustic listener vocoder studies. Doing so allowed this study to provide evidence against the hypothesized group differences in distractibility. It is still possible that individual listeners were using different strategies or exerting different amounts of effort to accomplish the dual tasks (Gosselin & Gagné, 2011), but these differences in strategy or effort were not revealed in the current protocol. The current results suggest that many of the difficulties experienced by CI listeners when listening to speech in noise with increased cognitive demand can be attributed to the spectral distortion of the signal.

Conclusions
The current study provides evidence that visual distractions irrelevant to the primary auditory task, as presented in the current study, do not affect the ability of listeners to repeat sentences with spectral degradation and background noise. However, the addition of a secondary task, which required listeners to divide their attention between the auditory and visual stimuli, detrimentally affected performance on both the primary auditory and secondary visual tasks. Contrary to expectations, experience with hearing loss and using a CI longer than 1 year did not enhance a listener’s ability to perform two tasks at once compared with NH listeners who had much less exposure to vocoded speech. Advancing age also did not affect the recognition of CI-processed or vocoded speech. Rather, the biggest difference between a group of listeners with CIs and those with NH presented with vocoded speech was the detrimental effect of accent, which was large overall, but greater for the listeners with CIs. This significant deficit merits further probing into the specific mechanisms, be they related to the device or to central auditory processing, that prevent listeners with CIs from understanding speech with altered timing characteristics, such as a Spanish accent.

Author Note
Portions of this work were presented at the 42nd MidWinter meeting of the Association for Research in Otolaryngology in Baltimore, MD in February 2019 and at the Conference on Implantable Auditory Prostheses 2019 in Lake Tahoe, CA in July 2019.

Acknowledgments
We would like to thank Ginny Alexander for helping with participant recruitment and cognitive testing. Ariella Shapiro for helping with stimulus development, and YouTube user, Exploring Alabama, for graciously allowing us to use their video in our stimulus. We would also like to thank Stefanie Kuchinsky for advice on statistical analyses, troubleshooting, and reporting of results. We are also grateful to Brittany Jaekel and Kristina Milvae for advice on statistical approaches and Zilong Xie for helpful comments on the article.

Declaration of Conflicting Interests
The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health under Award Numbers R01AG051603 (M. J. G.) and R01AG009191 (S. G.-S.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

ORCID iDs
Anna R. Tinnemore https://orcid.org/0000-0003-1191-365X
Matthew J. Goupell https://orcid.org/0000-0003-2662-3057

References
Alain, C., & Woods, D. L. (1999). Age-related changes in processing auditory stimuli during visual attention: Evidence for deficits in inhibitory control and sensory memory. *Psychology and Aging, 14*(3), 507. https://doi.org/10.1037/0882-7974.14.3.507
Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology, 4*, 328. https://doi.org/10.3389/fpsyg.2013.00328
Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of
Statistical Software, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Ben-David, B. M., Eidels, A., & Donkin, C. (2014). Effects of aging and distractors on detection of redundant visual targets and capacity: Do older adults integrate visual targets differently than younger adults? PLoS One, 9(12), e113551. https://doi.org/10.1371/journal.pone.0113551

Bent, T., Bradlow, A. R., & Smith, B. L. (2008). Production Tinnemore et al. 15

Beynon, G., & Munro, K. (1993). A discussion of current Carlozzi, N., Tulsky, D., Chiaravalloti, N., Beaumont, J., Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. The Journal of the Acoustical Society of America, 109(3), 1101–1109. https://doi.org/10.1121/1.1345696

Carlozzi, N., Tulsky, D., Chiavarotti, N., Beaumont, J., Weintraub, S., Conway, K., & Gershon, R. (2014). NIH Toolbox Cognitive Battery (NIHTB-GB): The NIHTB pattern comparison processing speed test. Journal of the International Neuropsychological Society, 20(6), 630–641. https://doi.org/10.1017/S1355617714000319

Champoux, F., Lepore, F., Gagné, J. P., & Théoret, H. (2009). Visual stimuli can impair auditory processing in cochlear implant users. Neuropsychologia, 47(1), 17–22. https://doi.org/10.1016/j.neuropsychologia.2008.08.028

Cohen, J. I., & Gordon-Salant, S. (2017). The effect of visual distraction on auditory-visual speech perception by younger and older listeners. The Journal of the Acoustical Society of America, 141(5), EL470–EL476. https://doi.org/10.1121/1.4983399

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. Journal of Experimental Psychology: General, 134(2), 222. https://doi.org/10.1037/0096-3445.134.2.222

Dupuis, K., Pichora-Fuller, M. K., Chasteen, A. L., Marchuk, V., Singh, G., & Smith, S. L. (2015). Effects of hearing and vision impairments on the Montreal Cognitive Assessment. Aging, Neuropsychology, and Cognition, 22(4), 413–437. https://doi.org/10.1080/13828158.2014.968084

Durlach, N. I., Mason, C. R., Kidd, G. Jr., Arborgast, T. L., Colburn, H. S., & Shinn-Cunningham, B. G. (2003). Note perceptual learning of distorted speech: Evidence from the Journal of Experimental Psychology: Human Perception and Performance, 29(5), 1868–1884. https://doi.org/10.1037/0096-1523.29.5.1868

Dye, M. W., Hauser, P. C., & Bavelier, D. (2009). Is visual selective attention in deaf individuals enhanced or deficient? The case of the useful field of view. PLoS One, 4(5), e5640. https://doi.org/10.1371/journal.pone.0005640

Exploring Alabama. (2017, July 4). You cannot stop watching this: San Antonio river walk people watching [Video File]. https://www.youtube.com/watch?v=cOgIGCZbVzK& t=1s

Finke, M., Sandmann, P., Kopp, B., Lenarz, T., & Büchner, A. (2015). Auditory distraction transmitted by a cochlear implant alters allocation of attentional resources. Frontiers in Neuroscience, 9, 68. https://doi.org/10.3389/fnins.2015.00068

Fitzgibbons, P. J., & Gordon-Salant, S. (1996). Auditory temporal processing in elderly listeners. Journal of the American Academy of Audiology, 7, 183–189. https://www.audiology.org/sites/default/files/journal/AAA_07_03_07.pdf

Forster, S., & Lavie, N. (2007). High perceptual load makes everybody equal. Psychological Science, 18(5), 377–381. https://doi.org/10.1111/j.1467-9280.2007.01908.x

Friedland, D. R., Runge-Samuelsen, C., Baig, H., & Jensen, I. (2010). Case-control analysis of cochlear implant performance in elderly patients. Archives of Otolaryngology–Head & Neck Surgery, 136(5), 432–438. https://doi.org/10.1001/archoto.2010.57

Fu, Q. J., & Nagaki, G. (2005). Noise susceptibility of cochlear implant users: The role of spectral resolution and smearing. Journal of the Association for Research in Otolaryngology: JARO, 6(1), 19–27. https://doi.org/10.1007/s10162-004-5024-3

Gifford, R. H., Shallop, J. K., & Peterson, A. M. (2008). Speech recognition materials and ceiling effects: Considerations for cochlear implant programs. Audiology and Neurotology, 13, 193–205. https://doi.org/10.1159/000113510

Goman, A. M., & Lin, F. R. (2016). Prevalence of hearing loss by severity in the United States. American Journal of Public Health, 106(10), 1820–1822. https://doi.org/10.2105/AJPH.2016.303299

Gordon-Salant, S., & Fitzgibbons, P. J. (1999). Profile of auditory temporal processing in older listeners. Journal of Speech, Language, and Hearing Research, 42(2), 300–311. https://doi.org/10.1044/jslhr.4202.300

Gordon-Salant, S., Yeni-Komshian, G. H., & Fitzgibbons, P. J. (2010). Recognition of accented English in quiet and noise by younger and older listeners. The Journal of the Acoustical Society of America, 128(5), 3152–3160. https://doi.org/10.1121/1.3495940

Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., & Barrett, J. (2006). Age-related differences in identification and discrimination of temporal cues in speech segments. The Journal of the Acoustical Society of America, 119(4), 2455–2466. https://doi.org/10.1121/1.2171527

Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., Cohen, J. I., & Waldroup, C. (2013). Recognition of accented and unaccented speech in different maskers by younger and older listeners. Journal of the Acoustical Society of America, 134, 618–627. https://doi.org/10.1121/1.4807817

Gosselin, P. A., & Gagné, J. P. (2011). Older adults expend more listening effort than young adults recognizing speech in noise. Journal of Speech, Language, and Hearing Research, 54(3), 944–958. https://doi.org/10.1044/1092-4388(2010/10-0069)

Goupell, M. J., Gaskins, C. R., Shader, M. J., Walter, E. P., Anderson, S., & Gordon-Salant, S. (2017). Age-related differences in the processing of temporal envelope and spectral cues in a speech segment. Ear and Hearing, 38(6),
Harmony, T., Bernal, J., Fernández, T., Silva-Pereyra, J., Fernandez-Bouzas, A., Marosi, E., Rodríguez, M., & Reyes, A. (2000). Primary task demands modulate P3a amplitude. *Cognitive Brain Research, 9*(1), 53–60. https://doi.org/10.1016/S0926-6410(99)00044-0

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory, 22*, 193–225. Academic Press. https://doi.org/10.1016/S0065-268X(08)60041-9

Hochberg, I., Boothroyd, A., Weiss, M., & Hellman, S. (1992). *Harmony, T., Bernal, J., Fernández, T., Silva-Pereyra, J., Hochberg, I., Boothroyd, A., Weiss, M., & Hellman, S. (1992).* Effects of noise and noise suppression on speech perception by cochlear implant users. *Ear and Hearing, 13*(4), 263–271. https://doi.org/10.1097/00003446-199208000-00008

Institute of Electrical and Electronics Engineers. (1969). *IEEE Transactions on Audio and Electroacoustics, 17*(3), 225–246. https://doi.org/10.1109/TAU.1969.1162058

Ito, H., Uno, H., Atsumi, B., & Akamatsu, M. (2001). Visual distraction while driving: Trends in research and standardization. *IATSS Research, 25*(2), 20–28. https://doi.org/10.1016/S0386-1112(14)60067-4

Ji, C., Galvin, J. J., Chang, Y. P., Xu, A., & Fu, Q. J. (2014). Perception of speech produced by native and nonnative talkers by listeners with normal hearing and listeners with cochlear implants. *Journal of Speech, Language, and Hearing Research, 57*(2), 532–554. https://doi.org/10.1044/2014_JSLSHR-I-12-0404

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Prentice-Hall.

Kausler, D. H., & Hakami, M. K. (1982). Frequency judgments by young and elderly adults for relevant stimuli with simultaneously present irrelevant stimuli. *Journal of Gerontology, 37*(4), 438–442. https://doi.org/10.1093/geronj/37.4.438

Lavie, N. (1995). Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance, 21*(3), 451-468. https://doi.org/10.1037/0096-1523.21.3.451

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning, 45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

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. *Journal of the American Geriatrics Society, 53*(4), 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Nelson, P. B., Jin, S. H., Carney, A. E., & Nelson, D. A. (2003). Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America, 113*(2), 961–968. https://doi.org/10.1121/1.1531983

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin, 116*(2), 220–244. https://doi.org/10.1037/0033-2909.116.2.220

Pearson, J. D., Morrell, C. H., Gordon-Salant, S., Brant, L. J., Metter, E. J., Klein, L. L., & Fozard, J. L. (1995). Gender differences in a longitudinal study of age-associated hearing loss. *The Journal of the Acoustical Society of America, 97*(2), 1196–1205. https://doi.org/10.1121/1.142231

Pichora-Fuller, M. K. (2003). Processing speed and timing in aging adults: Psychoacoustics, speech perception, and comprehension. *International Journal of Audiology, 42*(sup1), 59–67. https://doi.org/10.3109/14992020309074625

R Core Team. (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of Neurophysiology, 89*(2), 1078–1093. https://doi.org/10.1152/jn.00706.2002

Rees, G., Frith, C., & Lavie, N. (2001). Processing of irrelevant visual motion during performance of an auditory attention task. *Neuropsychologia, 39*(9), 937–949. https://doi.org/10.1016/S0028-3932(01)00016-1

Saltheouse, T. A. (1991). Mediation of adult age differences in cognition by reductions in working memory and speed of processing. *Psychological Science, 2*(3), 179–183. https://doi.org/10.1111/j.1467-9280.1991.tb00127.x

Simon, S. S., Tusch, E. S., Holcomb, P. J., & Daffner, K. R. (2016). Increasing working memory load reduces processing of cross-modal task-irrelevant stimuli even after controlling for task difficulty and executive capacity. *Frontiers in Human Neuroscience, 10*, 380. https://doi.org/10.3389/fnhum.2016.00380

Sladen, D. P., & Zappler, A. (2015). Older and younger adult cochlear implant users: Speech recognition in quiet and noise, quality of life, and music perception. *American Journal of Audiology, 24*(1), 31–39. https://doi.org/10.1044/2014_AJA-13-0066

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing, 26*(3), 263–275. https://doi.org/10.1097/00003446-200506000-00003

Strelnikov, K., Rouger, J., Demonet, J. F., Lagleyre, S., Fraysse, B., Deguine, O., & Barone, P. (2013). Visual activity predicts auditory recovery from deafness after adult cochlear implantation. *Brain, 136*(12), 3682–3695. https://doi.org/10.1093/brain/awt274

Tan, P. A., O’Kane, G., & Wingfield, A. (2002). Distraction by competing speech in young and older adult listeners. *Psychology and Aging, 17*(3), 453–467. https://doi.org/10.1037/0885-2974.17.3.453

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Audiovisual integration and lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing, 28*(5), 656–668. https://doi.org/10.1097/AUD.0b013e3181f27185

Verhaeghen, P., & Cerella, J. (2002). Aging, executive control, and attention: A review of meta-analyses. *Neuroscience & Biobehavioral Reviews, 26*(7), 849–857. https://doi.org/10.1016/S0149-7634(02)00071-4

Verhaeghen, P., Steitz, D. W., Sliwinski, M. J., & Cerella, J. (2003). Aging and dual-task performance: a meta-analysis.
Psychology and Aging, 18(3), 443–460. https://doi.org/10.1037/0882-7974.18.3.443

Voeten, C. (2020). buildmer: Stepwise elimination and term reordering for mixed-effects regression. R package version 1.5. https://CRAN.R-project.org/package=buildmer

Waddington, E., Jaekel, B. N., Tinnemore, A. R., Gordon-Salant, S., & Goupell, M. J. (2020). Recognition of accented speech by cochlear-implant listeners: Benefit of audiovisual cues. Ear and Hearing, Advance online publication. https://doi.org/10.1097/AUD.0000000000000842

Walden, B. E., Busacco, D. A., & Montgomery, A. A. (1993). Benefit from visual cues in auditory-visual speech recognition by middle-aged and elderly persons. Journal of Speech and Hearing Research, 36(2), 431–436. https://doi.org/10.1044/jshr.3602.431

Wang, J., Zhu, Y., Chen, Y., Mamat, A., Yu, M., Zhang, J., & Dang, J. (2020). An eye-tracking study on audiovisual speech perception strategies adopted by normal-hearing and deaf adults under different language familiarities. Journal of Speech, Language, and Hearing Research, 63(7), 2245–2254. https://doi.org/10.1044/2020_JSLHR-19-00223

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. Perception & Psychophysics, 63(8), 1293–1313. https://doi.org/10.3758/BF03194544

Wingfield, A. (1996). Cognitive factors in auditory performance: Context, speed of processing, and constraints of memory. Journal of the American Academy of Audiology, 7(3), 175–182. https://www.audiology.org/sites/default/files/journal/JAAA_07_03_06.pdf

Yi, H. G., Phelps, J. E., Smiljanic, R., & Chandrasekaran, B. (2013). Reduced efficiency of audiovisual integration for nonnative speech. Journal of the Acoustical Society of America, 134(5), EL387–EL393. https://doi.org/10.1121/1.4822320

Zelazo, P., Anderson, J., Richler, J., Wallner-Allen, K., Beaumont, J., Conway, K., Gershon, R., & Weintraub, S. (2014). NIH Toolbox Cognition Battery (CB): Validation of executive function measures in adults. Journal of the International Neuropsychological Society, 20(6), 620–629. https://doi.org/10.1017/S1355617714000472

Zhou, X., Innes-Brown, H., & McKay, C. M. (2019). Audiovisual integration in cochlear implant listeners and the effect of age difference. The Journal of the Acoustical Society of America, 146(6), 4144–4154. https://doi.org/10.1121/1.5134783