Word Learning in Deaf Adults Who Use Cochlear Implants: The Role of Talker Variability and Attention to the Mouth

Jasenia Hartman,[1,2] Jenny Saffran,[3] and Ruth Litovsky[1,4]

Objectives: Although cochlear implants (CIs) facilitate spoken language acquisition, many CI listeners experience difficulty learning new words. Studies have shown that highly variable stimulus input and audiovisual cues improve speech perception in CI listeners. However, less is known whether these two factors improve perception in a word learning context. Furthermore, few studies have examined how CI listeners direct their gaze to efficiently capture visual information available on a talker's face. The purpose of this study was two-fold: (1) to examine whether talker variability could improve word learning in CI listeners and (2) to examine how CI listeners direct their gaze while viewing a talker speak.

Design: Eighteen adults with CIs and 10 adults with normal hearing (NH) learned eight novel word-object pairs spoken by a single talker or six different talkers (multiple talkers). The word learning task comprised of nonsense words following the phonotactic rules of English. Learning was probed using a novel talker in a two-alternative forced-choice eye gaze task. Learners' eye movements to the mouth and the target object (accuracy) were tracked over time.

Results: Both groups performed near ceiling during the test phase, regardless of whether they learned from the same talker or different talkers. However, compared to listeners with NH, CI listeners directed their gaze significantly more to the talker's mouth while learning the words.

Conclusions: Unlike NH listeners who can successfully learn words without focusing on the talker's mouth, CI listeners tended to direct their gaze to the talker's mouth, which may facilitate learning. This finding is consistent with the hypothesis that CI listeners use a visual processing strategy that efficiently captures redundant audiovisual speech cues available at the mouth. Due to ceiling effects, however, it is unclear whether talker variability facilitated word learning for adult CI listeners, an issue that should be addressed in future work using more difficult listening conditions.

Key words: Audiovisual speech, Deaf adults with cochlear implants, Eye gaze, Face scanning, Talker variability, Word learning.

(Ear & Hearing 2023;XX;00–00)

INTRODUCTION

To acquire spoken words, learners must be able to accurately perceive the speech sounds that make up the word. Additionally, learners can utilize visual speech cues found on the talker's face to facilitate learning. For visually unimpaired people who are deaf, cochlear implants (CIs) not only grant listeners access to the auditory world but also allow them to combine visual cues with the auditory signal that is provided by the CI. Indeed, while CI recipients can perceive speech with solely auditory input (for review, see Dorman et al. 2002; Shannon 2002), they often misperceive speech sounds due to the degraded auditory input transmitted through the CI (Munson et al. 2003; Munson & Nelson 2005). Notably, listeners with CIs show improvements in speech intelligibility with the addition of visual input and tend to rely heavily on visual cues (Rouger et al. 2008; Tremblay et al. 2010; Stevenson et al. 2017).

While prior studies have highlighted the limitations of speech perception and reliance on visual speech cues in CI listeners, relatively few studies have examined these issues within the purview of spoken language learning. In particular, research on spoken word learning in CI listeners has not examined factors that might improve word learning. Relatedly, studies on speech perception in CI listeners have demonstrated the general benefits of audiovisual speech cues, but have not examined what portions of the face attract visual attention during word learning.

The current study aims to address two questions. First, would introducing variability into the acoustic input enhance word learning in CI listeners? Second, where on the talker's face do CI listeners look while learning new words? Given the high variability in outcomes and success of use with CIs in real-world listening environments, our broader goal is to understand the factors that may influence successful word learning (talker variability, audiovisual speech) in adults with CIs.

Word Learning in CI Listeners

Word learning is a core spoken language skill that consists of a complex array of cognitive and perceptual processes, including phonetic sensitivity, or access to fine phonetic details of the word forms. For people who are deaf, CIs allow listeners to develop phonetic categories and acquire spoken words. Despite these improvements, some CI listeners face challenges learning new words. One contributing factor to these difficulties is the CI processing strategy. Whereas the normal-hearing (NH) system consists of dozens of independent auditory filters, the CI system has up to 22 electrodes, with approximately eight independent channels stimulated at any time. As a result, CI listeners receive limited spectral information. Additionally, patient-specific factors, such as later implantation and less CI experience, lead to poorer word learning outcomes (Houston & Miyamoto 2010; Houston et al. 2012; Havy et al. 2013; Pimperton & Walker 2018). Even postlingually deafened CI adults show wide variability in language skills (see Peterson et al. 2010 for review). Finally, listeners' perceptual abilities in identifying speech sounds impact their ability to learn words. The current study focused on improving listeners' ability to perceive the speech sounds found within the word form.

CI listeners often misperceive speech sounds (Munson et al. 2003; Giezen et al. 2010), likely because of the degraded nature

[1]Neuroscience Training Program, University of Wisconsin-Madison, Madison, Wisconsin, USA; [2]Department of Psychology, Harvard University, Cambridge, MA, USA; [3]Department of Psychology, University of Wisconsin-Madison, Madison, Wisconsin, USA; and [4]Department of Communication and Science Disorders, University of Wisconsin-Madison, Madison, Wisconsin, USA.

<zdoi; 10.1097/AUD.0000000000001432>

of the spectral information in speech sounds (Lane et al. 2007; Winn & Litovsky 2015). In a recent analysis of spectral-temporal cues delivered through the clinical speech processors, Peng et al. (2019) found that pulsatile stimulation patterns may not provide the cue saliency needed for listeners with CIs to achieve the same level of accuracy in discriminating speech sounds as listeners with NH. CI listeners are thus more likely to exhibit less developed phonetic categories compared to NH listeners. Whereas listeners with NH show sharp phonetic categories, listeners with CIs show broad categories with shifted boundaries (Iverson 2003; Munson & Nelson 2005b; Desai et al. 2008). For this reason, the detail of word forms might be difficult to process and encode, thereby posing a challenge for word learning.

Prior word learning studies in listeners with CIs have focused primarily on children (Davidson et al. 2014; Quittner et al. 2016; Walker & McGregor 2013). One study found that 3- to 6-year-olds with CIs experienced more difficulty learning labels for objects that differed by a single phonetic feature than by multiple features (Havy et al. 2013). Similarly, 5- to 6-year-olds with CIs were more successful learning novel labels that exemplified acoustically salient contrasts, such as vowels, than perceptually difficult contrast, such as consonants (Giezen et al. 2016).

These findings underscore the challenges that CI listeners experience in acquiring new words, which may be due to their difficulties in discriminating between phonetic categories relative to NH children (e.g., Peng et al. 2019). However, one limitation of the methods used in these studies is that CI listeners are typically exposed to novel words by hearing the same speaker label the words. This approach might exacerbate the perceptual challenges faced by CI listeners by distorting listeners' perceptual space toward noncontrastive acoustic dimensions of the words to be learned. Variation within the acoustic signal might facilitate word learning by helping listeners to determine which acoustic dimensions are helpful for contrasting lexical items.

The Role of Variability in Learning

Studies with NH listeners suggest that variability plays an essential role in learning categories (Gómez 2002; Perry et al. 2010; Posner & Keele 1968), and, of particular relevance to word learning, in augmenting phonetic categories (Lively et al. 1993; Rost & McMurray 2009, 2010; Quam et al. 2017). For instance, Rost and McMurray (2009) examined the role of acoustic variability in learning phonologically similar words (e.g., /puk/ &/buk/). Infants were taught two novel word-object pairs spoken either by a single talker or by 18 different talkers. Whereas infants failed to learn the word-object pairs when both words were spoken by a single talker, they were successful when the words were spoken by multiple talkers. In a follow-up study, Rost and McMurray (2010) introduced variability along the contrastive cue (in this case, voicing for /puk/ and/buk/) while holding noncontrastive cues (talker and prosody) constant. In this condition, infants were unable to distinguish phonologically similar words. However, when the contrastive cue was held constant and the noncontrastive cue varied, learning was successful. The benefits of variability in learning also extend to adults. Lively et al (1993) found that Japanese native speakers learning English as a second language were able to form robust phonetic categories of the nonnative /r/-/l/ contrasts after learning from different talkers but not the same talker. Moreover, the authors found that variability allowed participants to generalize to new speakers.

The aforementioned studies highlight two benefits of talker variability in word learning. First, talker variability allows listeners to encode multiple exemplars of the lexical or phonemic categories. Different talkers pronounce the same word differently. Through variability, listeners are able to utilize these differences to organize their perceptual categories. Second, talker variability helps listeners to weigh the importance of different acoustic cues in distinguishing lexical categories. By introducing variation along the noncontrastive cues, such as prosody, listeners are able to learn that prosody is an irrelevant cue. In contrast, the relative invariance along the contrastive cue helps learners to realize the importance of such cue in contrasting the words. These two processes allow listeners to generalize to phonetic categories spoken by novel talkers. While talker variability has been shown to drive word learning in NH listeners, less is known about whether CI listeners would benefit from variability within the acoustic environment.

Notably, two studies have demonstrated that high variability training improves perceptual categorization in CI listeners (Miller et al. 2016; Zhang et al. 2021). Miller et al. (2016) examined the efficacy of high variability training on CI adults who were postlingually deafened. A group of nine CI adults were trained on consonant-vowel syllables spoken by multiple talkers and were then tested on their phonetic categorization of those syllables. The authors found that listeners who received high variability training exhibited sharper phonetic categories. Similarly, Zhang et al. (2021) found that high variability training improved tone perception in Mandarin-speaking CI children who were prelingually deafened. Moreover, the children in the high variability training group were able to generalize to tones produced by novel talkers. These findings are encouraging because they show that talker variability is able to induce tuning of CI listeners' perceptual categories. However, these studies only presented isolated syllables and not word forms. Examining the benefit of talker variability in a word learning context offers an additional dimension to auditory processing because listeners must be able to encode the speech form and retain it to associate labels with objects. Thus, one goal of the current study is to examine whether the benefits of talker variability extend beyond phonetic categorization to word learning in CI listeners, such that performance is better in conditions with variability than in conditions without variability.

Audiovisual Speech Processing

Most word learning studies with CI and NH listeners provide solely auditory input. However, in real-life situations, word learning is typically an audiovisual process that occurs primarily during face-to-face interactions. Moreover, the talker's face contains highly informative information that has been shown to support learning. The mouth is the primary source of phonetic information and moves in alignment with the audio signal. The eyes provide information about the social identity and referential intention of the talker. Given that the face contains a rich source of information, listeners must utilize a strategy to efficiently gather information.

For NH listeners, both the eyes and mouth attract the bulk of attention as listeners view a talker speak. However, many factors, such as the listener's age or the nature of the task, influence

which facial region listeners focus on. For instance, during the first year of life, infants shift their attention from the talker's eyes to the talker's mouth as they are faced with the challenge of acquiring their native language (Lewkowicz & Hansen-Tift 2012; Tenenbaum et al. 2013; Hillairet De Boisferon et al. 2018; Tsang et al. 2018), or when facing bilingual input (Birulés et al. 2019). Additionally, adult NH listeners focus on the talker's mouth or nose while listening to speech in noise (Munhall et al. 1998; Buchan et al. 2008; Król 2018; Lasing & McConkie 1999) or hearing sentences spoken by different talkers (Buchan et al. 2008). Although numerous studies with NH listeners have addressed where listeners focus while viewing a talker speak, relatively few studies have examined this question in CI listeners.

Studies using incongruent audiovisual speech stimuli provide insights about the facial regions that listeners with CIs attend to while viewing talking faces (e.g., Desai et al. 2008; Winn et al. 2013). When presented with McGurk stimuli (e.g., hearing/ba/ but seeing/ga/; McGurk & McDonald 1957) listeners with CIs tend to bias their response toward the visual domain, reporting a percept that corresponds to the visual input, whereas listeners with NH tend to report an illusory fused percept (e.g., /da/) or bias their response towards the auditory domain (Rouger et al. 2008; Tremblay et al. 2010). Later age of implantation and less experience with CIs are associated with greater bias toward the visual domain (Desai et al. 2008; Tremblay et al. 2010). This converging evidence indicates that CI listeners heavily rely on speech information coming from the visual domain, often weighing it more strongly than information coming from the auditory domain. However, these studies only provide indirect measures of audiovisual speech processing in CI listeners and do not interrogate the ways in which listeners direct their gaze to gather visual information. Examining CI listeners' preference for a particular facial region during word learning has the potential to provide insight into how listeners direct their gaze to support their learning. Thus, another goal of the current study was to examine CI listeners' visual attention to the talker's face.

Present Study

The purpose of the present study was to address two questions: (1) Does talker variability improve word learning in adults with CIs? (2) Which facial region of the talkers' faces do listeners with CIs focus on while learning new words? To address these questions, we exposed adults with CIs and NH to novel word-object pairings. During training, listeners heard and saw the same person (single speaker) or six different people (mixed gender) label the novel objects (within-subject design). Listeners were then tested on their word learning using items produced by a novel talker, in order to assess generalization. Throughout the learning and test phases, we tracked eye movements to obtain a fine-grained measure of language and audiovisual processing. In particular, we obtained a moment-by-moment assessment of listeners' attention to a particular region of a talker's face as well as their accuracy during the test of word learning. Although word learning is typically studied in children, we chose to focus on adults because CI adults also experience challenges in correctly perceiving speech sounds that may impact their ability to acquire words. Moreover, given the promising results from prior studies on the efficacy of high variability training on speech perception, we wanted to examine if talker variability would improve word learning for adults with CIs.

Talker Variability • We hypothesized that if CI listeners can capitalize on variability within the acoustic environment, then learning from multiple talkers would help listeners to determine which acoustic dimensions are relevant for distinguishing the words to be learned. Thus, talker variability would improve word learning test performance. However, if CI listeners are unable to detect variability, then it might not influence word learning. Because CI listeners often confuse similar-sounding words, we also manipulated the similarity of the words forms to examine whether variability might boost performance more when distinguishing minimal pairs compared to distinct pairs. In addition, we expected that overall performance on the word learning test would be worse for listeners with CIs compared to listeners with NH.

Audiovisual Processing • We expected that if CI listeners rely heavily on visual cues, then they might direct their gaze to the talker's mouth. Moreover, focusing on the mouth would suggest that CI listeners utilize a visual processing strategy that allows them to efficiently extract phonetic information. Alternatively, CI listeners may engage in eye gaze behavior similar to that of adult NH listeners, such that direct attention to the mouth is not required for accurate speech perception. Given that audiovisual information improves encoding of the auditory signal, we asked whether listeners who attended more to the talker's mouth during learning were more accurate at identifying the target object during the test of word learning. Finally, because talker variability was manipulated in the learning materials, we examined the interaction between talker variability and audiovisual speech processing. We predicted that listeners' fixation to the mouth might be modulated by talker variability: listeners would attend more to a talker's mouth when the talker varies than when it remains constant, consistent with previous findings (Buchan et al. 2008).

MATERIALS AND METHODS

Participants

Nineteen adult CI listeners (mean age: 57.3; range: 20–74) participated in the study (see Table 1). All were monolingual English speakers with at least 1 year of CI experience. This group consisted of 16 bilateral CI users, one unilateral CI user, and one hybrid CI user (acoustic + electric hearing in the CI ear). All participants had Cochlear Ltd CIs (Sydney, Australia). CI listeners were excluded from analysis for inability to track eye movement (n = 1) or for contributing less than the minimal number of learning trials (n = 2). Twelve NH adults (mean age: 60.2; range: 48–70) also participated. Due to COVID-19, we were unable to recruit additional NH participants. NH was indicated as audiometric thresholds of 25 dB for octaves between 250 and 3000 Hz and no greater than 40 dB at 4000 Hz (ANSI 1989). NH listeners were excluded from analysis due to failure in passing the hearing screening (n = 2) or for not contributing to the minimal number of learning trials (n =2). The final sample sizes were 16 CI and eight NH adults.

All participants had normal or corrected vision and achieved a typical score of 26 or above (Nasreddine et al. 2005; Goupell et al. 2017) on the Montreal Cognitive Assessment test for cognitive function, except for one CI participant who obtained a score of 25. However, we included this participant

TABLE 1. Demographics of CI participants

ID	Sex	Age	Onset of deafness (yrs)	Duration of HL (yrs)	Years with First CI	Years with Second CI	Device	Etiology	CNC Word Score (%)
ICP	M	56	4	42	10	7	Nucleus 7	Unknown	32
IAU	M	70	3	46	21	14	Nucleus 6	Unknown	53.1
ICI	F	61	46	4	11	10	Nucleus 6	Unknown	54
IAJ	F	73	12	38	23	16	L: Nucleus 6; R: Kanso	Unknown	70
ICJ	F	70	25	35	10	10	Nucleus 6	Hereditary	70
IDM	F	42	5	28	9	7	Nucleus 7	Unknown	74
IDL*	F	65	33	28	4	3.5	Nucleus 6	Unknown	74
ICY	M	66			4	4	—	—	74
IDJ	F	58	45	8	5	5	Nucleus 6	—	76
IBZ	F	52	38	1	14	12	Nucleus 6	Unknown	82
ICC	F	74	9	52	13	11	Nucleus 7	Congenital Progressive	82
IDA	F	52	8	38	6	5	Nucleus 6	Unknown	84
IBF	F	66	38	16	14	12	Nucleus 7	Hereditary	84
IDK†	M	64	16	34	14	—	Nucleus 6	Otosclerosis	88
ICM	F	63	23	34	9	7	Nucleus 6	Unknown	88
IDH	M	20	3	1.5	15	14	Nucleus 6	Unknown	96

*Hybrid (acoustic + electric hearing in CI ear) CI participant.
†Unilateral CI participant.
CI, cochlear implant; CNC, consonant-nucleus-consonant; HL, hearing level.

in the analysis due to the fact that she may have failed due to her hearing impairment (Dupuis et al. 2015). The experiments were approved by the local IRB. All participants were paid for their participation.

Visual and Auditory Stimuli

Eight novel objects (see Fig. 1) were selected from the NOUN database (Horst & Hout 2016). Each image was presented in high resolution (600 DPI) on a white background and aligned horizontally on a 19″ computer screen.

Speech stimuli consisted of eight novel words: /ditɑ/, /gitɑ/, /fomɑ/, /vomɑ/, /nodi/, /lodi/, /pibu/, and/tibu/. Words were selected and modified from the NOUN database (Horst & Hout 2016) and followed the phonotactic constraints of English. Each novel word was spoken in isolation by six native English speakers (four males, four females) raised in the Midwest. Multiple tokens of each word were recorded by every speaker.

Audio/visual speech stimuli were videorecorded with an iPad Air Pro (30 frames/s), resolution of 1920 × 1080. Each talker was filmed against a solid background. A microphone was placed 8 inches away from the talker to record the audio (44.1 kHz sampling rate). Audio recorded from the microphone was processed using Adobe Audition and replaced the original audio recorded from the iPad Air Pro. Four hundred milliseconds of silence were added before the onset of the word and 300 ms of silence was added after the offset of the word. Videos were edited with Adobe Premiere so that only the head and shoulders of the talker were visible. For every speaker, a single video of each word was selected based on the quality of the video (e.g., low-to-moderate eye blinking) and audio. Since the goal of the experiment was to expose CI listeners to acoustic variability, words were not matched for acoustic properties (e.g., duration, pitch contour) across talkers. The audio was synchronized to videos using Adobe Premiere. The mean length of the video was 1015 ms (range: 1000–1027 ms). The audio was scaled to 55 dB on a A-weighting scale using a sound level meter.

Word-Object Pairs

Each novel word was paired with a novel object (eight word-object pairings; see Fig. 1). There were two sets of four novel-word-object pairings. Set 1 consisted of the items /ditɑ/, /gitɑ/, /fomɑ/, and/vomɑ/. Set 2 consisted of the items /nodi/, /lodi/, /pibu/, and/tibu/. Each set was assigned to a learning condition (single versus multiple talkers), counterbalanced across participants. We chose to create two sets of four words to allow for within-subject study design and for our manipulation of test difficulty (see Procedure). Given the heterogeneity of cochlear implant listeners, a within-subject study design allows listeners to serve as their own control.

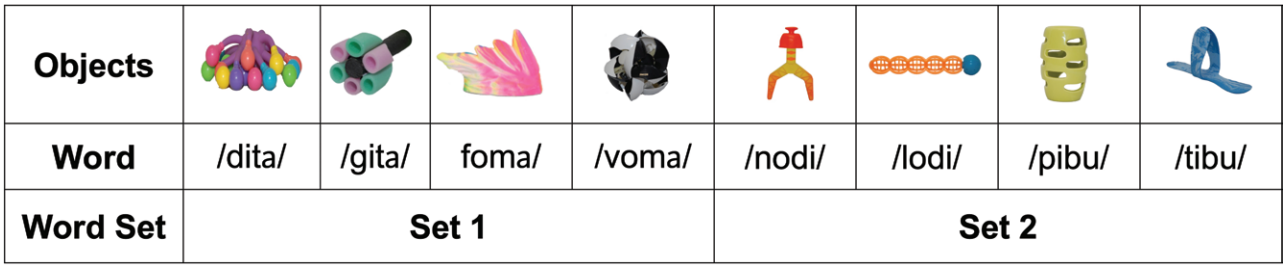

Fig. 1. The eight novel word-object pairings used. Each word set was counterbalanced across learning condition (single vs. multiple talker).

Apparatus

Participants were tested in a double-walled sound booth (Acoustic System, TX). Participants sat at a table with a 19-inch LCD monitor (1280 × 1240 pixels). Eye gaze was tracked with the EyeLink SR 1000 eye-tracker (SR Research, Kanata, ON, Canada) at a sampling rate of 1000 Hz. A chin rest was used to maintain the distance of the head to the monitor and to restrict head movement. A Babyface sound card delivered the audio signal to a speaker positioned at the front of the room. Audiovisual stimuli were presented using custom software written in MATLAB (Mathworks, Natick, MA). The Psychophysics Toolbox (v3.0.14; Brainard 1997; Pelli 1997; Kleiner et al. 2007) was used to maintain the synchronization of audiovisual stimulus presentation with an eye-tracking camera.

Procedure

Participants were seated 1 m from the computer screen. At the beginning of the experiment, the eye-tracker was calibrated by asking participants to look at nine different locations on the screen. After calibration, participants entered the learning phase, which consisted of two within-subjects conditions, a single-talker and a multiple-talker learning condition. Participants completed a learning phase followed by the test phase for one condition, and then the learning phase followed by the test phase for the other condition (order counterbalanced across participants). For the single-talker learning condition, participants were exposed to the novel word-object pairings spoken by a single male talker. In the multiple-talker learning condition, participants were exposed to the novel word-pairings spoken by six different talkers (three males, three females). Participants were instructed to try to learn the names of each object and to move their eyes freely. The learning phase began with the novel object appearing at the bottom left or bottom right of the screen. After 2000 ms, a video of the talker appeared in the center of the screen. There was approximately 400 ms of silence before the talker spoke the novel word in isolation (e.g., "tibu"). Following the offset of speech, the video and object image remained on the screen for an additional 1000 ms before disappearing. In the single-talker condition, each of the four objects were labeled six times, by a single speaker, for a total of 24 trials. For the multiple-talker condition, each of the six talkers labeled each of the four objects once, for a total of 24 trials. The labeling of each object was uniformly distributed across the learning phase to avoid all six presentations of a word-object pair from occurring at only one segment of the learning phase.

A test phase immediately followed each learning phase (see Fig. 2). Test trials consisted of two difficulty levels, Easy and Hard trials. Easy trials were defined as target and distractor labels that differed by several speech sounds (e.g., /ditɑ/ versus /vomɑ). Hard trials were defined as target and distractor labels that served as minimal pairs (e.g., /ditɑ/ versus /gitɑ/). Minimal pairs always differed in the onset consonant.

On each test trial, participants saw two objects at the bottom of the screen, one on each side. One object served as the target, whereas the other object served as the distractor. Participants heard and saw a novel female speaker who did not appear in either training phase. All labels were spoken in isolation (e.g., "tibu"). The timing of stimulus presentation was similar to the training trials, with the exception of the video and image remaining on the screen for an additional 3000 ms following the speech offset. The ISI was 500 ms between test trials. Participants were instructed to look at the target object. Easy and hard test trials occurred equally often. During each test phase, every object served as the target four times, for a total of 24 trials.

Eye Gaze Coding

Using a still frame of each video, areas of interest (AOIs) were defined by identifying the pixel locations of distinct reference points around the mouth and eyes. One reference point was coded for each eye, using the center of the pupil. For the mouth, four points were coded, one for each corner of the mouth, one on the midline of the upper lip on the vermillion border, and one on the midline of the bottom lip on the vermillion border. Rectangles centered around the reference point for each eye and the mouth were then used to define AOIs. Depending on the video, the rectangle for the mouth AOI was extended by 37 to 53 pixels horizontally and vertically to account for the talker speaking.

Eye movements were recorded as a measure of participants' behavioral responses. Eye gaze data were analyzed with respect to four AOIs: target object, distractor object, talker's eyes, and talker's mouth. If the gaze fell outside of any of these AOIs, or if tracking eye movement was unsuccessful, then eye gaze for that time point was considered as "away."

For the learning phase, gazes within the mouth or eye region were coded as 1 or 0, respectively, for each time point between 0 and 800 ms from the target word onset. This analysis window was chosen to account for listeners' gradual increase in fixations to the mouth at the onset of the auditory stimulus (Lansing & McConkie 2003). We focused on the eyes and mouth because these regions attract the bulk of attention in listeners while viewing a talker speak. For each group, average looks to the mouth were operationalized as the proportion of looks to the mouth relative to the total looks to the eyes and mouth, averaged across trials and the time window. Looks away from the screen and looks to the novel object were coded as NA and excluded from calculation to assess listeners' visual processing strategy while viewing a talker speak.

For the test phase, gaze within the target or distractor object regions was coded as 1 or 0, respectively, at each time point between 300 and 1800 ms from the target word onset. This analysis is consistent with standard eye gaze-based measurements of word learning (Fernald et al. 2008) and enables us to assess the robustness of listeners' newly learned lexical categories as the auditory stimulus unfolds. For each group, mean accuracy was operationalized as the proportion of time spent looking at the target object relative to the total looks to the target or distractor objects, averaged across trials and the time window (300–1800 ms following target word onset). Looks away from the screen and looks to either facial region were coded as NA and excluded from calculation to measure the robustness of listeners' novel word-object representations. For each measurement, trials were excluded if the participant was not fixating to the AOIs (objects, mouth, and eyes) for more than 50% of the critical windows. For each learning condition, participants were excluded from analysis if they did not contribute at least 12 trials per phase. On average, NH listeners contributed 23 trials (SD = 1.3) for the single-talker condition and 23 trials (SD = 2.0) for the multiple-talker condition for each phase. CI listeners contributed 23 trials (SD = 1.6) for the single-talker condition and 23 trials (SD = 2.2) for the multiple-talker condition for each phase.

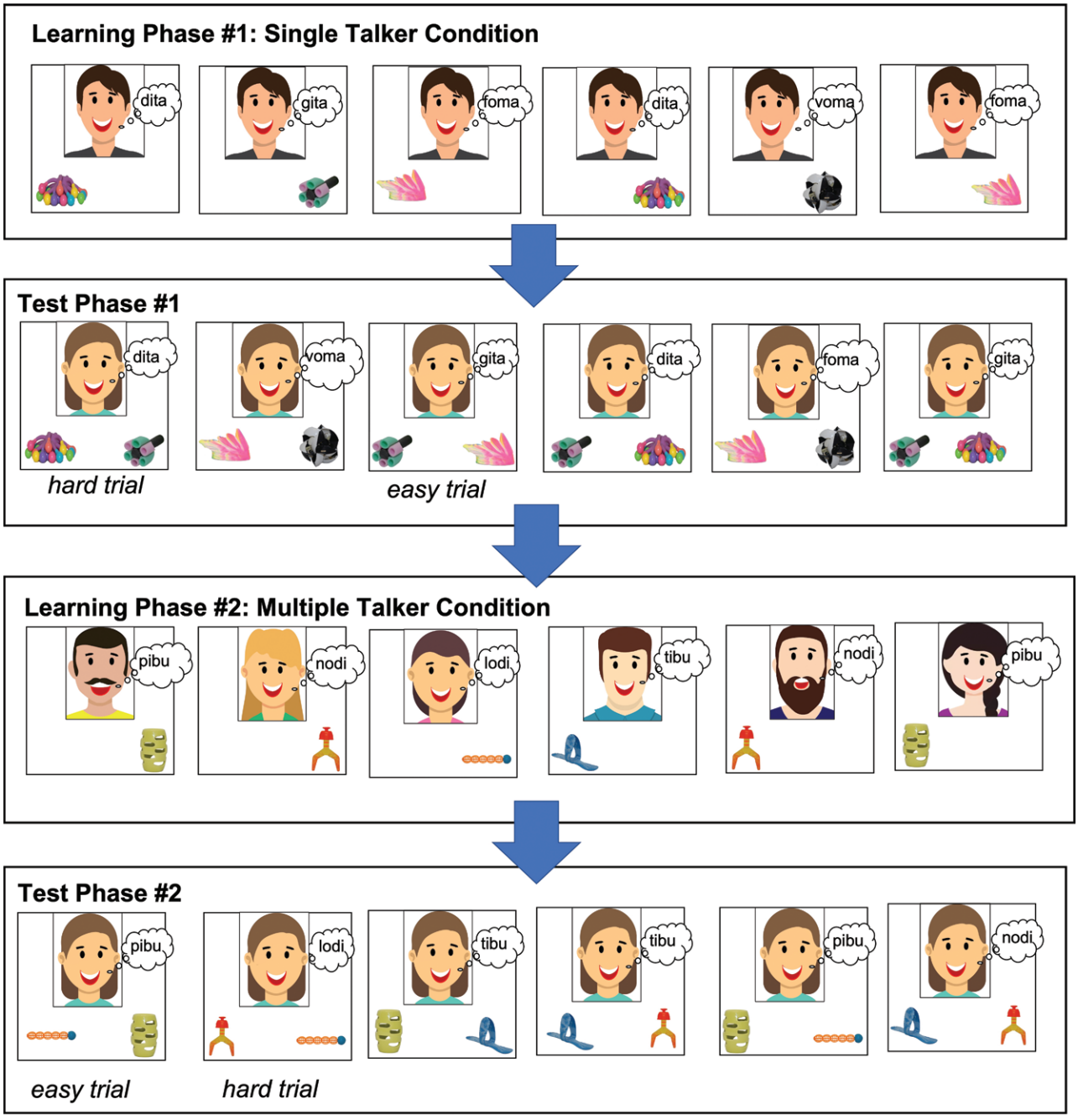

Fig. 2. Experimental paradigm. Learning phases were presented before each test phase. In the learning phases, participants saw and heard the label of a novel object from a same talker (single-talker condition) or from different talkers (multiple-talker condition). Objects were presented one at a time. In the test phase, participants saw two objects and heard the label of one of the objects, spoken by a novel talker.

RESULTS

Mean Accuracy (Test Phase)

First, we assessed the effects of word learning from single versus multiple talkers. We hypothesized that learning from multiple talkers would improve word learning in CI listeners by highlighting the contrastive cues that distinguish the words to be learned. We also predicted that CI listeners would learn words less accurately than listeners with NH.

Initially, we examined the time course of looks to the target object relative to all AOIs (eyes, mouth, target object, and distractor object). This visualization allowed us to get a global view of participants' gaze behaviors during the test phase. To visualize the data, we assigned a value of 1 to gazes within the target AOI and a value of 0 to gazes within the other AOIs. Time was centered from the onset of the target word (i.e., when the talker began speaking). As shown in Figure 3, both groups of listeners gradually increased their gaze to the target relative to all AOIs (eyes, mouth, target object, distractor object) following post-target word onset. Moreover, looks to the target object plateaued to 90% for NH listeners and to 80% for CI listeners, indicating that both groups performed well on the word learning tasks. However, for each group,

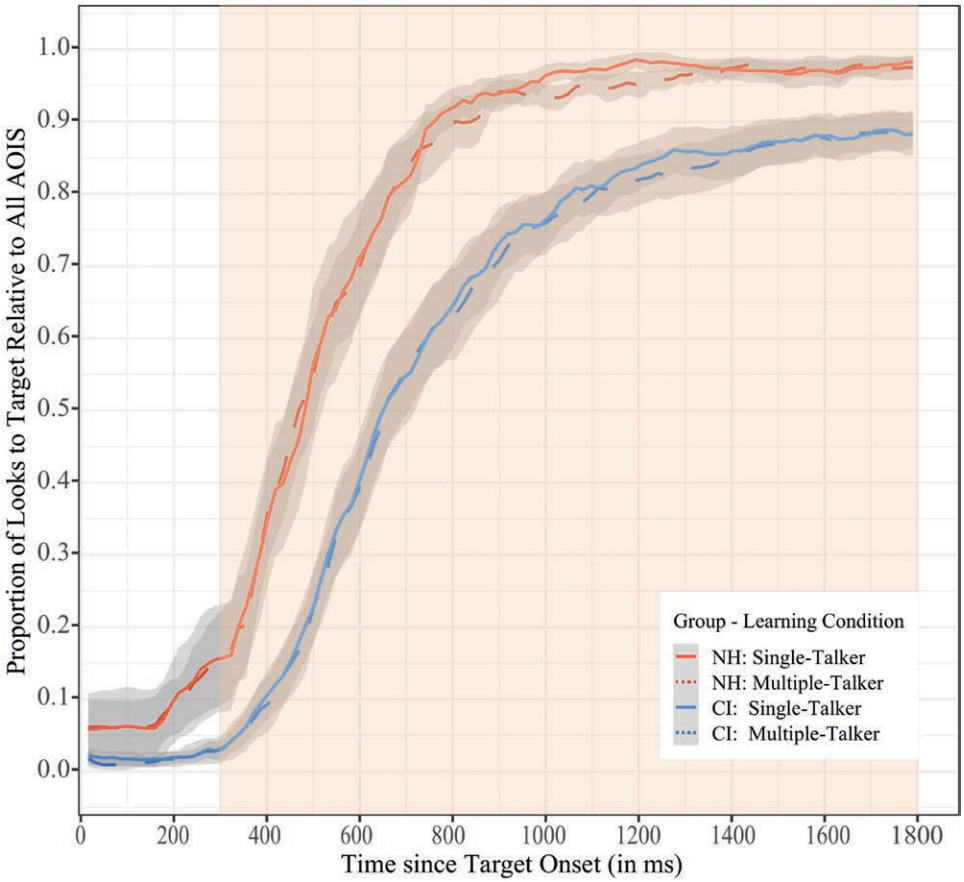

Fig. 3. Time course of fixation to the target by learning condition and hearing status for test phase trials. Proportion of looks to the target relative to the total looks to all AOIs (target, distractor, mouth, and eyes) for NH (red) and CI (blue) listeners. Data represents test trials following the single-talker (solid lines) or multiple-talker (dashed lines) learning condition. Data are averaged across trials. Shaded box represents time window of analysis. Gray ribbons around lines indicates ±1. AOIs indicates areas of interest; CI, cochlear implant; NH, normal hearing.

the single-talker and multiple-talker conditions resulted in similar gaze trajectories. To test our hypotheses, we recoded (following the coding system outlined under "Eye Coding" in Materials and Methods sections) and averaged the eye gaze data across the critical time window (300–1800 ms following target word onset) to examine the overall proportion of looks to the target relative to looks to the target and distractor (mean accuracy). Using a linear mixed effects model, we regressed mean accuracy on the fixed effects of training, test difficulty, and group. We also included three two-way interaction terms, training × group, test difficulty × group, and test difficulty × training, a three-way interaction term, training × group × test difficulty, and a by-subject random intercept and by-subject random slope for training and test difficulty. After this model failed to result in a singular fit, we reduced the random effects structure by removing the by-subject random slope for test difficulty.* Training condition was contrast coded as −0.5 for single-talker trials and 0.5 for multiple-talker trials. Test difficulty was contrast coded as −0.5 for easy test trials and 0.5 for hard test trials. The chance level was set to a proportion value of.50, such that if listeners successfully mapped the word-object pairings, then mean accuracy would be greater than chance (i.e., >50%).

*Final model: accuracy ~ training + test difficulty + group + training:group + test difficulty:group + test difficulty:group:training + (1+training|SubID).

Results from the test phrase indicate that the novel word learning task was fairly easy for both NH and CI listeners, as evidenced by ceiling effects. As seen in Figure 4, listeners with CIs and listeners with NH performed significantly above chance [b = 0.47, t(33.25) = 11.97, p < 0.001]. Contrary to our predictions, there was not a main effect of training on single versus multiple talkers [b = −0.006, F(1, 21.84) = 0.057, p = 0.8] nor an interaction effect between training and group [b = 0.01, F(1,21.84) = 0.057, p = 0.8]. For the CI group, mean accuracy reached 88% (range = 84.8%–91.3%) for the single-talker condition and 89% (range = 86.5%–91.6%) for the multiple-talker condition. For the NH group, mean accuracy reached 95.3% (range = 93.6%–97.0%) and 95% (range = 93.8%–97.1%) for the single-talker and multiple-talker condition, respectively (Fig. 4).

We also tested whether the extent to which talker variability improves word learning in CI listeners is modulated by the phonological similarity between the words. Using the same linear mixed effect model, we found a main effect of test difficulty [b = −0.04, F(1, 1075.0) = 24.7, p < 0.001] and a significant interaction between test difficulty and group [b = 0.07, F(1, 1075.0) = 7.99, p < 0.001]. Overall, performance was higher on the easy test trials [$\mu_{1/2}$ = 94.2%, range = 94.02%–96.2%] compared to the hard test trials [$\mu_{1/2}$ = 86.3%, range = 83.7%–89.0%]. In particular, for CI listeners, performance was higher [b = 0.10, t(1214.1) = 6.57, p < 0.001] on easy trials [$\mu_{1/2}$ = 94.2%, range

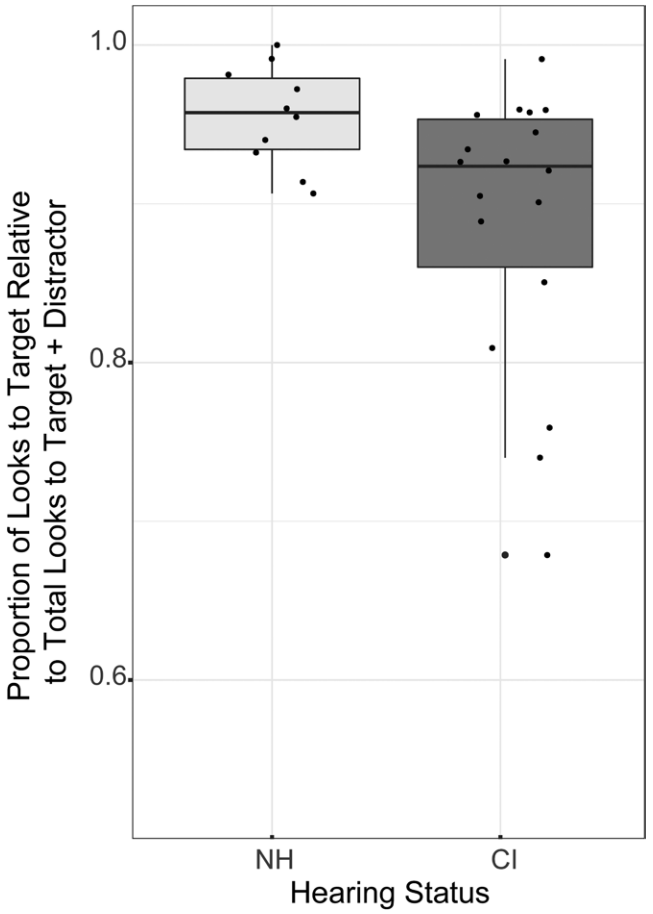

Fig. 4. Proportion of looks to the target relative to the total looks to the target and the distractor objects during the critical time window for NH and CI groups. Data represent the proportion during the test phases. The dark line represents the median. The upper and lower hinges represent the first and third quartiles, respectively (i.e., 25th and 75th percentiles). Data points represent the proportion for each participant. CI indicates cochlear implant; NH, normal hearing.

= 78.4%–100%] compared to hard trials [$\mu_{1/2}$ = 82.8%, range = 48.7%–99.1%]. Moreover, NH listeners were more accurate than CI listeners on both easy trials [NH: $\mu_{1/2}$ = 96.9%, range = 90.1%–100%; $b = -0.10$, $t(36.8) = -3.22$, $p = 0.01$] and hard trials [NH: $\mu_{1/2}$ = 93.8%, range = 89.5%–100%; $b = -0.13$, $t(37.3) = -4.10$, $p = 0.001$], as seen in Figure 5. These findings indicate that CI listeners are able to distinguish phonologically distinct words better than phonologically similar-sounding words, albeit to a lesser extent than those with NH. Finally, we did not see a significant two-way interaction of training and test difficulty [$b = 0.013$, $F(1, 1076.3) = 0.10$, $p = 0.74$] nor a three-way interaction of training, test difficulty, and hearing group [$b = -0.008$, $F(1,1076.4) = 0.02$, $p = 0.80$]. That is, each group performed similarly on easy and hard test trials, regardless of talker variability. In summary, excellent performance by many of the listeners meant that it became challenging to draw conclusions about whether talker variability facilitates word learning in CI listeners, especially in distinguishing phonologically similar words.

Attention to the Talker's Mouth (Learning Phase)

Next, we assessed whether listeners with CIs attended to a talker's mouth more than listeners with NH during the learning phase. We hypothesized that, while learning new words, listeners with CIs would attend to a talker's mouth more than listeners with NH, because listeners with CIs rely more heavily on visual cues during audiovisual speech processing than listeners with NH. We also predicted that listeners' fixation to the mouth would be modulated by speaker variability: listeners would attend more to a talker's mouth when the talker varies across trials than when the talker remains constant.

Similar to the test phase, we initially examined the time course of fixations to the mouth relative to all AOIs to get a global view of listeners' gaze pattern while learning novel words. For this analysis, gazes within the mouth region were coded as 1, whereas gaze within the other AOIs (the eyes or target object) were coded as 0. Time was centered from the onset of the target word (i.e., when the talker began speaking). As seen in Figure 6, both listeners with CIs and with NH gradually increased their looks to the mouth relative to all AOIs (talker's eyes, mouth, and target object), following the onset of the target word during the learning phase. However, listeners with CIs showed more looks to the mouth than listeners with NH. To test our hypotheses, we recoded (following the procedure in "Eye Coding" under Materials and Methods section) and averaged the eye gaze data across the critical time window (0 to 800 ms following the onset of the target word) to examine the proportion of looks to the mouth relative to total looks to mouth and eyes. If CI listeners focused equally on the mouth and eyes, then we would expect the proportion to be close to 0.5. However, if CI listeners rely heavily on visual speech cues coming from the talker's mouth, then the proportion would be greater than 0.5.

Proportion of looks to the mouth was regressed on hearing group, learning condition (contrast coded as −0.5 for single-talker trials and 0.5 for multiple-talker trials), and an interaction of learning condition and hearing group.†‡ We included a by-subject random intercept and a by-subject random slope for learning condition. Proportion of looks to the mouth were significantly higher [$b = 0.27$, $F(1,23.03) = 6.99$, $p < 0.05$] for listeners with CIs [$\mu_{1/2}$ = 99.7%; range = 10.7%–100%] than for listeners with NH [M = 70.9%, range = 81.6%–99.3%], as shown in Figure 7. The main effect of training condition almost reached significance [$b = -0.02$, $F(1, 23.26) = 4.06$, $p = 0.056$]. Interestingly, the effect of training reached significance [$F(1, 23.26) = 4.69$, $p < 0.05$] after RAU transformation. Additionally, contrary to our prediction, the interaction between learning condition and hearing group was not significant [$b = -0.12$, $F(1,24.71) = 3.06$, $p = 0.09$]. Within each group, proportion of looks to the mouth was similar for the multiple-talker condition [NH: $\mu_{1/2}$ = 47.8%, range: 8.2%–99.3%; CI: $\mu_{1/2}$ = 99.9%; range = 20.1%–100%] and the single-talker condition [NH: $\mu_{1/2}$ = 80.1%; range = 12.4%–98.5%; CI: $\mu_{1/2}$ = 99.5%; range = 10.7%–100%], as shown in Figure 8. Altogether, these results show that CI listeners focus more on the talker's mouth than NH listeners while learning new words. However, the proportion of looks to the mouth was unaffected by the number of talkers.

†Model: mouth ~ group + training + group:training + (1+training|SubID).

‡The data from the learning phase also violated the assumption of normality. Thus, we transformed the data using rationalized arcsine transformation. Because the analyses yielded similar findings regardless of whether the data were untransformed or transformed, the untransformed data are being reported.

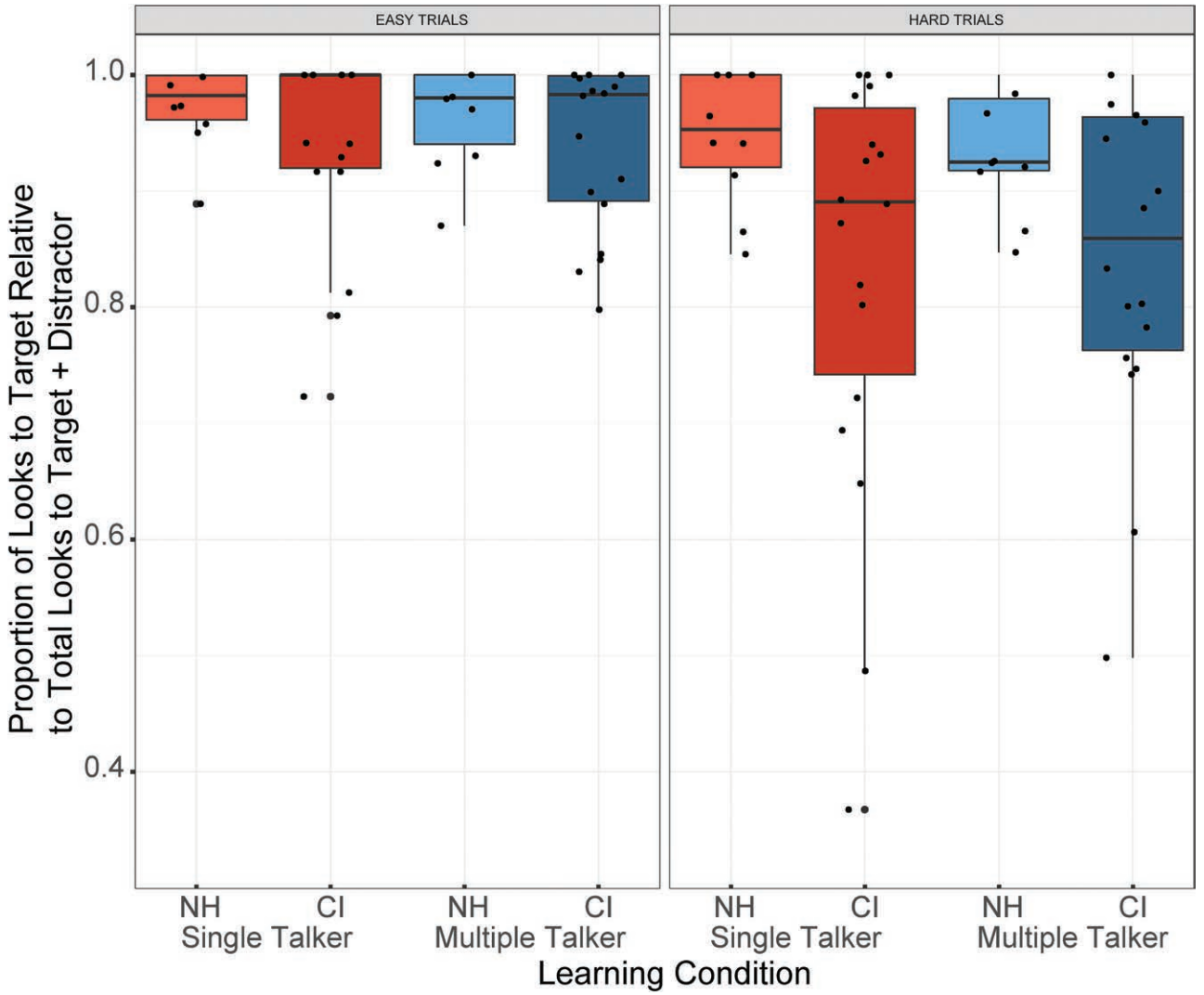

Fig. 5. Proportion of looks to the target relative to the total looks to the target and distractor objects during the critical time window for NH and CI groups. Data represent the proportion for the easy and hard test trials after learning from a single talker (red) or multiple talkers (blue). The dark line represents the median. The upper and lower hinges represent the first and third quartile, respectively (i.e., 25th and 75th percentiles). Data points represent the proportion for each participant. CI indicates cochlear implant; NH, normal hearing.

Relationship Between Looks to Mouth During Training and Performance on Test Trials

Finally, we were interested in examining the relationship between listeners' attention to the talker's mouth during the learning phase and their accuracy on the test phase. We hypothesized that listeners who attended more to the mouth during the learning phase would be more accurate in identifying the target object during testing. However, given the lack of variance on the test trials, the current study was unable to examine this relationship.

DISCUSSION

The purpose of this study was to examine important aspects of speech perception in CI listeners: the effects of talker variability on word learning and eye gaze behavior while viewing a talker speak. The task was easier for CI listeners than had been expected, which resulted in high percent correct scores (i.e., ceiling effects); thus, we were not able to draw conclusions about whether talker variability improves word learning for CI listeners. However, we found that CI listeners attended more to the talker's mouth than NH listeners while learning new words.

Talker Variability and Word Learning in CI Listeners

Our results suggest that the word learning task used here was too easy for the NH and CI group tested in this study. Three factors might have contributed to the ceiling effects. First, our population were experienced language users. Prior studies showing a benefit of talker variability on word learning have focused on young children (Rost & McMurray 2009) and second language learners (Davis 2015; Logan & Pisoni 1995). In our study, most of our participants were deafened after they had already acquired spoken language (mean age of onset of hearing loss = 22 years of age) and were likely to be highly proficient speakers of English. Thus, they might have already developed

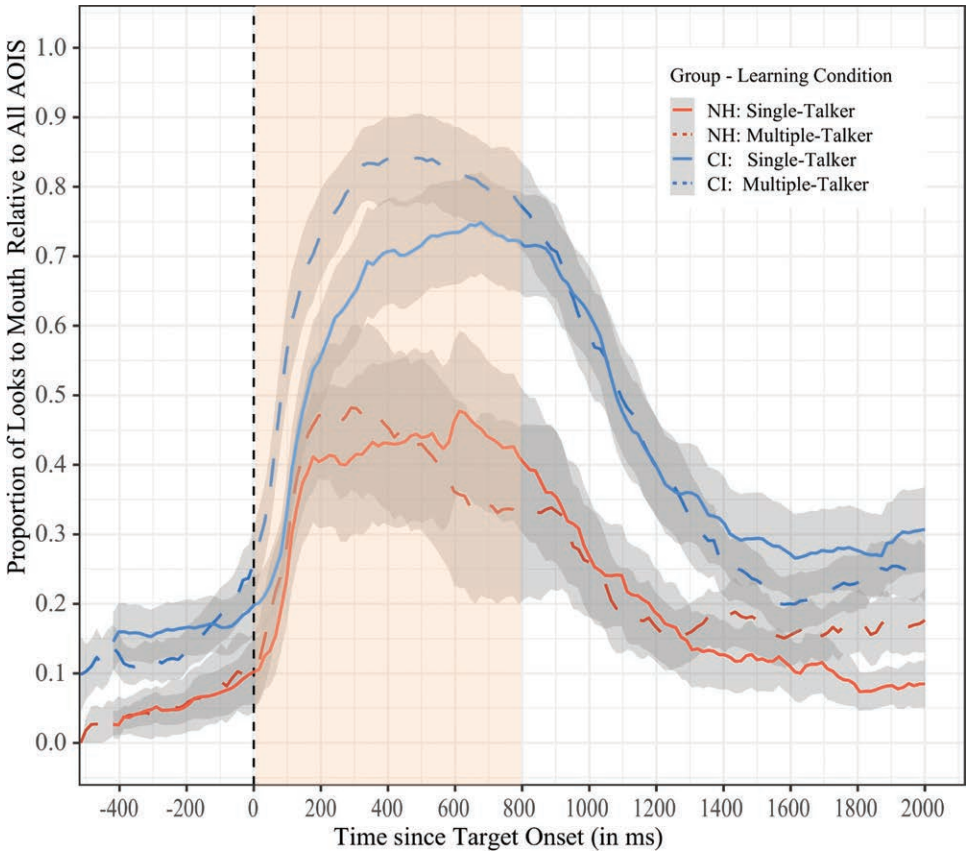

Fig. 6. Proportion of looks to the mouth relative to the total looks to all AOIs (target, mouth, and eyes) for NH (red) and CI (blue) listeners during the single-talker (solid lines) and multiple-talker (dashed lines) training trials. Data are averaged across trials. Shaded box represents time window of analysis. Ribbons around lines indicate ±1 SE. AOIs indicates areas of interest; CI, cochlear implant; NH, normal hearing.

robust phonetic representations that facilitated word learning and encoding. Second, most of the participants tested would be considered "star performers," or listeners with good auditory proficiency. Prior studies showing a benefit of a high variability stimulus training set only tested CI listeners with poor auditory proficiency. Most of our participants scored 80% or higher on the consonant-nucleus-consonant test (see Table 1), indicating that our population consisted of listeners with good CI proficiency. Future studies should enroll participants with a broad range of auditory proficiency to determine if the strong effects of variability depends on auditory proficiency. Finally, the ceiling effect may be attributed to the presence of visual speech cues throughout the experiment. In our study, CI listeners received both acoustic variation and visual speech. It is possible that, in our study, learning from different talkers did not add any benefit beyond the presence of visual cues. Future studies should compare whether talker variability improves word learning in CI listeners when provided with only auditory input versus auditory-visual input.

It is difficult to compare our findings to prior studies that have found improvements in perceptual learning for CI listeners following exposure to a highly variable stimulus set (Miller et al. 2016; Zhang et al. 2021). Unlike the current study, the goals of prior studies were to assess whether learning is possible with the training paradigm, and not the variability, per se. Although the findings from the current study are inconclusive, this study did find a benefit associated with talker variability for some participants. Thus, it is possible that learning from multiple talkers could be beneficial for some CI recipients, although who would most likely benefit is still unclear. To address this question, we calculated a difference score between the talker conditions for each participant and explored the relationship between these scores and three CI demographics: age of participant, onset of deafness, and years of auditory deprivation. Unfortunately, we did not observe a relationship between the benefits of talker variability and any of the demographic factors (data not shown). This lack of relationship is most likely a result of the small sample and limited variability in performance on the word learning task. Future studies with a larger and more diverse pool of CI listeners are needed to assess which type of CI listeners would benefit from talker variability.

Although our findings are inconsistent with prior studies showing a benefit of talker variability on perceptual learning, our results are in line with previous work showing no benefit of variability. For example, Davis (2015) found that talker variability did not improve perception or production of novel words for native English-speaking adults. Similarly, Bulgarelli and Weiss (2021) found that adult NH listeners were equally able to learn an artificial grammar following exposure to a single talker or eight different talkers. Recent work with infants has also revealed limitations of talker variability on early word learning (Bulgarelli & Bergelson 2022, 2023). For instance, talker variability does not help 9-month olds reject mispronunciations of newly learned words (Bulgarelli & Bergelson 2022) nor does it help 14-month olds learn distinct novel words (Bulgarelli & Bergelson 2023). Thus, the current study adds to a growing literature revealing that talker variability is not always helpful for word learning.

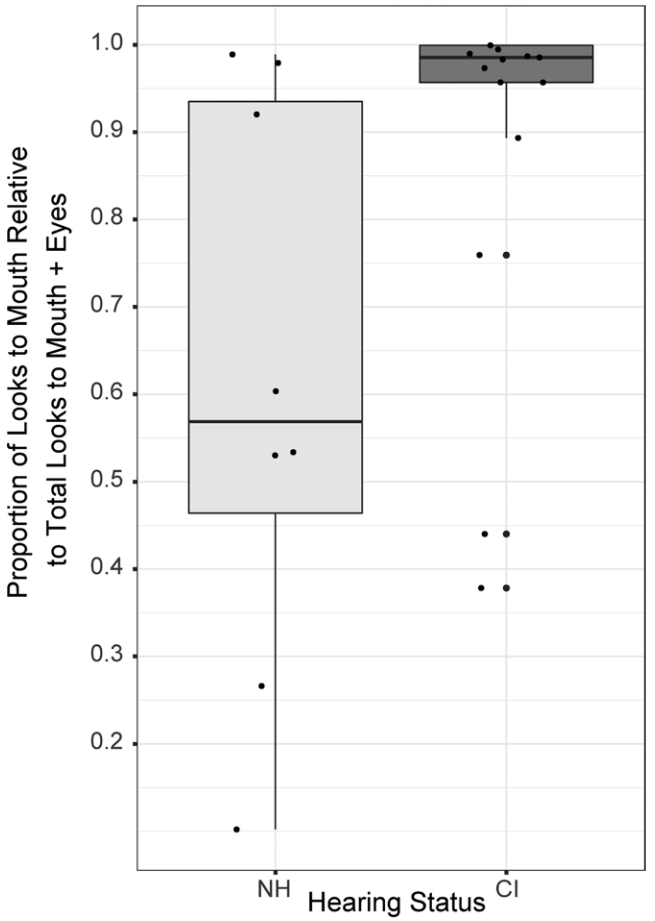

Fig. 7. Proportion of looks to the mouth relative to the total looks to the mouth and eyes during the critical time window for NH and CI groups. Data represent the proportion during the training phases. The dark line represents the median. The upper and lower hinges represent the first and third quartile, respectively (i.e., 25th and 75th percentiles). Data points represent the proportion for each participant. Data points represent the proportion for each participant. CI indicates cochlear implant; NH, normal hearing.

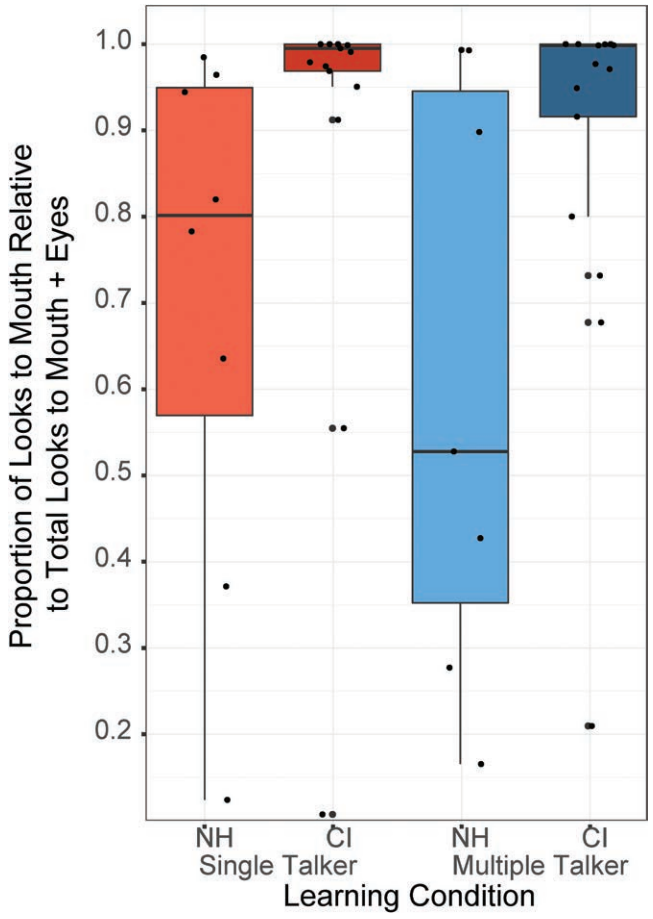

Fig. 8. Proportion of looks to the mouth relative to the total looks to the eyes and mouth during the critical time window for the NH and the CI group in the learning conditions. Data represent the proportion during the training phases. The dark line represents the median. The upper hinges represent the first and third quartile (i.e., 25th and 75th percentiles). Data points represent the proportion for each participant. Data points represent the proportion for each participant. CI indicates cochlear implant; NH, normal hearing.

Turning to the effect of test difficulty, our finding that CI listeners are able to distinguish phonologically distinct words (easy items) better than phonologically similar words (hard items) is in line with previous research showing that CI listeners experience difficulty differentiating between similar-sounding words (Giezen et al. 2010; Havy et al. 2013). However, prior studies have focused on infants and school-aged children with CIs. Our results show that CIs provide insufficient phonological cues to children and adults alike.

One limitation of the current study is words were presented in isolation. We chose to present words in isolation to determine whether acoustic variability alone is sufficient enough to bolster word learning in CI listeners. However, in typical learning settings, words are presented in a sentential or phrasal context. Moreover, listeners with hearing impairments rely heavily on context to compensate for the impoverished acoustic signal (Winn 2016; Holmes et al. 2018). We speculate that if novel words were presented in a sentence or phrase, our findings would remain unchanged if context was weighed more than acoustic variability. Future studies should assess the role of these factors when both cues are presented simultaneously. Another limitation is the difference in sample size and age between each group due to COVID-19. Future studies should match groups by sample size and age.

Audiovisual Speech Processing in CI Listeners

Although both hearing groups' performances reached ceiling, they employed a different visual processing strategy to learn the novel words. Whereas the NH group shifted their gaze equally between the mouth and eyes, the CI group focused primarily on the talker's mouth. Our finding that CI listeners focus more on the talker's mouth than NH listeners extends findings from previous research showing that CI listeners rely more on cues coming from the visual domain than NH listeners (Rouger et al. 2007, 2008; Tremblay et al. 2010). While these studies show that CI listeners weigh visual cues heavily, past research has not shown how this reliance impacts listeners' visual processing strategy while viewing a talker speak. Particularly important is that while NH listeners focus on the mouth during early development (Lewkowicz &Hansen-Tift 2012; Tenenbaum et al. 2013; Hillairet De Boisferon et al. 2018) or under adverse listening conditions (Munhall 1998; Vaitikiotis-Bateson et al. 1998; Król 2018), the current study shows that

CI listeners appear to focus on the mouth even when listening conditions are ideal. This emphasis on visual information may be due to the reliability of visual cues compared to auditory cues. Because of the limited spectro-temporal information conveyed through their processors, CI listeners receive less cues about the acoustic signal. Attending to the mouth might help listeners disambiguate the acoustic signal, given that the mouth region is a primary source for redundant linguistic information. Additionally, CI listeners' attention to the mouth might be a remnant of their period of auditory deprivation. Several studies have shown that CI listeners maintain a high level of speech-reading performance even years after implantation (Strelnikov et al. 2009). Thus, our results suggest that CI listeners utilize a gaze strategy in which they can efficiently extract visual speech information.

In the current study, even CI listeners who experienced short periods of auditory deprivation fixated to the talker's mouth more than 90% of the time. This finding is consistent with evidence from Rouger et al (2007) showing that CI listeners who experienced sudden deafness and were implanted one year later performed similarly in a lipreading task as those who experienced longer periods of auditory deprivation. One might assume that a longer period of auditory deprivation compared to a shorter period might force listeners to become more reliant on visual speech cues. However, our results suggests that any period of deafness might propel listeners to adopt a strategy of attending to the talker's mouth to access speech.

Surprisingly, attention to the mouth was similar when learning from multiple talkers or the same talker for both hearing groups. Prior studies suggest that attention to the mouth increases as the learning conditions becomes more challenging. For example, Buchan et al. (2008) observed modest effects of talker variability on the distribution of eye gaze in NH listeners, such that listeners will fixate more to the mouth when talker varies across trials than when the talker remains constant. Unlike the current study, their study consisted of a large sample size (128 participants). Thus, differences in findings between the current study and the Buchan et al. study might be due to number of participants recruited for each study. It is important to note that RAU transformation of the learning phase data yielded a significant main effect of training. However, this finding should be interpreted with caution, given the small sample size for each group. Further studies with a larger sample size are needed to determine if attention to the mouth is modulated by talker variability.

One limitation of the current study is that we did not find a relationship between CI listeners' attention to the mouth during learning and their performance on the test trials. Given that attention to the mouth improves audiovisual encoding, one might assume that listeners who attend more to the mouth during learning would perform better on the test trials than those to focus on the mouth to the lesser extent. We might not have observed a relationship between these two variables due to the small variability in looks to the mouth and accuracy on the test. Additionally, this lack of correlation might also suggest that direct attention to the mouth is not necessary for accurate speech perception. For example, Lansing & McConkie (2003) found that the number of fixations to the mouth does not correlate with accuracy on a sentence recognition task for NH adults. Future research with a more diverse population of CI listeners is needed to reveal how CI listeners' eye gaze strategy impacts language processing.

Finally, in the current study, NH listeners looked equally to the mouth and eyes while the talker was speaking. This result deviates from Vatikiotis-Bateson et al (1998) showing that in quiet conditions, NH listeners focus on the talker's mouth approximately 35% of the time. This discrepancy between the two studies may be due to the fact the current study recruited an older age population of NH listeners. Our population might have experienced some age-related difficulties with auditory processing and may have relied more on audiovisual cues to facilitate language processing. Nonetheless, the current findings show that CI listeners rely more on visual speech cues to learn novel words than NH listeners.

Clinical Implications

The results of the present study advance the field and are of potential clinical importance for CI listeners. The current study sought to go beyond testing phonetic discrimination in CI listeners, and, instead examined ways to bolster word learning in the moment. While word learning is a primary skill of childhood, it continues throughout the lifespan, and challenges in word learning from auditory information may hinder language processing in adults with CIs. Moreover, word learning provides a different type of window into auditory processing than measures of speech perception. To learn words, listeners must be able to encode the speech form and retain it in order to associate labels with objects. Our task, in particular, called for robust representations of the speech input because listeners were required to generalize from the voices presented during training to a new voice during testing. Due to the ease of our task, the study cannot conclude whether talker variability improves word learning for adult CI listeners. These results do not negate the benefit of talker variability in other settings. In fact, as demonstrated by prior studies (Miller et al. 2016; Zhang et al. 2021), there may be some clinical benefit, particularly for CI adult listeners with poor auditory proficiency or for CI children who are still developing phonetic categories. Thus, future studies should examine how talker variability might influence successful word learning in CI children.

Additionally, our results suggest that when audiovisual cues are available, listeners will utilize the visual cues. Current clinical assessments are administered auditorily. This method may not be capturing how listeners perform in real-life situations. Our results show that listeners rely heavily on visual speech cues and will direct their gaze to talkers' mouths to extract phonetic information. Thus, listeners depend heavily on the mouth to perceive speech. However, there may be a cost-benefit to attending primarily to the mouth. While the mouth conveys linguistic cues, the eyes convey affective and social cues that are also important for efficient language processing. If CI listeners are focusing solely on the mouth, they might miss out on information available at the eyes. Additionally, our results may also provide insight into the challenges encountered by CI listeners in understanding speech throughout the COVID pandemic. Some patients believe that masks are dampening the auditory signal. However, because face masks block the talker's mouth, CI listeners are no longer able to access redundant audiovisual speech cues, which may also lead to challenges in speech perception. Thus, clinicians could counsel CI listeners as to why they are experiencing difficulty in speech understanding

and encourage self-advocacy in finding solutions for better communication.

Overall, our results are consistent with prior work suggesting that adults with CIs are particularly focused on facial information during language processing, and extend those findings by emphasizing the particular importance of the mouth. To our knowledge, this study is the first to directly assess face-scanning behavior in adult CI listeners during online language processing. Future research is needed to assess the potential cost of attending to mouth for CI listeners.

ACKNOWLEDGMENTS

We would like to thank the participating individuals with cochlear implants. We would also like to thank the following people: Daniel Bolt for help in statistical analysis; Won Jang for assistance in programming; and Ron Pomper and Ellen Peng for their comments on previous versions of this article.

This work was funded by the National Institute of Health Grant R01DC00308 (awarded to R.L.) and the National Science Foundation GRFP-DGE-1256259 (awarded to J.H.). This study was supported in part by a core grant to the Waisman Center from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (U54 HD090256) and in part by the Neuroscience Training Program (NIH/NINDS T32 NS105602).

J.H. designed and performed experiments, analyzed data, and wrote the paper; J.S. and R.L. designed experiments. All authors discussed the results and implications and commented on the manuscript at all stages.

Ruth Litovsky is the Editor in Chief of Ear and Hearing, thus the paper was handled by a Guest Editor.

Address for correspondence: Jasenia Hartman, Department of Psychology, Harvard University, 33 Kirkland St, Cambridge, MA 02138 USA. Email: jaseniahartman@fas.harvard.edu

Received March 29, 2022; accepted August 6, 2023

REFERENCES

American National Standards Institute. (1989). American National Standards for audiometers (ANSI S3.6-1989). New York: American National Standards Institute.

Birulés, J., Bosch, L., Brieke, R., Pons, F., & Lewkowicz, D. J. (2019). Inside bilingualism: Language background modulates selective attention to a talker's mouth. *Dev Sci, 22,* e12755.

Buchan, J. N., Paré, M., Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Res*, *1242*, 162–171.

Bulgarelli, F., & Weiss, D. J. (2021). Desirable difficulties in language learning? How talker variability impacts artificial grammar learning. *Lang learn, 71,* 1085–1121.

Bulgarelli, F., & Bergelson, E. (2022). Talker variability shapes early word representations in English-learning 8-month olds. *Infancy*, *27*, 341–368.

Bulgarelli, F., & Bergelson, E. (2023). Talker variability is not always the right noise: 14 month olds struggle to learn dissimilar word-object pairs under talker variability conditions. *J Exp Child Psychol*, *227*, 105575.

Davidson, L. S., Geers, A. E., Nicholas, J. G. (2014). The effects of audibility and novel word learning ability on vocabulary level in children with cochlear implants. *Cochlear Implants Int*, *15*, 211–221.

Davis, A. (2015). The interaction of language proficiency and talker variability in learning (Doctoral dissertation). http://hdl.handle.net/10150/556484

Desai, S., Stickney, G., Zeng, F.-G. (2008). Auditory-visual speech perception in normal-hearing and cochlear-implant listeners. *J Acoust Soc Am*, *123*, 428–440.

Dorman, M. F., Loizou, P. C., Spahr, A. J., Maloff, E. (2002). Factors that allow a high level of speech understanding by patients fit with cochlear implants. *Am J Audiol*, *11*, 119–123.

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. *Neuropsychol, Dev, and Cogn. Section B, Aging, Neuropsychol, and Cogn, 22,* 413–437.

Fernald, A. E., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening: Using eye movements to monitor spoken language comprehension by infants and young children. In Sekerina I, Fernandez E, Clahsen H, (Eds.), *Developmental Psycholinguistics: On-line Methods in Children's Language Processing* (pp. 97–135). John Benjamins Publishing Company.

Giezen, M. R., Escudero, P., Baker, A. (2010). Use of acoustic cues by children with cochlear implants. *J Speech Lang Hear Res*, *53*, 1440–1457.

Giezen, M. R., Escudero, P., & Baker, A. E. (2016). Rapid learning of minimally different words in five- to six-year-old children: effects of acoustic salience and hearing impairment. *J of child lang, 43,* 310–337.

Gómez, R. L. (2002). Variability and detection of invariant structure. *Psychol Sci*, *13*, 431–436.

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 Hear*, *38*, e335–e342.

Havy M, Nazzi T, Bertoncini J. (2013). Phonetic processing during the acquisition of new words in 3-to-6-year-old French-speaking deaf children with cochlear implants. *J Commun Disord, 46,* 181–192.

Hillairet De Boisferon, A., Tift, A. H., Minar, N. J., Lewkowicz, D. J. (2018). The redeployment of attention to the mouth of a talking face during the second year of life. *J Exp Child Psychol*, *172*, 189–200.

Holmes, E., Folkeard, P., Johnsrude, I. S., & Scollie, S. (2018). Semantic context improves speech intelligibility and reduces listening effort for listeners with hearing impairment. *Int J Audiol, 57,* 483–492.

Horst, J. S., & Hout, M. C. (2016). The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research. *Behav Res Methods*, *48*, 1393–1409.

Houston, D. M., & Miyamoto, R. T. (2010). Effects of early auditory experience on word learning and speech perception in deaf children with cochlear implants: Implications for sensitive periods of language development. *Otol Neurotol*, *31*, 1248–1253.

Houston, D. M., Stewart, J., Moberly, A., Hollich, G., Miyamoto, R. T. (2012). Word learning in deaf children with cochlear implants: effects of early auditory experience. *Dev Sci*, *15*, 448–461.

IJsseldijk, F. J. (1992). Speechreading performance under different conditions of video image, repetition, and speech rate. *J Speech Hear Res*, *35*, 466–471.

Iverson, P. (2003). Evaluating the function of phonetic perceptual phenomena within speech recognition: An examination of the perception of/d/–/t/ by adult cochlear implant users. *J Acoust Soc Am*, *113*, 1056–1064.

Król, M. E. (2018). Auditory noise increases the allocation of attention to the mouth, and the eyes pay the price: An eye-tracking study. *PLoS One*, *13*, e0194491.

Lane, H., Denny, M., Guenther, F. H., Hanson, H. M., Marrone, N., Matthies, M. L., Perkell, J. S., Stockmann, E., Tiede, M., Vick, J., Zandipour, M. (2007). On the structure of phoneme categories in listeners with cochlear implants. *J Speech Lang Hear Res*, *50*, 2–14.

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Percept Psychophys*, *65*, 536–552.

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proc Natl Acad Sci USA*, *109*, 1431–1436.

Lively, S. E., Logan, J. S., Pisoni, D. B. (1993). Training Japanese listeners to identify English/r/ and/l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *J Acoust Soc Am*, *94*, 1242–1255.

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

Miller, S., Zhang, Y., Nelson, P. (2016). Neural correlates of phonetic learning in postlingually deafened cochlear implant listeners. *Ear Hear*, *37*, 514–528.

Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Percept Psychophys*, *60*, 926–940.

Munson, B., & Nelson, P. B. (2005). Phonetic identification in quiet and in noise by listeners with cochlear implants. *J Acoust Soc Am*, *118*, 2607–2617.

Munson, B., Donaldson, G. S., Allen, S. L., Collison, E. A., Nelson, D. A. (2003). Patterns of phoneme perception errors by listeners with cochlear implants as a function of overall speech perception ability. *J Acoust Soc Am*, *113*, 925–935.

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

Peng, Z. E., Hess, C., Saffran, J. R., Edwards, J. R., Litovsky, R. Y. (2019). Assessing fine-grained speech discrimination in young children with bilateral cochlear implants. *Otol Neurotol*, *40*, e191–e197.

Perry, L. K., Samuelson, L. K., Malloy, L. M., Schiffer, R. N. (2010). Learn locally, think globally: Exemplar variability supports higher-order generalization and word learning. *Psychol Sci*, *21*, 1894–1902.

Peterson, N. R., Pisoni, D. B., Miyamoto, R. T. (2010). Cochlear implants and spoken language processing abilities: Review and assessment of the literature. *Restor Neurol Neurosci*, *28*, 237–250.

Pimperton, H., & Walker, E. A. (2018). Word learning in children with cochlear implants: Examining performance relative to hearing peers and relations with age at implantation. *Ear Hear*, *39*, 980–991.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *J Exp Psychol*, *77*, 353–363.

Quam, C., Knight, S., Gerken, L. (2017). The distribution of talker variability impacts infants' word learning. *Lab Phonol*, *8*, 1–27.

Quittner, A. L., Cejas, I., Wang, N. Y., Niparko, J. K., Barker, D. H. (2016). Symbolic play and novel noun learning in deaf and hearing children: Longitudinal effects of access to sound on early precursors of language. *PLoS One*, *11*, e0155964.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Dev Sci*, *12*, 339–349.

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, *15*, 608–635.

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O., Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proc Natl Acad Sci USA*, *104*, 7295–7300.

Rouger, J., Fraysse, B., Deguine, O., Barone, P. (2008). McGurk effects in cochlear-implanted deaf subjects. *Brain Res*, *1188*, 87–99.

Shannon, R. (2002). The relative importance of amplitude, temporal, and spectral cues for cochlear implant processor design. *Am J Audiol*, *11*, 124–127.

Stevenson, R. A., Sheffield, S. W., Butera, I. M., Gifford, R. H., Wallace, M. T. (2017). Multisensory integration in cochlear implant recipients. *Ear Hear*, *38*, 521–538.

Strelnikov, K., Rouger, J., Barone, P., Deguine, O. (2009). Role of speechreading in audiovisual interactions during the recovery of speech comprehension in deaf adults with cochlear implants. *Scand J Psychol*, *50*, 437–444.

Tenenbaum, E. J., Shah, R. J., Sobel, D. M., Malle, B. F., Morgan, J. L. (2013). Increased focus on the mouth among infants in the first year of life: A longitudinal eye-tracking study. *Infancy*, *18*, 534–553.

Tremblay, C., Champoux, F., Lepore, F., Théoret, H. (2010). Audiovisual fusion and cochlear implant proficiency. *Restor Neurol Neurosci*, *28*, 283–291.

Tsang, T., Atagi, N., Johnson, S. P. (2018). Selective attention to the mouth is associated with expressive language skills in monolingual and bilingual infants. *J Exp Child Psychol*, *169*, 93–109.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Percept Psychophys*, *60*, 926–940.

Walker, E. A., & McGregor, K. K. (2013). Word learning processes in children with cochlear implants. *J Speech Lang Hear Res*, *56*, 375–387.

Winn, M. B., & Litovsky, R. Y. (2015). Using speech sounds to test functional spectral resolution in listeners with cochlear implants. *J Acoust Soc Am*, *137*, 1430–1442.

Winn, M. B., Rhone, A. E., Chatterjee, M., Idsardi, W. J. (2013). The use of auditory and visual context in speech perception by listeners with normal hearing and listeners with cochlear implants. *Front Psychol*, *4*, 824.

Winn M. B. (2016). Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. *Trends Hear, 20,* 2331216516669723.

Zhang, H., Ding, H., & Zhang, Y. (2021). High-variability phonetic training benefits lexical tone perception: An investigation on Mandarin-speaking pediatric cochlear implant users. *J Speech Lang Hear Res, 64,* 2070–2084.