

The Role of Talker Variability and Visual Speech on Word Learning in Adverse Listening
Conditions

By

Jasenia T. Hartman

A dissertation submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

(Neuroscience)

at the

UNIVERSITY OF WISCONSIN-MADISON

2022

Date of final oral examination: 06/03/2022

The dissertation is approved by the following members of the Final Oral Committee:

Ruth Y. Litovsky, Professor, Communication Sciences and Disorders

Jenny Saffran, Professor, Psychology

Caroline Niziolek, Assistant Professor, Communication Sciences and Disorders

Meyer Jackson, Assistant Professor, Neuroscience

Federick J. Gallun, Professor, Otolaryngology

Acknowledgements

Thank you to my village. I would not have made it through this journey if it were not for you. Mom, I can't thank you enough for the sacrifices you made to ensure I had a good education. I am a diligent and hard worker because of you. To my sisters, Chrissy, Briana, and Ashley, and to my godmothers, Dawn and Martina - thank you for the laughs. At times, pursuing this PhD was stressful, but you have always reminded me to enjoy life. To my friends back home, thank you for your words of encouragement. To my best friends, Stephanie and Scarlet, thank you for lending an ear. I can always count on you two to be there for me in times of need. To my sister-friend, Michelle, thank you for serving as a paragon of black excellence in research. I aspire to reach the same level of excellence as you. To my friends in Madison, thank you for creating a space for me. You have made my time in Madison enjoyable. To my loving partner, Rachel, thank you for being my biggest cheerleader and for feeding me.

Thank you to my co-advisors, Ruth Litovsky and Jenny Saffran, for giving me the freedom and support to explore my research interest. Your knowledge and advice have helped to gain self-confidence as a scientist. Thank you to my committee members, Erick Gallun, Meyer Jackson, and Carrie Niziolek, for your feedback on my work. Thank you to my labmates, especially Tanvi and Taylor, for the countless discussions on language development. To my former mentors, thank you for providing me with the tools to navigate academic spaces.

Table of Contents

| | |
|--|----|
| ACKNOWLEDGEMENT | i |
| ABSTRACT | iv |
| CHAPTER 1 | 1 |
| Introduction..... | 1 |
| Thesis Overview..... | 3 |
| CHAPTER 2 | 6 |
| Face-Scanning Behavior in Children and Adults During Language Processing | |
| <i>Introduction</i> | 6 |
| <i>Face-scanning behavior in adults during language processing</i> | 7 |
| <i>Face-scanning behavior in children during language processing</i> | 12 |
| <i>Neural bases of face-to-face communication</i> | 18 |
| <i>Future directions</i> | 21 |
| <i>Conclusion</i> | 22 |
| CHAPTER 3 | 24 |
| Word Learning in Deaf Adults who Use Cochlear Implants: The Role of Talker | |
| Variability and Attention to the Mouth..... | |
| <i>Abstract</i> | 24 |
| <i>Introduction</i> | 25 |
| <i>Methods</i> | 33 |
| <i>Results</i> | 40 |
| <i>Discussion</i> | 50 |
| <i>Acknowledgements</i> | 57 |

| | |
|---|-----------|
| CHAPTER 4..... | 58 |
| The Role of Talker Variability and Visual Speech on Word Learning in Noise in Typical- Hearing Adults..... | |
| <i>Introduction.....</i> | <i>58</i> |
| <i>Methods.....</i> | <i>63</i> |
| <i>Results.....</i> | <i>68</i> |
| <i>Conclusion.....</i> | <i>77</i> |
| CHAPTER 5..... | 80 |
| General Discussion..... | |
| <i>The role of talker variability.....</i> | <i>81</i> |
| <i>The role of visual speech on learning.....</i> | <i>82</i> |
| <i>Visual attention to talking faces.....</i> | <i>84</i> |
| <i>The role of visual speech on learning in noise.....</i> | <i>86</i> |
| <i>Conclusion.....</i> | <i>86</i> |
| REFERENCES..... | 88 |

Abstract

Word learning is a core language skill that is characterized by mapping sounds to meaning. To successfully acquire newly spoken words, listeners must accurately perceive the speech sounds found in the word form. Studies have shown that talker variability (learning from different talkers) and visual speech cues facilitate the acquisition of spoken words. While the benefits associated with talker variability and visual speech cues have been examined in ideal learning situations, it is unclear whether these cues can be useful in suboptimal learning settings, such as listening with a cochlear implant (CI) or in noise. Using two novel word learning tasks, one in quiet (Study 1) and one in noise (Study 2), this study examined the role of talker variability and visual speech on word learning in adverse listening conditions.

Eighteen adults with CIs and forty-nine adults with TH ($N = 10$ for Study 1, $N = 39$ for Study 2) were taught novel word-object associations spoken by a single talker or by six different talkers. Across both tasks, learners saw an image of the talker. However, the presence of visual cues were manipulated for Study 2. Learning was probed using a novel talker in a two-alternative force choice task. Learners' eye movements to the mouth (Study 1) and to the target object (Studies 1 and 2) were tracked over time.

In both studies, talker variability did not enhance word learning for adult CI and NH listeners. Both groups were successful at acquiring newly learned words, regardless of whether they learned from the same talker or from different talkers. Compared to listeners with NH, listeners with CIs focused significantly more on the talker's mouth. In the presence of background noise, listeners with NH performed better when presented with audiovisual cues during learning than only audio cues. In summary, these outcomes suggest visual cues, but not talker variability, may facilitate learning under adverse listening conditions.

Chapter 1

Introduction

A fundamental skill of spoken language acquisition is the ability to associate labels with objects. This skill, also referred to as word learning, emerges during infancy and continues to develop across the lifespan. To successfully learn new words, listeners must encode the speech sounds found in the word form. Both acoustic variability (e.g. hearing different talkers label an object) and visual speech cues have been shown to support learning.

Studies have shown that talker variability plays a pivotal role in learning. For example, Rost & McMurray, (2009, 2010) have shown that increasing variability allows infants to develop robust representations of newly formed words. Similarly, Lively et al. (1993) have found that adult second language learners succeeded in learning lexical categories when exposed to multiple talkers than to a single talker. Talker variability is beneficial because it draws learners' attention to the relevant acoustic dimension that contrast the words to be learned. In the presence of variation, many acoustic dimensions, such as prosody and pitch, vary substantially across listeners. However, these cues are uninformative in distinguishing the phonetic categories of the words. Other acoustics cues show little variation between talkers and are informative in distinguishing phonetic categories. Through variability, listeners are able to weigh which acoustic dimension is important for distinguishing phonetic categories of the words. As a result, learners are able to develop robust representation of words that they can generalize to new instances.

Just as talker variability is beneficial for learning, so is the presence of visual cues. Studies have shown that audiovisual cues facilitate language acquisition by providing redundant information to the auditory signal (Chandrasekaran et al., 2009; Mcdaniel et al., 2018;

Tenenbaum et al., 2013, 2015; Thomas & Jordan, 2004). More importantly, these studies have posited that learners' visual processing strategies towards a talking face influences their language outcome. For example, Tenenbaum (2013) found that infants show increasing interest to the mouth around the age in which language production emerges. This study also showed large variation in face-scanning behavior between infants. This finding suggests that learners who attend more to the mouth may be better at capturing and benefiting from phonetic details available at the mouth.

In summary, both talker variability and audiovisual speech play an important role in facilitating spoken language acquisition. Whereas audiovisual cues provide redundant speech cues, talker variability allows listeners to infer the structure of the phonetic categories. However, the role of talker variability and audiovisual speech has been examined in ideal listening conditions. Most learning situations occur in suboptimal listening conditions which could impact the integrity of the acoustic signal.

One example of a suboptimal acoustic condition is listening with a cochlear implant (CI). Unlike the acoustic signal transmitted through the normal-hearing system, the signal provided by the CI is impoverished and degraded. For example, unlike the NH which contains 32 auditory filters, the CI contains up to 22 electrodes, 8 of which can be active at the same time. Additionally, the CI system contains "dead regions", or areas in the cochlear where there no spiral ganglion cells, due to degeneration of these cells caused by deafness. As a consequence, some CI listeners experience challenges in encoding the phonological information words, and thus, acquire words less accurately than their NH peers (Davidson et al., 2014a, 2014b; Houston et al., 2012a, 2012b; Pimperton & Walker, 2018; Walker & McGregor, 2013).

Another example of learning with an impoverished acoustic signal is listening with background noise. Studies have shown that noise can sometimes makes spoken language acquisition difficult for both adults and children (Morini & Newman, 2020; Riley & McGregor, 2012). One reason for this reduction in learning is due to the fact that noise masks the acoustic signal. For example, Bidelman et al (2019) found that noise weakens categorical perception. Additionally, noise hampers learning by distracting listeners from attending to the target sound. Thus, similar to listening with CI, noise creates a challenging learning situation by interfering or distracting listeners from encoding speech sounds of the word forms.

Given that talker variability and audiovisual speech augments phonological processing, it is possible that these cues might facilitate learning, even if the acoustic input is degraded. Indeed, within the last decade, studies have shown that high stimulus variability improves speech for CI listeners (Miller et al., 2016; Zhang et al., 2021). Additionally, in regards to audiovisual input, studies have shown that seeing a talker's face improves categorical perception (Bidelman et al., 2019) and speech recognition (Basirat et al., 2018; Bergeson-dana et al., 2005; Kaiser et al., 2003; Kirk & Pisoni, 2002) in adverse listening conditions. Specifically, in the presence of noise, listeners will adapt a visual processing strategy, by directing their gaze to the nose or mouth, as a means of efficiently capture audiovisual cues. However, none of these studies have examined the benefits of these cues in a word learning context. Word learning serves as an interesting topic because the acoustic signal needs to be reliable so that listeners can form associations between unfamiliar words and corresponding meaning.

Thesis Overview

The purpose of this dissertation is to explore the role of talker variability and audiovisual speech on learning in adverse listening conditions. In this outlined studies, we focused on adults.

Although word learning is typically associated with young children, it is also important in adulthood (e.g. acquiring technical terminology for career). To address these aims, we taught listeners novel word-object pairings using word learning tasks. The format of the dissertation is as follows:

Chapter 2 consists of a theoretical review paper on how individuals direct their visual attention while viewing a talker speak. In this review, I argue that efficient language processing requires individuals to dynamically scan a talking face to capture both linguistic and social information. This review will describe the findings and limitations of behavioral studies on face-scanning behavior in children and adults. It will also discuss recent advances in neuroimaging in elucidating the role of linguistic and social cues on communication.

Chapter 3 will explore the role of talker variability and attention to the mouth on word learning in adult CI listeners. Specifically, it will address whether talker variability enhances word learning in CI listeners. This chapter will also compare how NH and CI listeners direct their gaze while viewing a talking face.

Chapter 4 will examine the role of talker variability and visual speech on word learning in noise with NH listeners. The purpose of this chapter is to 1) examine the contribution of talker variability and audiovisual speech on word learning in noise, and 2) to assess if learning in noise is enhanced when these cues are combined compared to presented alone. This study will provide insight into how individuals use information in a noisy environment to facilitate learning.

Chapter 5 will summarize findings from chapters 3 and 4. It will also discuss the limitations of each study and future directions.

Altogether, this dissertation will contribute to the literature on word learning. More specifically, it will explore the extent to which talker variability and audiovisual speech cues support learning in adverse listening conditions.

Chapter 2. Face-Scanning Behavior in Children and Adults During Language Processing

Introduction

Social interactions are typically an audiovisual process; in most situations, communication occurs face-to-face. Viewing a talker speak is extremely beneficial. For example, audiovisual speech has been shown to improve speech understanding in quiet and in noise (Sumbly & Pollack, 1954, see Peelee & Sommers, 2015 for review). It has also been shown to speed up processing of the auditory signal (Hisanaga et al., 2016; van Wassenhove et al., 2005). Because visual cues play an important role in speech perception, it is important to understand how perceivers capture information from a talking face to effectively process speech.

To efficiently process speech, listeners must capture and integrate the different types of information available on the face. Most studies on audiovisual speech processing have focused on the visual cues available at the mouth. The mouth has been a topic of interest because of the complementary cues it provides to acoustic signals. Some studies have postulated that the mouth region is sufficient for speech recognition (Brooke & Summerfield, 1983; IJsseldijk, 1992; Jordan & Thomas, 2011; Marassa & Lansing, 1995; Thomas & Jordan, 2004). However, these studies ignore critical aspects of communication: the social, affective, and indexical information conveyed by the eyes. Indeed, communication is not solely a linguistic phenomenon, but also contain social cues, such as the referential intent and identity of the talker.

Thus, the present review argues that efficient language processing requires individuals to dynamically scan a talking face to capture both linguistic and social information. This review paper will consist of four sections. In the first section, I will provide an overview of adults' gaze strategies while viewing a person speak. In the second and third sections, I will describe eye-tracking studies with children and non-invasive neuroimaging studies examining the role of

audiovisual cues on language processing, respectively. Using the first section as a critical lens, I will explain how the current metrics used to examine infants' attention to the eyes and mouth might underestimate how children direct their gaze while viewing a talker speak. Similarly, for the third section, I will discuss the limitations and recent advances of neuroimaging studies in elucidating the neural correlates of face-to-face interactions. Finally, the fourth section will summarize the key findings and discuss future directions of the field. The overarching goal of this review is to provide insight as to why a dynamic gaze strategy facilitates successful language processing.

Face-scanning behavior in adults during language processing

Eye-tracking studies suggest that adults have developed a flexible gaze strategy that allows them to extract different streams of information. When instructed to identify words or stress pattern, adults fixated on the mouth (Lansing and McConkie, 1999). In contrast, when asked to judge the intonation of a sentence (i.e., question/ statement sentences), adults tend to fixate on the eyes (Cvejic et al., 2012; Kim et al., 2014; Lansing & McConkie, 1999). Furthermore, occluding a facial region can impact performance. For example, in the same study, Lansing and McConkie (1999) found that occluding the eyes reduced performance on the intonation task. This pattern reflects the importance of the eyes for conveying intonation. Studies have shown that changes in fundamental frequency contour correlate with eyes and eyebrow motion (Cvejic et al., 2012; Kim et al., 2014; Srinivasan & Massaro, 2003). Thus, observers use cues from the upper region of the face to make judgements about intonation.

Prosodic information is not confined to the upper region of the face. Consistent with the findings from Lansing and McConkie, Cvejic et al (2012) found that the upper facial region allowed for accurate discrimination of intonation (question/statement). In contrast, the lower

region allowed individuals to judge whether emphasis is placed on an entire utterance or one part of a sentence (i.e., broad compared to focused statements). Thus, successful speech understanding involves listeners knowing where to direct their gaze to capture different aspects of speech.

In addition to capturing prosodic information, perceivers must also determine the affective state of the talker. Buchan et al. (2007) found that the eyes conveyed information about the emotional state of the talker. In their study, the authors demonstrated that while viewing an expressive speaker, adults directed their gaze to the talker's eyes to evaluate the affective state of the speaker, but focused on the talker's mouth to perceive the words being said. In contrast, Blais et al (2017) found that perceivers focused more on the center of the face while viewing dynamic expressive faces. However, different facial regions were most useful at categorizing certain facial expressions. For example, both the eyes, eyebrows, and mouth were useful for categorizing anger. However, for expressions that convey disgust and fear, the mouth was most useful. Altogether, these findings confirm that adults are sensitive to the distribution of different information across the face. As a result, they tend to orient their attention to a particular facial region that allows them to maximally extract specific aspects of speech and emotions.

In addition to adapting a flexible gaze strategy, adults also show a dynamic gaze pattern while viewing a talker speak. For instance, Lansing & McConkie (2003) monitored the spatial and temporal characteristics of adults' eye movements while viewing videos of a talker reciting a monologue. This methodical approach has the advantage of allowing researchers to capture the sequence and duration of eye fixations towards a particular facial region as speech unfolds. The authors found that, prior to, and after a speech period, adults tended to direct their gaze to the talker's eyes. Moreover, as the talker was speaking, adults showed a higher proportion and a

longer duration of looks to the talker's mouth. Interestingly, after the first trial, adults show increased fixations to the mouth one second prior to the onset of speech, suggesting that they fixate to the mouth in anticipation of a speech event. These findings were interpreted as reflecting "two forces of operation." The first operation, the "eye primacy effect", initially draws attention to the talker's eyes. This initial eye contact may serve as a signal of one's intention to communicate. It may have also developed through human evolution or as a learned experience. The second operation, the "information source attention effect," draws attention to the talker's mouth, possibly as a means of capturing redundant audiovisual speech cues. Altogether, these findings underscore how adults will employ a gaze strategy that allows them to seek information from different parts of the face.

While these studies reveal the gaze strategy employed by adults while viewing a speaker talk, they do not confirm whether direct gaze to a particular facial region facilitates language processing. If the mouth is the primary source of linguistic cues, then one might assume that direct fixation to mouth would correlate with performance on speech tasks. Similarly, if the eyes are a primary source for indexical, social, and affective content, then one might assume a relationship between attention to the eyes and emotion or talker recognition. However, if no relationship exists between attention to a facial region and performance on a task, then this might suggest that individuals can still capture information from a particular region without focusing on said region.

Some studies have explored this very question. Much of this work has focused on the relationship between adults' attention to the mouth and speech recognition. Using correlational analysis, no relationship between attention to the mouth and performance on speech tasks has been found. For example, Lansing and McConkie (2003) found that the number of fixations to

the mouth did not correlate with accuracy on a sentence recognition tasks. Other studies have explored whether fixations at different distances from the mouth might reduce audiovisual integration. For example, Pare et al. (2003) examined how fixating to different facial regions (mouth, eyes, and hairline) would impact adults' susceptibility to the McGurk effect, which occurs when an auditory signal (e.g., /ba/) is paired with a visual stimuli (e.g., /ga/), leading to an illusory response (e.g., /ga/). In the study, the authors found no significant difference in the McGurk effect when adults fixated at either the eyes or mouth, and only a slight difference between fixations positioned at the hairline instead of the mouth. Altogether, these findings suggest that direct gaze to the mouth is not required for either optimal speech understanding nor audiovisual integration.

One circumstance in which direct attention to the mouth would be optimal is when listening in a noisy situation. In noisy environments, direct access to redundant audiovisual cues could circumvent the deleterious effects of speech being masked and difficult to understand. Interestingly, while noise increases adults' attention to the mouth, it does not propel them to look exclusively at this region. For example, Vatikiotis-Bateson (1998) examined how eye movements vary as a function of noise level. Attention to the mouth increased as the noise level increased. However, even at high levels, adults only fixated to the mouth 60% of the time. Similarly, Yi et al. (2013) showed that audiovisual speech recognition in noise is unaffected when participants are instructed to focus on the hairline while viewing a single talker speak. These findings are consistent with the implications reported in the previous paragraph by showing that even in adverse listening environments, direct gaze is not required for accurate speech perception. Instead, listeners may be able to use their peripheral vision to capture audiovisual speech cues.

In line with this idea, studies have found that under adverse listening conditions, adults adopt a strategy where they focus on a central vantage point, such as the nose. This strategy may be advantageous by allowing listeners to keep the mouth and eyes in their field of view. For example, Buchan et al. (2007) found that adults fixated more on the nose while performing the emotion and word recognition tasks in noise. On a speech intelligibility in noise task, Buchan et al. (2008) observed longer looks to the nose when background noise was added or when the talker varied across trials. Thus, adults will utilize a strategy where they can maximally capture dynamic cues from both the mouth and eyes.

It is worth noting that these gaze strategies may be culture-specific. In fact, most of these studies recruited native English speakers only, with the exception of the Vatikiotis-Bateson et al (1998) study. When visual speech cues are uninformative in one's native language, gaze pattern seems to differ between cultures. Such is the case between Japanese and English speakers. For example, Hisanaga and colleagues (2015) found that Japanese speakers focused less on mouth in anticipation and during speech events, compared to English speakers. The mouth is less relevant for Japanese speakers because visual information is less useful for speech perception in this language. Whereas English consonants can be divided into 5 or 6 viseme groups, Japanese consonants can only be divided into three. Thus, one's native language can influence individuals direct their attention to capture these cues available on the face.

In summary, the eyes and mouth convey different types of information that can support language processing. Whereas the eyes convey prosodic, affective, identity, and social cues, the mouth conveys linguistic cues. Furthermore, adults are sensitive to the distribution of information across the face and will direct their gaze to the appropriate area to complete a task-related goal. Moreover, while viewing a talker speak, adults tend to display a dynamic gaze

strategy, directing their gaze to specific facial regions at different time points. Finally, adults can capture cues from a particular facial region, without direct fixation to a region.

Face-scanning Behavior in Children during Language Processing

Throughout development, children are faced with the challenge of learning their native language. Children who are both sighted and hearing have access to cues from both the visual and auditory modalities. Seeing a talker's face can support language acquisition. Specifically, the mouth provides redundant cues to the auditory signal. Studies have focused on how infants direct their gaze while viewing someone speak.

Eye-tracking studies have shown that as infants develop a growing interest in language, they focus more on the talker's mouth (Hillaiet de Boisferon et al., 2017; Lewkowicz & Hansen-Tift, 2012; Pons et al., 1982; Tenenbaum et al., 2013). Lewkowicz and Hansen-Tift (2012) found that selective attention to the mouth emerges around 8 months, which coincides with the onset of canonical babbling. Moreover, infants continue to show a preference to the mouth until 12 months of age, where they begin to equally attend to the eyes and mouth. The authors suggested that direct attention to the mouth at 8 months might help infants to acquire the speech sounds found in their native language. Once sufficient expertise with their native language is achieved, infants no longer require direct access to audiovisual speech cues available at the mouth. Instead, they can shift their attention to the eyes to capture social cues. In a longitudinal study, Tenenbaum et al. (2013) also found an increase in attention to the mouth during the second half of the first year of life. In this study, infants saw a woman label and/or look at an object. Whereas attention to the mouth gradually increased between 6 and 12 months, attention to the eyes gradually decreased. In sum, infants show a great deal of attention to the mouth in the

second half of the first year of life. By directly attending to the mouth, infants gain access to redundant audiovisual speech cues which could facilitate language acquisition.

Language expertise is not completely attained by 12 months of age. In fact, children's understanding of their native language continues to develop into adolescence (Marchman & Fernald, 2008; Rigler et al., 2015). As children grow, they must acquire several components of language, including, but not limited to, vocabulary. Some studies have explored children's face-scanning behavior past 12 months of age. Hillairet De Boisferon et al. (2018) found that attention to the mouth re-emerges around 14- and 18-months of age, around the same time children display a growth spurt in vocabulary. Morin-Lessard et al. (2019) found that children's preference to the mouth still persists at the age of 5. Altogether, these findings suggest that children may still show greater attention to mouth, even at later ages. This strategy may help them to develop their lexicon.

Beyond the need to acquire their native language, monolingual infants show a preference the speaker's mouth while listening to an unfamiliar language (Barenholtz et al., 2016; Kubicek et al., 2013; Lewkowicz & Hansen-Tift, 2012; Pons et al., 1982). Both Lewkowicz and Hansen-Tift (2013) as well as Pons, Bosch, and Lewkowicz (2019) found that 12-month-old infants showed a preference to the mouth while listening to a non-native speaker. Similarly, Kubicek et al (2013) found that 12-month-old monolinguals who heard a talker speaking in a non-native language focused on the mouth region while viewing silent-talking videos. In contrast, monolinguals who were exposed to a talker speaking in their native language focused on the eye region while viewing silent-talking videos. These findings have been interpreted to reflect infants' attempt to disambiguate unfamiliar speech. However, it is unclear what infants gain from attending to the mouth while listening to a non-native language.

For bilinguals, direct attention to the mouth may help them overcome the challenge of learning two languages. Compared to monolinguals, bilingual infants show earlier and extended periods of attention to the mouth. Pons et al. (2015) showed that at 4 months of age, bilinguals equally attend to mouth and eyes, while monolinguals of the same age focus solely on eyes. At 12 months of age, bilingual infants still preferentially attended to the mouth, whereas monolinguals focused equally on the mouth and eye region. Additionally, Ayneto and Sebastian-Galles (2017) found that bilingual infants show a bias to look at the mouth, even in non-linguistic contexts. In this study, monolinguals and bilinguals saw videos of dynamic expressive faces. Compared to monolinguals, bilingual infants looked longer at the mouths of expressive faces. The authors suggested that the bilingual experience triggers infants to adopt a strategy that may generalize to non-linguistic scenarios. Collectively, these findings demonstrate the gaze strategy employed by bilingual infants to overcome the challenge of learning two languages.

Interestingly, the degree of perceptual similarity between two languages influences bilingual infants' gaze patterns. Birulés et al. (2019) demonstrated that bilinguals who are learning perceptually similar languages attend more to the mouth than those who are learning perceptually distinct languages. Morin-Lessard et al (2019) also found that, when viewing a Russian speaker, bilingual English and French-speaking children focused equally on the mouth and eyes. However, when viewing an English or French speaker, they focus more on the mouth. It is unclear why children, particularly the monolingual ones, showed a different gaze pattern for Russian than their non-dominant language. According to the authors, this difference in pattern between the two languages could be attributed to the idiosyncrasies of the speaker. For example, children may have focused on the Russian speaker's eye more because she blinked more than the English and French speakers. However, these findings could be attributed to the fact that this

experiment took place in Montreal, where both English and French are spoken. In this setting, children might be exposed to English or French, regardless of the language(s) spoken at home. Thus, attending to the mouth when the language spoken is either English or French might help children to disambiguate which language is being spoken. Nonetheless, these findings suggest that language background modulates bilingual infants' gaze pattern while viewing a talking face.

So far, these studies have revealed that, as infants show emerging interest in language, they begin to shift their attention to the mouth. However, the onset and offset of this preference are still debatable. Moreover, infants' preferences to the mouth also depend on their language experiences. Based on these findings, one might assume that children do not show a dynamic gaze pattern like adults. While this is a reasonable assumption, one should proceed with caution in interpreting these results for several reasons.

First, the analyses used to calculate infants' gaze pattern to a talking face might bias how researchers interpret the results. In the studies described above, infants' attention to a particular region is typically calculated using one of two measurements, proportion of total looking time (PTLT) and dwell time. Whereas the former measures the total time directed to each area of interest (AOI) relative to the total looking time spent on the face, the latter calculates the total time spent looking at each area. For PTLT or dwell time, a zero-difference score or nonsignificant effect of AOI has been interpreted as reflecting equal attention to the eyes or mouth. Conversely, a non-zero difference score or significant effect of AOI has been interpreted as a preference to one AOI. Here, the underlying assumption is that by selectively attending to the mouth, infants are ignoring the eyes. However, upon reexamination of the data in these studies, infants also focus on the eyes, albeit to a lesser extent than the mouth. Thus, infants may

in fact exhibit dynamic gaze patterns. However, the current metrics for face-scanning studies in infants may fail to capture it.

Second, in these studies, infants' eye movements are averaged across the length of the entire video, rather than time-locked to different speech events (e.g., onset of the sound). With the former analysis window, researchers cannot capture the temporal characteristics of infants' eye movements to a talking face. Thus, the current literature cannot assess when children's gaze pattern display adult-like behaviors, such as the eye primacy effect.

Third, the nature of stimuli may also bias infants to focus on the mouth. In most of these studies, stimuli are recorded in a non-interactive setting. That is, they typically record a talker alone rather than in a natural context, such as interacting with an infant or adult listener. When presented with stimuli that were recorded in an interactive setting, 5- and 8-month-olds do show greater attention to the talker's eyes (Smith et al., 2013). This result differs from the developmental trajectory postulated by Hansen-Tift and Lewkowicz (2013), and underscores the importance of using naturalistic stimuli to measure infants' eye gaze to dynamic faces.

Fourth, while selective attention to the mouth might help infants acquire their native language(s), communication is not just a linguistic process. It also consists of social cues, which are distributed across the face. As mentioned in the introduction, the eyes convey information about the social identity and referential intent of the talker. Many studies have demonstrated that infants' ability to capture information available at the eye also supports language acquisition (Brooks & Meltzoff, 2005; Morales et al., 1998; Morales, Mundy, Delgado, Yale, Messinger, et al., 2000; Morales, Mundy, Delgado, Yale, Neal, et al., 2000; Mundy, 1998; Tenenbaum et al., 2015). Using parental reports, Morales et al (1998) found that infants who followed the adult's gaze at 6 months had better receptive and expressive vocabulary at a later age than infants who

followed gaze to a lesser extent. Similarly, Brooks and Meltzoff (2005) found that looks to the intended object and simultaneous vocalizations at 11 months predicted language scores at 14- and 18-months. Altogether, these studies underscore the role of gaze following on language acquisition.

One limitation of gaze following studies is that they provide an indirect measure of children's attention to a particular facial region. Because infants' eye gazes are not measured in these studies, it is unclear whether infants are directly looking at the talker's eyes. Another limitation is that adults in these studies are silent throughout the experiment. Thus, information available from the eyes is not competing with information available from the mouth. Nonetheless, these studies highlight the importance of capturing information available at the eyes.

Two studies have addressed the shortcomings highlighted in the past few paragraphs by using stimuli that contains information available at the eyes and mouth. Fort et al. (2018) explored whether infants in their second year of life can perceive additional nonspeech information coming from the eyes or mouth. In this study, 15- and 18-month-old infants watched a video of a woman talking and performing a nonspeech movement sequentially. The nonspeech movement consisted of the speaker either raising her eyebrows (EB) or protruding her mouth (MP). The authors reported that during speech events, infants attended to the mouth. For the EB condition, both bilingual and monolingual infants at both ages looked at the talker's eyes. Using growth curve analysis, the authors revealed that monolingual infants and bilingual 18-month-old infants look towards the eyes in anticipation of the EB event. This study is one of the few to show how eye movements unfold over the time course of the experiment. It also shows that around the second year of life, infants adopt a strategy that allows them to capture information

from both the eyes and mouth. Similarly, language outcomes are better predicted when both gaze following and attention to the mouth are taken into account than when only one of these processes are considered (Tenenbaum et al, 2015). Altogether these findings suggest that infants can process information from the eyes and mouth.

In summary, eye-tracking studies with infants have implied that infants selectively attend to the mouth as a means of acquiring their native language. However, these studies might be overestimating infants' preferences to the mouth due to their methodological and analytical limitations. Some studies have begun to show that infants are able to capture information from both the eyes and mouth. Altogether, these studies shed light into how infants direct their visual attention while viewing a talker speak.

Neural bases of face-to-face communication

To further understand the relative contribution of the eyes and mouth in facilitating speech perception, it is important to understand the neural basis of communication. Unlike behavioral studies, neuroimaging studies have revealed the brain areas involved in face-to-face communication. Functional magnetic resonance imaging (fMRI) and functional near-infrared resonance (fNIRS) studies have demonstrated that seeing a talker's face activates a network of cortical areas (Calvert et al., 1997, 2000; Dick et al., 2010; Jiang et al., 2012; Pekkola et al., 2006). These areas include primary sensory areas as well as higher association areas, such as the inferior frontal gyrus and the supramarginal gyrus, to name a few. Evidence from event related potential (ERP) studies also confirm that visual speech modulates auditory processing (Besle et al., 2008; Paris et al., 2016; van Wassenhove et al., 2005).

Articulatory movements have been shown to influence auditory processing at different stages (Besle et al., 2008; Brunellière et al., 2013; Calvert et al., 2000; van Wassenhove et al.,

2005). At the phonetic level, visual speech has been shown to modulate activity related to early auditory processing. Electrophysiological studies have found that visual speech reduces activity and timing related to auditory processing at around 100 ms (N1) and 200ms (P2) following the onset of the acoustic signal. The reduced magnitude and faster latency has been presumed to reflect predictive coding of the upcoming speech signal. Besle and colleagues (2008) also found that lip movements activate the second auditory cortex 150ms after the acoustic signal. These visual activations occur before activations of other parts of the brain, suggesting that lip movements modulate auditory processing through a direct feedforward process. At the semantic level, Brunelliere and colleagues found that visual speech might also impacts lexical processing, as evidenced by the attenuation of the N200 peak in a sentence context. Moreover, the audiovisual modality increased the amplitude of the late part of the N400, known to reflect word processing at the lexical level.

Recently, studies have also examined the neural effects of social cues, specifically eye contact, on communication. Hernández-Gutiérrez et al. (2018) found that seeing the whole face of a talker increased the amplitude of the late component of the N400. Moreover, compared to the whole face, occluding either the eyes or mouth attenuated the amplitude of the late component. Interestingly, the authors only found a significant difference in amplitude between the whole face and covered mouth condition. These findings suggest that both the eyes and mouth contribute to sentence processing.

Dual brain studies using fMRI or fNIRs have also elucidated how eye contact during face-to-face communication influences neural activity. These studies typically involve scanning two people simultaneously as they make eye contact with one another or look at still image of a person. One of the advantages of this technique is that one can examine cross-brain effects. In

other words, one can see whether there is synchrony between brain regions as two people make eye contact. Some studies have shown that eye-to-eye contact activates a system of cortical areas, including frontal temporal and parietal areas. In fact, eye-to-eye contact activates areas involved in expressive and receptive language. Moreover, eye-to-eye contact increases the synchrony or coherence between brain regions involved in expressive and receptive language, such as left superior temporal gyrus. These findings suggest that eye contact may prime brain regions that are typically involved in communication. It is unclear what the nature of the coherence entails. These brain regions can be coherent to rapidly process socially, relevant social cues or it can function to process this information simultaneously. One limitation of these studies is that the stimuli consist of still images. Jiang et al. (2012) also examined neural features of face-to-face communication in a naturalistic context. The authors found that synchrony in the left inferior frontal cortex when interacting partners are facing each other and taking turns in a conversation than when their backs are to each other or when only one person is speaking. Altogether, these findings suggest that nonverbal social cues are also associated with specialized language networks.

Collectively, these neuroimaging studies highlight the relative contribution of the eyes and mouth in facilitating communication. Whereas articulatory movements speed up auditory processing, the eye region primes brain areas involved in communication. However, these studies do not provide a clear link between listeners' gaze distribution to a talking face and neural correlates of face-to-face communication. Evidence from behavioral studies suggest that adults can capture audiovisual information without fixating directly on the talker's mouth. On the other hand, evidence from electrophysiological studies seem to imply that articulatory movements are responsible for the attenuation of the N1 and P2 components of the auditory

ERP. If listeners can efficiently process speech without looking at the mouth, then visual speech should still modulate early auditory processing under these interactions.

One study sought to address this question. Kaplan and Jesse, (2019) examined whether visual speech can facilitate early auditory processing when listeners fixate to the talker's eyes. To ensure that fixations were maintained at the eye region, participants were trained to always fixate to a rapid serial visual representation (RSVP) of abstract shapes at the eye region. Using ERP measurements, the authors found a reduction in N1 and P2 amplitudes for audiovisual compared to audio only speech. Thus, listeners can obtain sufficient visual speech information while fixating on the eyes.

Future directions

While researchers have examined how individuals direct their visual attention while viewing a talking face, there are still questions that remained unanswered.

- At what age do children begin to show adult-like gaze patterns to a talking face? Most studies examining infants' visual attention to a talking face focus on infancy and early childhood. However, children are still learning language into adolescence. Thus, it is unclear how adolescents direct their gaze to the talking face. By focusing on later ages, studies may begin to shed light into how children's gaze strategy changes throughout development.
- How do children direct their gaze to a talking face during online language learning? Using parental reports, several studies have found a relationship between infants' attention to the mouth or gaze following abilities and later language outcome. Little is known whether infants who direct their attention to the mouth acquire language better.

By addressing this question, studies will uncover the role of audiovisual processing on language acquisition.

- What is the neural mechanism underlying expressive talking faces? Behavioral studies have reported hemispheric differences in processing facial expressions of emotions. Thompson, Malloy, and LeBlanc (2009) found that a high degree of emotional prosody directs viewers to the right side or eye region of the talker's face. Conversely, a neutral prosody directs views to the left side or mouth region of the talker's face. One explanation for this finding is that observers orient their attention to the side of the face that is contralateral to the primary activated hemisphere. Attention to the eyes increases activation of the right hemisphere, which is specialized for emotion processing. Further research is needed to connect listeners' gaze behavior to the neural correlates of emotion processing.
- Does varying the gaze location away from the eye influence emotion processing? Studies have shown that direct attention to the mouth is not necessary for perceivers to capture audiovisual speech information. However, it is unclear whether efficient emotion processing also possible without directly attending to the eyes. Such findings will reveal whether overt attention to the eyes is necessary for emotion processing.

Conclusion

The purpose of this review was to provide insight as to why a dynamic gaze strategy facilitates successful language processing. Eye-tracking studies with adults show that adults are sensitive to the distribution of different information across the face. As such, they will utilize a strategy where they can capture information from the eyes and mouth. Under adverse listening conditions, adults fixate on the talker's nose as a means to capture dynamic cues from both the

mouth and eyes. More importantly, adults can sufficiently capture audiovisual speech information without directly attending to the mouth.

Eye-tracking studies with children, on the other hand, suggest that direct attention to the mouth facilitates language acquisition. However, the methodological and analytical approaches used in developmental studies might be overestimating infants' preference to the mouth. Studies addressing these shortcomings have begun to show that both the eyes and mouth facilitate language acquisition.

Finally, neuroimaging studies have revealed the relative contribution of the eyes and mouth on speech communication. Whereas articulatory movements speeds up auditory processing, the eye region primes brain areas involved in communication.

Chapter 3

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

Jasenia Hartman^{1*}, Jenny Saffran² and Ruth Litovsky^{1,3}

¹Neuroscience Training Program, University of Wisconsin-Madison; Madison, WI 53706

²Department of Psychology, University of Wisconsin-Madison; Madison, WI 53706

³Communication and Science Disorders, University of Wisconsin-Madison; Madison, WI 53706

*Corresponding author – Email: jhartman3@wisc.edu

Under review for *Ear and Hearing*

ABSTRACT

Objectives: Although cochlear implants (CI) allow people who are deaf to develop spoken language skills, many CI listeners experience difficulty learning new words. This difficulty may be due to the challenges listeners face in perceiving speech sounds. Two factors have been shown to improve word learning in listeners with normal-hearing listeners: 1) learning from different talkers (talker variability) and 2) availability of visual speech cues. Our study addressed the question of whether talker variability and audiovisual speech cues would improve word learning in adult CI listeners.

Design: 18 adults with CIs and 10 adults with NH learned 8 novel word-object pairs spoken by a single talker or six different talkers (multiple talkers). The word learning task consisted 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: Talker variability did not enhance word learning in adult listeners with either CIs or NH. Both groups performed near ceiling during the test phase, regardless of whether they learned from the same talker or from different talkers. However, compared to listeners with NH, CI listeners eye gaze focused significantly more on the talker's mouth while learning the words. Conclusions: For adult CI and NH listeners, talker variability is not required to facilitate word learning. However, unlike NH listeners who can successfully learn words without focusing on the talker's mouth, CI listeners tend 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.

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 people who are deaf, cochlear implants (CIs) not only grant listeners access to the auditory world, but also offer the opportunity to integrate auditory and visual cues. 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 (Stevenson et al., 2017; Rouger et al., 2008; Tremblay et al., 2010).

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 by 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, audio-visual 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 implant device itself. Whereas the normal hearing (NH) system consist of dozens of independent auditory filters, the CI system has up to 22 electrodes, with approximately 8 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 (Havy et al., 2013; Houston & Miyamoto, 2010; Houston et al., 2012, Pimperton et al., 2018). Even post-lingually deafened CI adults show wide variability in language skills (see

Peterson, Pisoni, & Miyamoto, 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 (Giezen et al., 2010; Munson et al., 2003), likely because of the degraded nature 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 (Desai et al., 2008; Iverson, 2003; Munson & Nelson, 2005b). 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. 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, Nazzi, & Bertoncini, 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, Escudero, & Baker (2015).

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, the methods used to study word

learning in CI listeners might also exacerbate these perceptual challenges. CI listeners are typically exposed to repetitions of a single token of the word spoken by the same person. Such methods may distort listeners' perceptual space towards acoustic dimensions that are irrelevant in distinguishing 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 on learning

Studies with NH listeners suggest that variability plays an essential role in learning categories (Gomez, 2002; Perry et al., 2010; Posner & Keele, 1968), and, of particular relevance to word learning, in augmenting phonetic categories (Lively et al., 1992; Quam et al., 2017; Rost & McMurray, 2009, 2010). 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 extends to adults. Lively et al (1993) found that Japanese native speakers learning English as a second language were able to learn familiar words with /r/-/l/ contrasts after learning from different talkers compared to 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 weight 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 9 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 in order 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 (Hillaret De Boisferon et al., 2018; Lewkowicz & Hansen-Tift, 2012; Tenenbaum et al., 2013; 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 (Buchan et al., 2008b; Król, 2018; Munhall, 1998) or hearing sentences spoken by different talkers (Buchan et al., 2008a). While 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 towards 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 towards 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 weighting 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 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, and (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).

METHOD

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 one year of CI experience. This group consisted of 16 bilateral CI users, 1 unilateral CI user, and 1 hybrid CI user (acoustic + electric hearing in the CI ear). All participants had Cochlear Ltd CIs (Sydney, Australia). One CI listener was excluded from the analyses due to the inability to track their eye movements. 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 25dB HL for octaves between 250 and 3000Hz and no greater than 40dB HL at 4000Hz (ANSI, 1989). Two NH listeners were excluded from analysis due to failure in passing the hearing screening.

| ID | Sex | Age | Onset of Deafness (yrs) | Duration of HL(yrs) | Years w/ 1 st CI | Years w/ 2 nd CI | Device | Etiology | CNC scores |
|------------------|-----|-----|-------------------------|---------------------|-----------------------------|-----------------------------|---------------------|------------------------|------------|
| IDH | M | 20 | 3 | 1.5 | 15 | 14 | Nucleus 6 | Unknown | 96 |
| IDM | F | 42 | 5 | 28 | 9 | 7 | Nucleus 7 | Unknown | 74 |
| IBZ | F | 52 | 38 | 1 | 14 | 12 | Nucleus 6 | Unknown | 82 |
| IDA | F | 52 | 8 | 38 | 6 | 5 | Nucleus 6 | Unknown | 84 |
| ICP | M | 56 | 4 | 42 | 10 | 7 | Nucleus 7 | Unknown | 32 |
| IDJ | F | 58 | 45 | 8 | 5 | 5 | Nucleus 6 | -- | 76 |
| ICI | F | 61 | 46 | 4 | 11 | 10 | Nucleus 6 | Unknown | 54 |
| ICM | F | 63 | 23 | 34 | 9 | 7 | Nucleus 6 | Unknown | 88 |
| IDK ^a | M | 64 | 16 | 34 | 14 | -- | Nucleus 6 | Otosclerosis | 88 |
| IDL ^b | F | 65 | 33 | 28 | 4 | 3.5 | Nucleus 6 | Unknown | 74 |
| IBF | F | 66 | 38 | 16 | 14 | 12 | Nucleus 7 | Hereditary | 84 |
| ICY | M | 66 | -- | -- | 4 | 4 | -- | -- | 74 |
| IAU | M | 70 | 3 | 46 | 21 | 14 | Nucleus 6 | Unknown | 53.1 |
| ICJ | F | 70 | 25 | 35 | 10 | 10 | Nucleus 6 | Hereditary | 70 |
| IAJ | F | 73 | 12 | 38 | 23 | 16 | L:Nucleus 6/R:Kanso | Unknown | 70 |
| ICC | F | 74 | 9 | 52 | 13 | 11 | Nucleus 7 | Congenital Progressive | 82 |

Table 1. Demographics of CI listeners

All participants had normal or corrected vision, and achieved a typical score of 26 or above (Goupell et al., 2017; Nasreddine et al., 2005) on the Montreal Cognitive Assessment test for cognitive function, except for 1 CI participant who obtained a score of 25. However, we included this participant 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.









| Objects |  |  |  |  |  |  |  |  |
|----------|--|--|--|--|--|---|--|--|
| Word | /dita/ | /gita/ | /foma/ | /voma/ | /nodi/ | /lodi/ | /pibu/ | /tibu/ |
| Word Set | word set 1 | | | | word set 2 | | | |

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

Speech stimuli consisted of 8 novel words: /dita/, /gita/, /foma/, /voma/, /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 (4 males, 4 females) raised in the Midwest. Multiple tokens of each word were recorded and a single token of each word was selected from each speaker.

Audio/visual speech stimuli were videorecorded with an iPad Air Pro (30 frames/sec, resolution of 1920 x 1080). Each talker was filmed against a solid background. A microphone was placed 8 inches away from the talker to record the audio (44.1kHz sampling rate). Audio recorded from the microphone was processed using Adobe Audition and replaced the original audio recorded from the iPad Air Pro. 400ms of silence was added before the onset of the word and 300ms 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. Audio was synchronized to videos using Adobe Premiere. Mean length of video was 1015ms (range:1000ms-1027ms). Audio was scaled to 55dB on a A-weighting scale using a sound level meter.

Word-object pairs

Each novel word was paired with a novel object (8 word-object pairings; see Figure 1). There were 2 sets of 4 novel-word object pairings. Set 1 consisted of the items /dita/, /gita/, /foma/, and /voma/. Set 2 consisted of the items /nodi/, /lodi/, /pibu/, and /tibu/. Each set was assigned to a learning condition (single vs. 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.

Apparatus

Participants were tested in a double-walled sound booth (Acoustic System, Tx, USA). Participants sat at a table with a 19-inch LCD monitor (1,280 by 1,240 pixels). Eye gaze was tracked with the EyeLink SR 1000 eye-tracker (SR Research, Kanata, ON, Canada) at a sampling rate of 1,000 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, USA). The Psychophysics Toolbox (v3.0.14) was used to maintain the synchronization of audiovisual stimulus presentation with 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 9 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 (3 males, 3 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 2000ms, a video of the talker labelling the object appeared in the center of the screen. The video and image remained on the screen for 1000ms before disappearing. In the single-talker condition, each of the 4 objects was labelled 6 times, by a single speaker, for a total of 24 trials. For the multiple-talker condition, each of the 6 talkers labelled each of the 4 objects once, for a total of 24 trials. The labelling 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 following each learning phase (see Figure 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., /dita/ vs. /voma/). Hard trials were defined as target and distractor labels that served as minimal pairs (e.g., /dita/ vs /gita/). 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). 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 4 times, for a total of 24 trials.

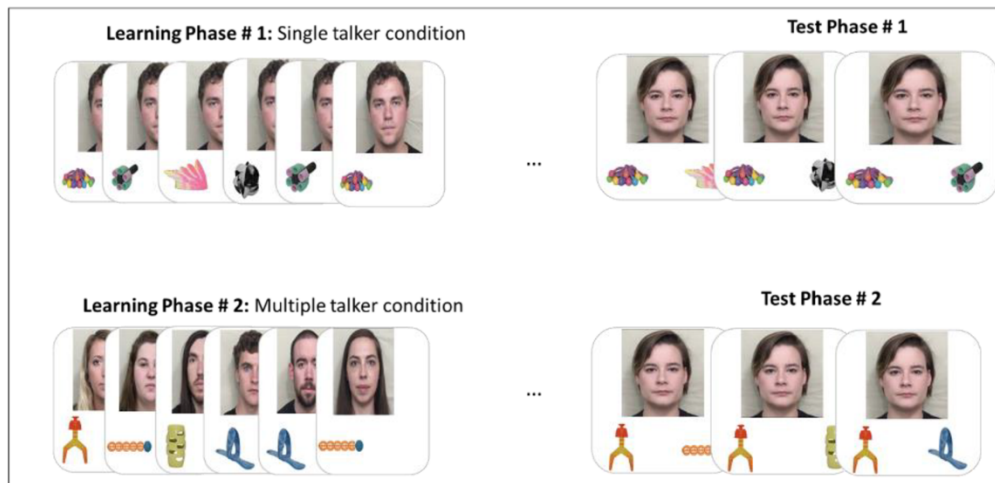


Figure 2. Experimental paradigm. Learning phases were presented before each test phase. In the learning phases, participants saw and heard the label of a novel-object from a same talker (single-talker condition) or from different talkers (multiple-talker condition). Objects were presented one at a time.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

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, 4 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 areas of interest (AOIs). Depending on the video, the rectangle for the mouth AOI was extended by 37-53 pixels horizontally and vertically to account for the talker speaking.

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, proportion of looks to the mouth was calculated as proportion of looks to the mouth relative to the total looks to eyes and mouth during the time window of 0 to 800ms from the onset of the target word. This analysis window was chosen to account for listeners gradually increasing their 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 the test phase, mean accuracy was calculated as proportion of time spent looking at the target object out of the total time spent looking at either of the two objects during a critical window of 300 to 1800 ms following onset of the target word (Fernald et al., 2008). For each measurement, trials were excluded if the participant was not fixating to any AOIs (objects, mouth, and eyes) for more than 50% of the critical window. 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.

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.

For the test phase, the time course of looks to the target object, as illustrated by frame-by-frame data, provides a fine-grained measure of language processing. 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) after the onset of the target word. To analyze this data, we collapsed the data within the critical time window of 300-1800 ms after word onset to examine the proportion of looks to the target relative to looks to the target and distractor (mean accuracy). This analysis is consistent with standard eye gaze-based measurements of word learning (Fernald et al., 2008).

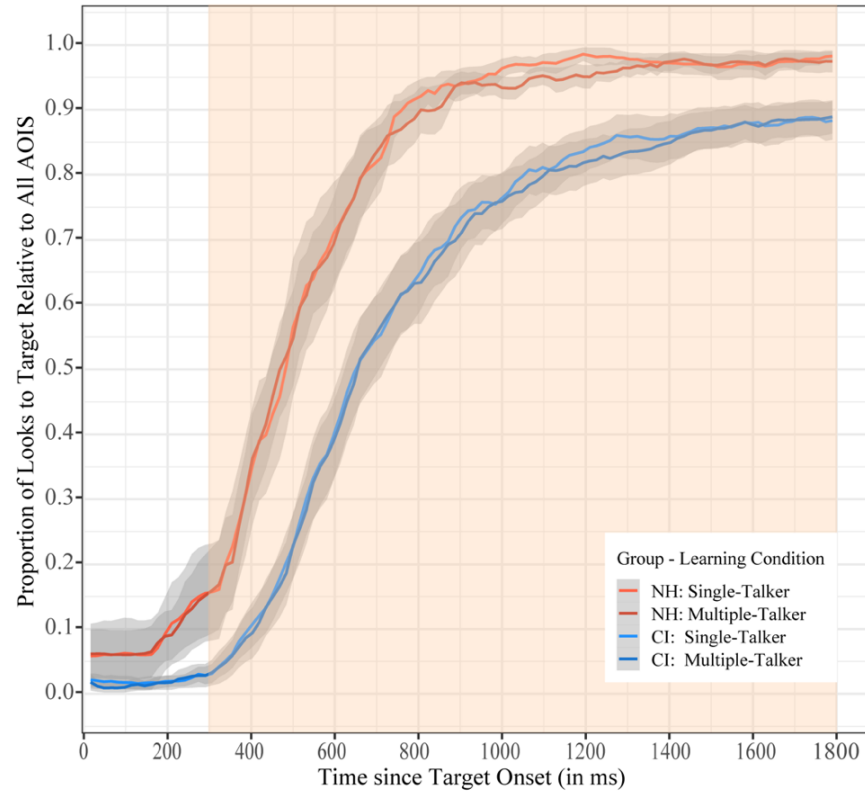


Figure 3. Time Course of Fixation to the Target by Learning Condition and Hearing Status for Test Phase Trials. Proportion of looks to the target relative to the total looks to all AOIs (target, distractor, mouth, and eyes) for NH and CI listeners. Data represents test trials following the single-talker or multiple-talker learning condition. Data is averaged across trials. Shaded box represents time window of analysis. Gray ribbons around lines indicates ± 1 SD.

We analyzed mean accuracy using linear mixed model effects. All analyses were conducted in R using the package lme4. We first conducted a model that regressed mean accuracy (defined as proportion of looks to target divided by looks to target and distractor within the critical time window of 300ms to 1800ms) on the fixed effects of training condition, test difficulty, and group. We also included two interaction terms, training condition x group and test difficulty x group, as well as by-subject random intercept and by-subject random slope for training and test difficulty. After this model resulted in a singular fit, we reduced the random

effects structure by removing the by-subject random slope for test difficulty.¹ 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.

As seen in Figure 4, listeners with CIs and listeners with NH performed significantly above chance [$b = 0.39$, $T(25.82) = 22.34$, $p < .0001$]. Contrary to our predictions, there was not a main effect of training on single vs. multiple talkers [$b = .004$, $F(1, 24.67) = .124$, $p = 0.8$] nor an interaction effect between training and group. Thus, learning from multiple talkers did not enhance word learning in CI or NH adult listeners (Fig 5). There was a significant main effect of test difficulty [$b = -0.10$, $F(1, 1211.78) = 24.72$, $p < .0001$] and group [$b = .07$, $F(1, 25.56) = 4.45$, $p < .05$].

¹ Final Model: Accuracy \sim training + testdifficulty + group + training*group + testdifficulty*group + (1+training|SubID)

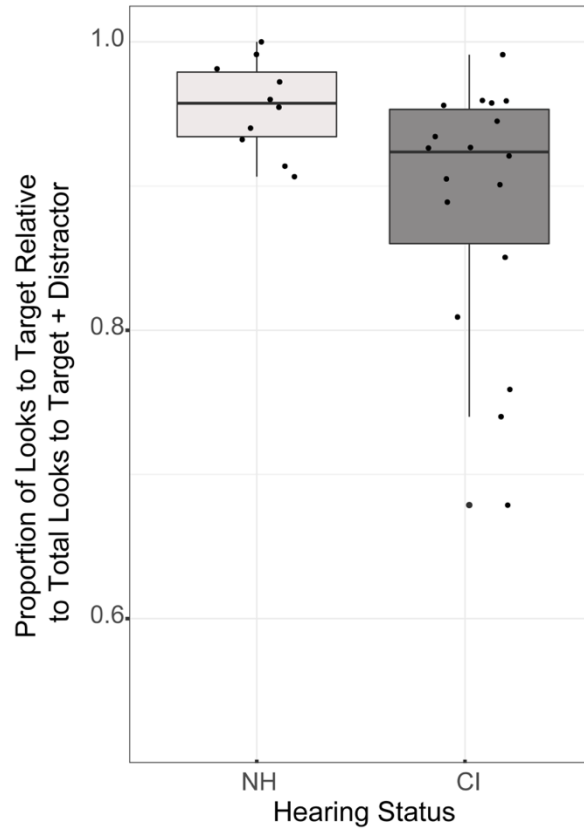


Figure 4. Proportion of looks to the target relative to the total looks to the target and the distractor objects during the critical time window for NH and CI groups. Data represents the proportion during the test phases. The dark line represents the median. The upper hinges represent the first and third quartile (i.e., 25th and 75th percentiles). Data points represent the proportion for each participant.

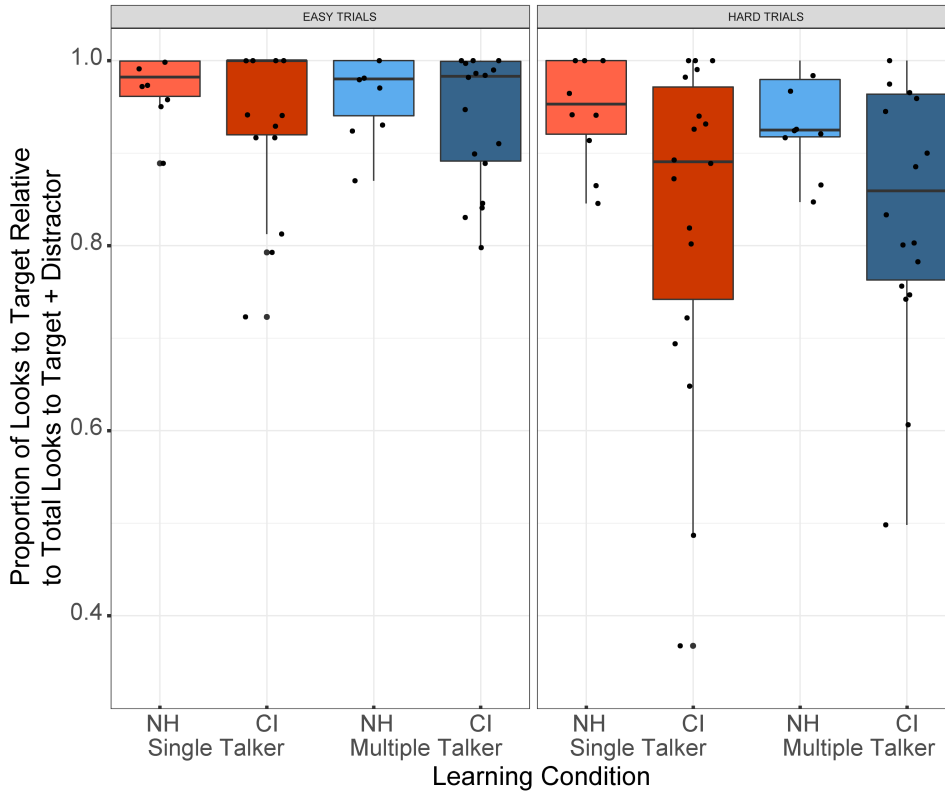


Figure 5. Proportion of looks to the target relative to the total looks to the target and distractor objects during the critical time window for NH and CI groups. Data represents the proportion for the easy and hard test trials after learning from a single-talker (red) or multiple talkers (blue). The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

Next, we collapsed the data by test difficulty and learning condition to examine whether the extent to which talker variability improves word learning in CI listeners is modulated by the phonological similarity between the words. We predicted a larger difference in accuracy for similar sounding words (Hard items) compared to phonologically-distinct words (Easy items). Using the same regression approach, we found a significant interaction between test difficulty and group [$b = .07$, $F(1, 1211.78) = 7.99$, $p < .01$]. In particular, for CI listeners, performance was higher [$b = 0.10$, $t(1214.1) = 6.57$, $p < .0001$] on Easy trials [$\mu_{1/2} = 98.6\%$;

range = 70.6-100%] compared to Hard trials [$\mu_{1/2}$ = 89.9%; range = 33.9-100%]. Moreover, listeners with CIs were less accurate than NH listeners on both Easy trials [$\mu_{1/2}$ = 98.4%; range = 86.7-100%; b = -0.10, $t(36.8)$ = -3.22, p = .01] and Hard trials [$\mu_{1/2}$ = 93.9%; range = 85.9-100%; b = -0.13, $t(37.3)$ = -4.10, p = .001]. We also analyzed the test phase data after using rationalized arcsine transformation (RAU) to take into consideration possible ceiling effects. Analysis of the transformed data rendered the same significant effects as the untransformed data, with the exception of a main effect of group.²

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 remains constant.

As seen in Fig. 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 analyze these patterns of results, we collapsed the data across the critical time window (0 to 800ms following the onset of the target word) and examined the proportion of looks to the mouth relative to total looks to mouth and eyes. This proportion 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.^{2,3} We included a by-subject random intercept and a by-subject random slope for learning condition. For at least one learning phase, three participants (2 NH, 1 CI) attended to the video on three or fewer trials. Therefore, they were excluded from analysis. Proportion of looks to the mouth were significantly higher [$b = 0.27$, $F(1,23.03) = 6.99$, $p < .05$] 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 Fig. 7. The main effect of training condition almost reached significance [$b = -.02$, $F(1, 23.26) = 4.06$, $p = .056$]. Interestingly, the effect of training reached significance [$F(1, 23.26) = 4.69$, $p < .05$] after RAU transformation. Additionally, contrary to our prediction, the interaction between learning condition and hearing group was not significant [$b = -0.12$, $F(1,24.71) = 3.06$, $p = .09$]. 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 Fig. 8. Altogether, these results suggest that CI listeners rely more on visual speech information than NH listeners when presented with audiovisual speech information. However, the proportion of looks to the mouth was unaffected by the number of talkers.

² After our initial transformation still failed to meet the assumption of non-normality, we transformed the original data using a binary transformation. Data was analyzed with a generalized linear mixed model applied using the glmer function in R. Because the analyses yielded similar findings regardless of whether the data were untransformed or transformed, the untransformed data are being reported.

² Model: Mouth ~ group + training + group*training + (1+training|SubID)

³ The data from the learning phase also violated the assumption of non-normality. Thus, the same transformations from the test phase were also conducted for analysis of the learning phase and revealed the same yielded same significance effect as the untransformed data.

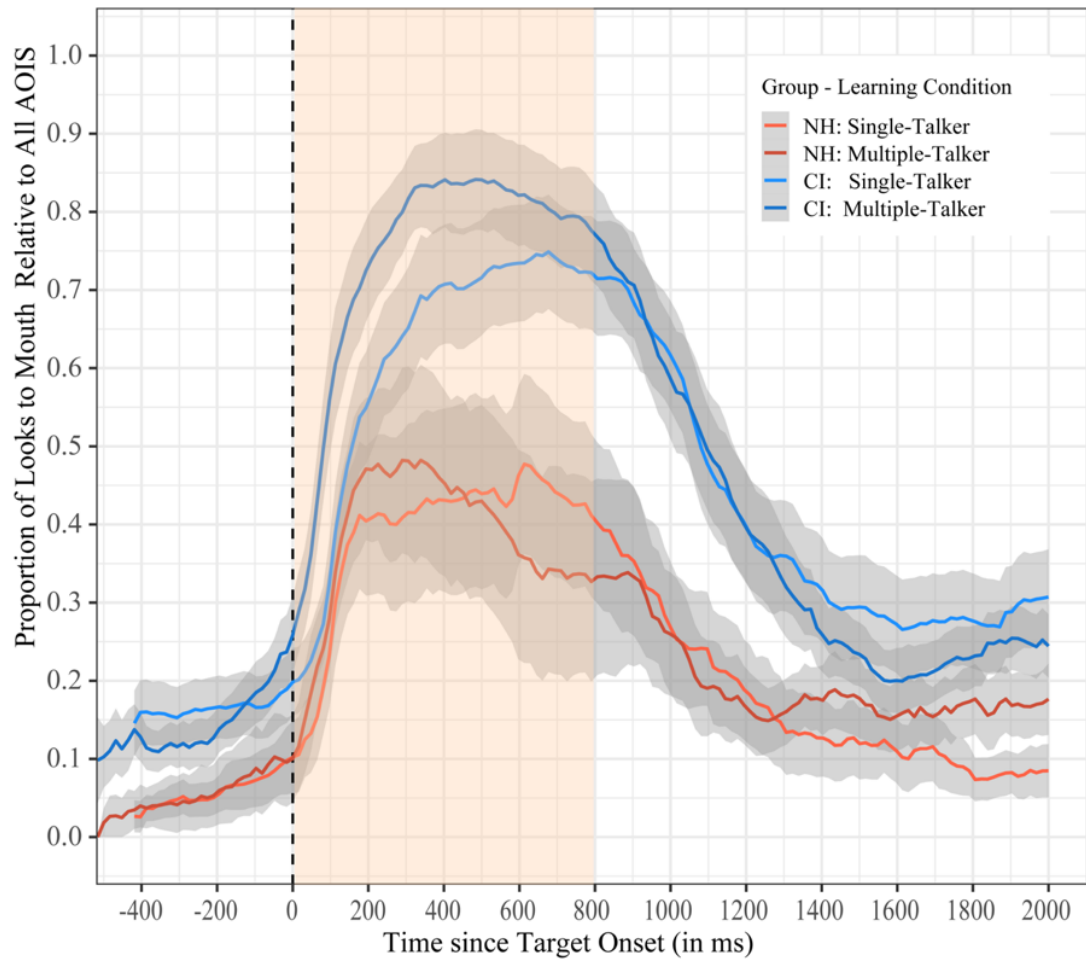


Figure 6. Proportion of looks to the mouth relative to the total looks to all AOIs (target, mouth, and eyes) for NH and CI listeners during the single-talker and multiple-talker training trials. Data is averaged across trials. Shaded box represents time window of analysis. Ribbons around lines indicate ± 1 SE

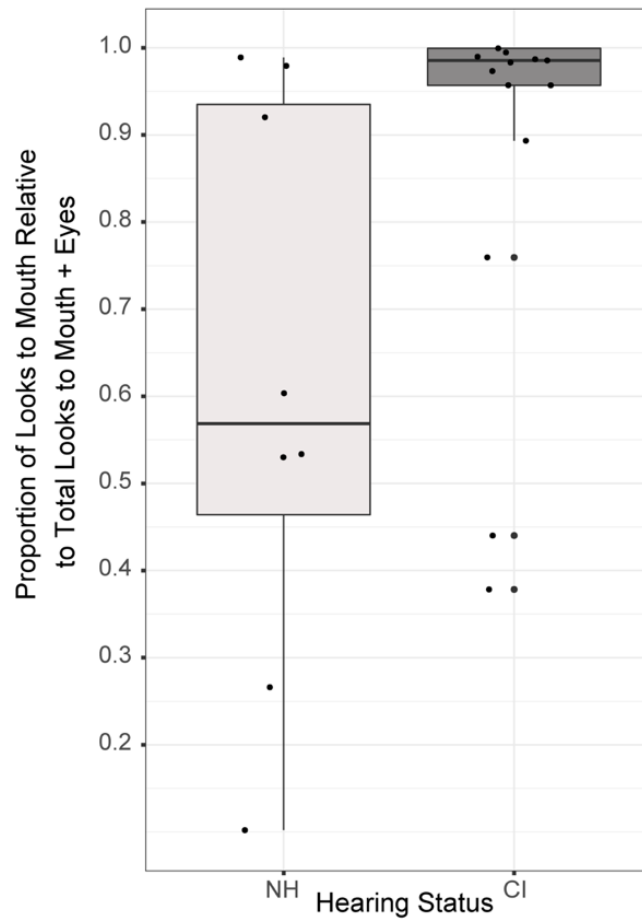


Figure 7. Proportion of looks to the mouth relative to the total looks to the mouth and eyes during the critical time window for NH and CI groups. Data represents the proportion during the training phases. The dark line represents the median. The upper hinges represent the first and third quartile (i.e.,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

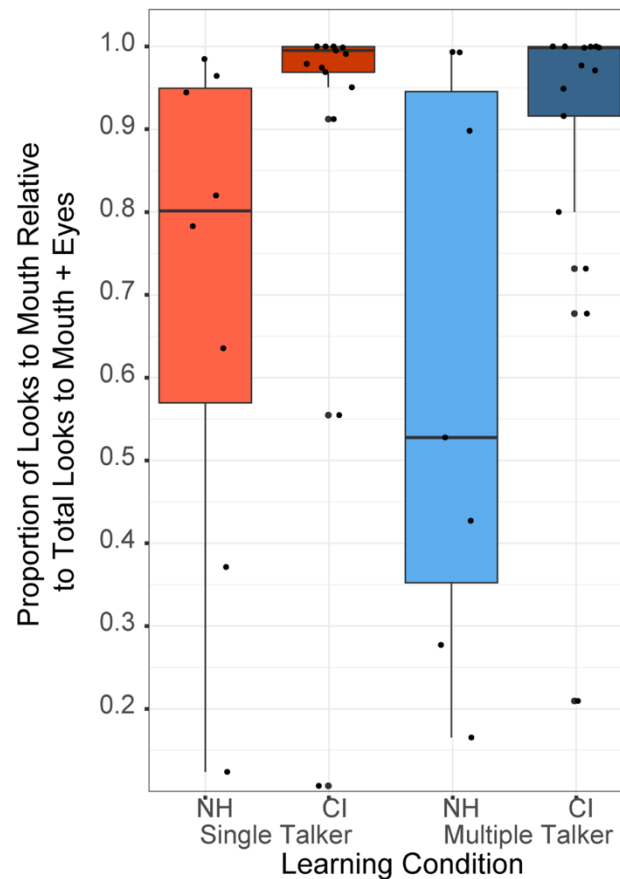


Figure 8. Proportion of looks to the mouth relative to the total looks to the eyes and mouth during the critical time window for the NH and the CI group in the learning conditions. Data represents the proportion during the training phases. The dark line represents the median. The upper hinges represent 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.

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. To test this hypothesis, we conducted a linear regression. Mean accuracy (proportion of looks to the target relative to looks

to target and distractor) was regressed on proportion of looks to the mouth during the training phase and group, as well as an interaction term of proportion of looks to the mouth and group. There were no main effects of training phase [$b = .07$, $F(1, 22) = 0.16$, $p = 0.70$] or group [$b = 0.11$, $F(1,22) = 3.01$, $p = 0.10$], nor an interaction effect of proportion of looks to the mouth during training and group [$b = -.06$, $F(1,22) = 0.19$, $p = 0.67$]. That is, attention to the mouth during training did not predict accuracy in identifying the target object during test for either hearing group.

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. Our results revealed that contrary to our prediction, talker variability did not improve word learning for CI or NH listeners. Additionally, as expected, overall performance on the word-learning task was higher for NH listeners compared to CI listeners. Finally, 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

Prior studies have found that CI listeners experience difficulty acquiring spoken words. This difficulty has been attributed to their perceptual issues in identifying the speech sounds that make up the word. Studies with NH and CI listeners have shown that talker variability improves learning of speech contrasts (Logan & Pisoni, 1995; Miller et al., 2016; Rost & McMurray, 2009; 2010; Zhang et al., 2021). Through variability, listeners are able to tune into the relevant acoustic dimension that distinguishes the perceptual categories. However, our results show that talker variability did not enhance word learning in adult listeners with CIs or with NH. This

finding differs from prior studies with NH listeners (Rost & McMurray, 2009, Logan & Pisoni, 1995) and with CI listeners (Miller et al, 2016; Zhang et al., 2021).

The discrepancy between our results and prior studies may be due to several reasons. First, it might be due to a difference in the population tested. Prior studies examining the role of talker variability on word learning have focused on children (Rost & McMurray, 2009) and second language learners with NH (Davis et al., 2015; Logan & Pisoni, 1995). These populations are less proficient speakers of English, and may require acoustic variation to develop robust phonetic categories. 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 thus likely highly proficient speakers of English. CI listeners tested here might have already developed robust phonetic representations that facilitated word learning and encoding. Thus, talker variability may be especially helpful for establishing new phonetic categories and early word learning, but less useful for later word learning, even given the challenges of learning via a CI. Still, our results were surprising given that high variability training has been shown to improve phonetic categorization for post-lingually deafened CI listeners (Miller et al, 2016). One explanation for the discrepancy between our results and the Miller et al study may be due to the nature of the task. In our study, CI listeners received both acoustic variation and visual speech. In contrast, in the Miller et al (2016), listeners were trained and tested with solely auditory input. It is possible that, in our study, the benefit of talker variability might have been washed out by the presence of visual input. Because CI listeners relied heavily on visual speech cues in the current study, they might not have needed to utilize the variation in the auditory input to support their learning. Thus, future studies should investigate whether talker variability improves word learning in CI listeners when provided with solely auditory input.

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 (Havy et al., 2013; Giezen et al., 2010). However, prior studies have focused on infants and school-aged children with CIs. To our knowledge, our study is the first to show that this difficulty in learning minimal pairs persists into adulthood for CI listeners. Interestingly, exploratory analysis on hard test trials revealed that talker variability yielded a 13% increase in performance for the /dita/-/gita/ word pair, a phonetic contrast that differs in place of articulation. This result corroborates with the Miller et al (2016) study demonstrating that high variability training improves CI listeners' sensitivity to this same phonetic feature. These results are encouraging because CI listeners often confuse words that differ by place of articulation. Thus, talker variability might be beneficial for helping CI listeners to learn difficult phonetic contrasts, such as place of articulation.

One limitation of the current study is that it did not assess how much variability is sufficient to improve perceptual learning. In our study, listeners were exposed to six talkers, whereas prior studies (Miller et al, 2016, Zhang et al, 2021) used an adaptive approach. In these studies, CI listeners were initially trained with two talkers and then took an identification quiz. If listeners scored 90% correct on the quiz, additional talkers were added to the training block. Based on this procedure, it is possible that listeners may show improvements in phonetic categorization with low variability during training. Future studies should assess how much variability is sufficient to improve word learning outcomes in CI listeners.

Another limitation is that when we transformed the data using RAU, the main effect of group became insignificant. It is possible that aggregating the proportion of looks to the target

across the time window might not have allowed us to capture subtle differences between the NH and CI group. In our future study, we will use growth curve analysis to quantify differences in the shape in the time course of listeners' fixations to the target.

Audiovisual speech processing in CI listeners

Our finding that CI listeners direct their gaze to the talker's mouth more than NH listeners is in line with 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). Particularly important is that 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 (Król, 2018; Munhall, 1998; Vaitikiotis-Bateson et al., 1998) whereas 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 have a "noisy" representation of the auditory 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 adapt a strategy of attending to the talker's mouth in order to access speech. These results make sense given the mouth is the primary source of linguistic information.

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. (2008b) 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 word recognition task in which participants had to repeat back the sentences spoken by the talker. Thus, differences in findings between the current study and the Buchan et al. study may be due to the nature of the task used in each study. However, the effect of training on attention to the mouth should be interpreted with caution given that the RAU transformation of the learning phase data yielded a significant main effect of training. Future studies should increase the sample size to assess whether there is an effect of training on attention to the mouth.

Although CI listeners performed less accurately on the word learning task than NH listeners, the CI group was still highly successful in learning the word-object pair associations. It is interesting to note the strategy they used to reach high performance differed from NH listeners. Whereas NH listeners were able to successfully learn the words with minimal attention to the mouth during the learning phase, CI listeners relied on the visual speech cues as they were

learning the words. Interestingly, we did not find a relationship between CI listeners' accuracy during the test phase and the proportion of looks to the mouth during the learning phase. In this study, we allowed listeners to freely view the screen rather than constraining their eye gaze to one particular part of the screen. Thus, future studies should examine whether moving CI listeners' gaze away from the mouth impacts language processing.

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 by 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. Our results suggest that learning from multiple talkers might not help adult CI listeners retune their attention to relevant contrastive acoustic dimensions. Thus, exposing adult CI listeners to multiple talkers might not improve their word-learning abilities. Although we did not see any benefits of talker variability in adult CI listeners, this paradigm might be beneficial for children with CIs, as phonetic categories are still developing throughout childhood. 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 led 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.

CONFLICT OF INTEREST STATEMENT

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest in the subject matter or materials discussed in this manuscript.

FUNDING STATEMENT

This work was funded by the National Institute of Health Grant R01DC00308 (awarded to Ruth Y. Litovsky) and the National Science Foundation GRFP-DGE-1256259 (awarded to Jasenia Hartman). 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).

ACKNOWLEDGEMENTS

We would like to thank the following people: Ron Pomper and Daniel Bolt for help in statistical analysis; and Won Jang for assistance in programming.

Chapter 4. The Role of Talker Variability and Visual Speech on Word Learning in Noise in Normal-Hearing Adults

INTRODUCTION

Word learning involves mapping sound to meaning. To successfully acquire spoken words, listeners must accurately perceive the speech sounds that make up the word forms. Several studies have shown that learning can be facilitated by both auditory cues, such as acoustic variability (Gomez, 20002; Lively, Logan, & Pisoni, 1992; Perry et al., 2010; Quam et al., 2017 Rost & McMurray, 2009; 2010) as well as visual cues such as viewing a talker speak (Tenenbaum et al., 2015; Tsang et al., 2018; Weatherhead et al., 2021).

When learning new words, individuals must develop a well-defined category that can accommodate for surface-level changes, such as the word being said by a different talker. Several studies have shown that talker variability leads to robust categorical learning and generalization to new instances (Gomez, 20002; Lively, Logan, & Pisoni, 1992; Perry et al., 2010; Quam et al., 2017; Rost & McMurray, 2009; 2010). For example, both infants (Quam et al., 2017; Rost & McMurray, 2009; 2010) and second-language learning adults (Lively, Logan, & Pisoni, 1992) succeed in learning lexical categories when exposed to multiple talkers compared to a single talker. Talker variability is beneficial for learning because it allows listeners to tune into the relevant acoustic dimension that contrast the words to be learned, while ignoring the irrelevant ones. Moreover, it is the variation in the noncontrastive cues that helps listeners to tune into the relevant acoustic dimension. For example, Rost and McMurray (2010) assessed whether infants could successfully learn minimal pairs, /buk/ and /puk/, when variation was available either in the contrastive acoustic cue (in this case, voice-onset time, VOT) or along the noncontrastive cues (in this case, pitch). The authors found that infants were able to

successfully learn minimal pairs when variation occurs along the non-contrastive acoustic dimension, but not the contrastive one. Additionally, Quam et al (2017) found that infants do not succeed in learning minimal pairs when male and female talkers said different words. These findings demonstrate that variability along the noncontrastive acoustic dimensions helps listeners to develop robust phonetic categories.

Training paradigms with highly variable acoustic stimuli also induces learning in deaf people who use CIs (Miller et al, 2016; Zhang et al, 2021). In these studies, the training set included syllables spoken by multiple talkers. Adult CI listeners' identification and discrimination of perceptual categories showed improvements post-training compared to pre-training. These studies underscore the idea that high variability training can induce learning. However, because the goals of these studies were to assess whether learning is possible with the training paradigm, and not the variability per se, it is unclear whether listeners would have performed similarly if exposed to variable input spoken by a single talker.

Notably, Hartman, Saffran, and Litovsky (2022, in prep) examined whether talker variability could enhance word learning in adult CI listeners. In this study, listeners were taught novel word-object pairings spoken by a single talker or by 6 different talkers (multiple talker condition). The authors found that talker variability did not enhance learning for adults with normal-hearing (NH) or CIs; for each group, performance was similar between the two talker conditions. There are two possible reasons for this null result. First, the authors considered ceiling effects as a possible factor; for CI listeners, performance reached 89%, regardless of the learning condition, suggesting that the task was easy. More so, this ceiling effect may be due to the fact that learning occurred in quiet listening conditions. Thus, under an ideal listening situation, talker variability might not have been required to accurately perceive the speech

sounds in the word form. Alternatively, this null effect could be due to the presence of a video of the talker during the learning phase. In fact, the authors observed that CI listeners attended more to the talker's mouth than NH listeners. It is possible that, for CI listeners, visual speech cues might have been a more important cue than talker variability or that talker variability may not have been helpful beyond seeing the talker speak.

Visual cues support perception by providing information about the timing and content of the auditory signal, such as the onset and temporal amplitude envelope of speech, as well as place of articulation (see Peelle & Sommers, 2015 for review). Moreover, eye-tracking studies have shown that adults listeners will direct their gaze to the mouth before and during a speech event, as a means to capture relevant linguistic cues available at the mouth (Lansing & McConkie, 2003; Hisanga et al., 2016). Additionally, in regards to learning, studies have found a strong correlation between attention to the mouth and language outcomes (Tenenbaum et al., 2015; Tsang, Atagi, & Johnson, 2018). For example, infants who attend more to the mouth have better receptive and expressive language scores than those who attend less often. These studies suggest that infants who direct their gaze to the mouth are better able to capture phonetic information available at the mouth. Interestingly, no study has directly tested whether learning is enhanced with audiovisual compared to auditory only input, particularly under degraded listening conditions.

Most studies that have demonstrated the benefit of acoustic variability and audiovisual speech on learning have investigated these information sources in isolation. However, in typical learning environments, both talker variability and audiovisual speech occur simultaneously. Not only does communication typically occur face-to-face, it also occurs with different people. Moreover, both of these inputs have been shown to disambiguate the acoustic signal, albeit

through different mechanisms. Whereas audiovisual cues provide redundant speech cues, talker variability allows listeners to infer the structure of the phonetic categories (Rost & McMurray, 2009, 2010). Thus, to fully capture how individuals learn words, it is important to assess how listeners utilize multiple sources of information within their environment.

Noise presents an interesting arena to explore these ideas because it creates a challenging listening environment by reducing the reliability of the acoustic signal. For example, Bidelman, Sigley, & Lewis (2019) found that noise weakens phonetic categorization, possibly by interrupting the mapping of acoustic to internal phonetic representations. Furthermore, several factors, such as the masker type (Hawley et al., 2004; Tun et al., 1991), spatial location of the interferers relative to the target (Jones & Litovsky, 2011; Misurelli & Litovsky, 2015), and the perceptual similarity between the talker and masker (Brungart et al., 2000; Durlach et al., 2003; Kidd et al., 2016; Misurelli & Litovsky, 2015), can exacerbate the effect of noise on perception. Despite the deleterious effects of noise, in some situations, listeners are able to accurately perceive speech in noise and learn spoken words (Blaiser et al., 2014; McMillan & Saffran, 2016; Riley & McGregor, 2012).

One factor that has been shown to ameliorate the effects of noise is audiovisual cues. For example, in unfavorable signal-to-noise ratios (SNR), speech intelligibility is higher with audiovisual input compared to audio only input (Grant et al., 1998; Sumby & Pollack, 1954; Tye-Murray et al., 2007). Furthermore, on a categorical perception task, the presence of visual speech counteracts the deleterious effects of noise by sharpening categorical perception (Bidelman, Sigley, & Lewis, 2019). Thus, audiovisual speech cues reduce the ambiguity of the auditory signal, especially when the signal is impoverished.

It is also unclear whether talker variability would be beneficial for word learning in noise, and how variation would interact with audiovisual speech when both sources are present simultaneously in the environment. Because noise distorts the acoustic signal, listeners might not be able to tune into the relevant acoustic dimensions of the words to be learned. Alternatively, the presence of noise might increase the difficulty of the task, propelling listeners to utilize variation to enhance their learning. Indeed, it seems when the task is difficult (e.g., acquiring a native or foreign language), talker variability is able to bolster learning. Moreover, with regards to the interaction between audiovisual speech and variability, it is possible that these two inputs might work antagonistically. For example, on a speeded task, talker variability slows word recognition, even when audiovisual speech is present (Heald & Nusbaum, 2014). Thus, extra talker information in the visual display can increase processing load. Alternatively, these two sources of information could work synergistically to improve learning in noise. Whereas audiovisual speech cues provide redundant information about the acoustic signal, talker variability would allow listeners to infer the statistical structure of the lexical contrasts. Thus, listeners could utilize these inputs to support their learning.

The present study aims to address two questions 1) does talker variability and/or audiovisual speech enhance word learning in noise and 2) do listeners learn word better when both audiovisual speech and talker variability are presented concurrently rather than in isolation? To address these questions, we exposed NH adult listeners to novel word-object pairings. Half the listeners were taught by a single talker whereas the other half were taught by 6 different talkers (between-subjects). During training, listeners either saw a still image (audio only) or video of a talker speaking (within-subjects). Word learning was then tested by displaying pairs of objects from the training phase and tracking participants' eye-gaze while the target object was

labeled. For the test trials, we also manipulated the degree of similarity between the distractor and target labels.

For talker variability, we hypothesize that the presence of noise may increase the task difficulty for NH listeners, propelling listeners to rely on talker variability to tune into the relevant acoustic dimension. Furthermore, we predicted that talker variability would be more beneficial for learning similar sounding words, consistent with findings reported in Rost & McMurray (2009). Alternatively, the presence of noise may hamper listeners' ability to utilize talker variability by masking the acoustic signal. For audiovisual speech, we hypothesized that audiovisual speech cue will facilitate word learning in noise by providing complementary information to the auditory signal. Finally, we hypothesize that word learning would be enhanced when both inputs occur in combination, given that audiovisual speech provides redundant speech cues and talker variability highlights the relevant acoustic contrast of the words to be learning. Findings from this study will provide basic understanding on how adults make use cues in environment to support learning and are of clinical importance to individuals with CIs.

METHODS

Participants

54 participants between the ages of 18-26 years old were recruited and randomly assigned to either the Single Talker or Multiple Talker condition. Participants were native English speakers, had normal or corrected to normal vision, and normal hearing (audiometric thresholds of 20dB for octaves between 0.25 and 8 kHz). Participants were recruited through academic departments (course credit) or campus job postings (paid). Five participants were excluded due to failing the hearing screening. An additional ten were excluded due to contributing less than half of the test trials for each modality condition, resulting in a final

sample size of 39 (N = 21 for multiple talker, N = 18 for single talker) This study was approved by the campus Institutional Review Board.

Visual 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.

Auditory stimuli

Speech stimuli consisted of 8 novel words: /dita/, /gita/, /foma/, /voma/, /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 8 native English speakers (4 males, 4 females) raised in the Midwest. Multiple tokens of each word were recorded and a single token of each word was selected from each speaker. One speaker was assigned to the Single Talker condition, six speakers were assigned to the Multiple Talker condition, and one speaker was assigned to the test trials. The talker assigned to the Single Talker condition was randomized across participants; however, the speaker for the test trials always remained the same. The Single Talker assignment was randomized to ensure that any differences we found were not due to a particular speaker.

Audio/visual speech stimuli were videorecorded with an iPad Air Pro (30 frames/sec, resolution of 1920 x 1080). Each talker was filmed against a solid background. A microphone was placed 8 inches away from the talker to record the audio (44.1kHz sampling rate). Audio recorded from the microphone was processed using Adobe Audition and replaced the original audio recorded from the iPad Air Pro. 400ms of silence was added before the onset of the word and 300ms 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. Audio was synchronized to videos using Adobe Premiere. Mean length of video was 1015ms (range:1000ms-1027ms). Still images of each talker were generated by extracting the first frame of each video. Noise stimuli consisted of IEEE sentences spoken by a male and female speaker. Each speaker spoke a different IEEE sentence. The talkers were combined to produce a two-talker masker. To ensure that the listening condition for the Multiple Talker manipulation would be consistently difficult across trials, we chose the maskers to consist of a male and female talker. The masker was set at 65dB and presented simultaneously with the target stimuli to yield an SNR of -10dB. We selected this SNR level based on piloting; it was chosen to be challenging without being impossible for listeners.

Word Object Pairs

Each novel word was paired with a novel object (8 word-object pairings; see Figure 1). There were 2 sets of 4 novel-word object pairings. Set 1 consisted of the items /dita/, /gita/, /foma/, and /voma/. Set 2 consisted of the items /nodi/, /lodi/, /pibu/, and /tibu/. Each set was assigned to a learning modality (audio only vs. audiovisual), 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).

Apparatus

Participants were tested in a double-walled sound booth (Acoustic System, Tx, USA). Participants sat at a table with a 19-inch LCD monitor (1,280 by 1,240 pixels). Eye gaze was tracked with the EyeLink SR 1000 eye-tracker (SR Research, Kanata, ON, Canada) at a sampling rate of 1,000 Hz. A chin rest was used to maintain the distance of the head to the monitor and to restrict head movement. A sound card (Babyface) delivered the audio signal to a

loudspeaker positioned at the front of the room. Audiovisual stimuli were presented using custom software written in MATLAB (Mathworks, Natick, MA, USA). The Psychophysics Toolbox (v3.0.14) was used to maintain the synchronization of audiovisual stimulus presentation with eye-tracking camera.

Procedure

Participants were seated 1 m from the computer screen. The procedure began with a 9-point calibration: Participants were asked to look at 9 different locations on the screen. Prior to the start of the experiment, participants were randomly assigned to the single talker or multiple talker condition. Participants assigned to the single talker condition heard all novel word-object pairings spoken by the same talker, whereas participants assigned to the multiple talker condition heard all pairings spoken by six different talkers. The experiment began with the learning phase which consisted of two within-subject conditions, an audio only and an audiovisual condition. Participants completed a learning phase followed by the test phase for one presentation modality, and then the learning phase followed by the test phase for the presentation modality (order counterbalanced across participants). In the audio only condition, participants saw a still image of the target talker and heard the talker label the word. In the audiovisual condition, they heard and saw the talker label the word. Participants were instructed to attempt learning the names of the novel object. The learning phase began with the novel object appearing at the bottom of the screen. After 2000ms, the two-talker babble was presented (-10 dB SNR). Simultaneously, the participant saw an image or video of the talker appear in the center of the screen and heard the talker label the novel object. The video and image remained on the screen for 1000ms before disappearing. In the single-talker condition, each of the 4 objects was labelled 6 times by a single speaker, for a total of 24 trials. For the multiple-talker condition, each of the 6

talkers labelled each of the 4 objects once, for a total of 24 trials. The labelling 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 following each learning phase (see Figure 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., /dita/ vs. /voma/). Hard trials were defined as target and distractor labels that served as minimal pairs (e.g., /dita/ vs /gita/). Minimal pairs always differed in the onset consonant. Note that all test trials occurred in quiet.

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 citation form. They 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 4 times, for a total of 24 trials.

Eye-gaze coding

Eye gaze data were analyzed for test trials with respects to two areas of interests (AOI): target object and distractor object. If the gaze fell outside of these AOIs, or if tracking eye movement was unsuccessful, then eye-gaze for that time point was considered as “away.” Mean accuracy was calculated as proportion of time spent looking at the target object out of the total time spent looking at either of the two objects during a critical window of 300 to 1800 ms following onset of the target word (Fernald et al., 2008). For this measurement, trials were excluded if the participant was not fixated to either object for more than 50% of the critical window.

RESULTS

In this study, we sought to examine the individual and synergistic contributions of talker variability and audiovisual speech on word learning in noise in NH adults. We analyzed participants' mean accuracy (proportion of looks to target relative to total looks to target and distractor) using linear mixed effects model and approximated Kenwood-Roger's degrees of freedom. The model included the fixed effects of talker condition, presentation mode, test difficulty, and the interaction effects of mode and test difficulty, talker condition and test difficulty, and mode and talker condition. For each independent variable, levels (e.g., audio only compared to audiovisual) were contrast coded as -0.5 or 0.5. For the random effects structure, we included a by-subject random intercept and by-subject random slope for presentation mode. All analyses were conducted in R using the package lme4, following the pre-registered report (<https://osf.io/ry34a>). In the following sections, we will examine the effects of talker variability and/or audiovisual speech on word learning in noise.

Talker variability on word learning in noise

First, we examined whether talker variability would improve word learning in noise for NH listeners. We hypothesized that the presence of noise may increase the task difficulty for NH listeners, propelling listeners to rely on talker variability to tune into relevant acoustic dimension. Alternatively, the presence of noise may hamper listeners' ability to utilize talker variability by masking the acoustic signal. The time course of language processing, as illustrated by frame-by-frame data, provides a fine-grained measure of language processing. As seen in Figure 1, listeners in the single talker group showed greater proportion of looks to the target than those in the multiple talker group. To analyze the data, we averaged the proportion of looks to the target (mean accuracy) within the critical time window of 200 to 1800ms after word onset. This

analysis is consistent with the standard eye gaze-based measurements of word learning (Fernald et al., 2008).

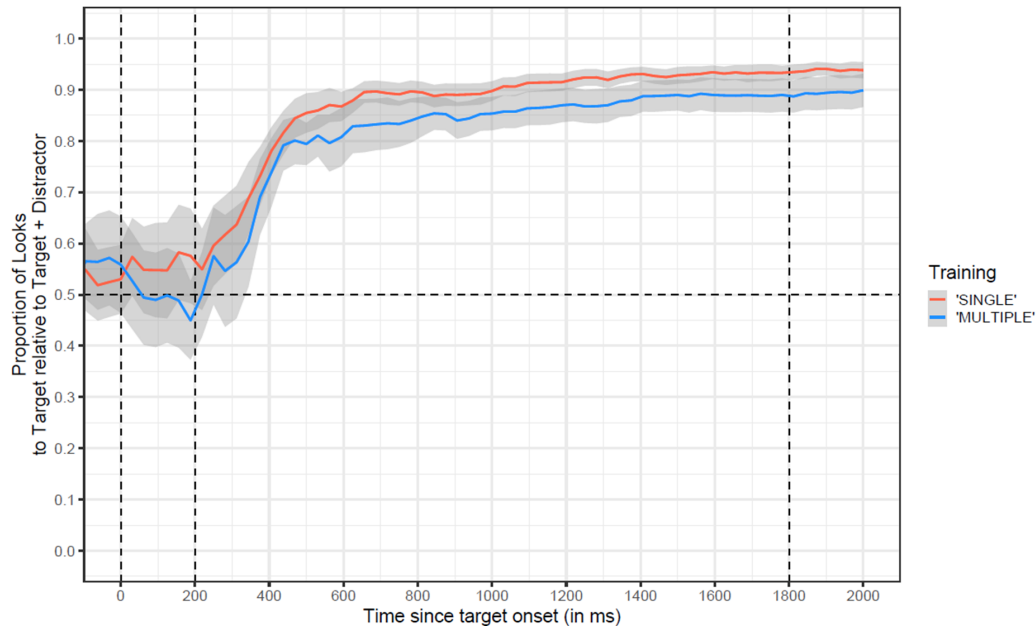


Figure 1. Time Course of Fixations by Talker Condition. Time course of fixations to target object following the single talker (red) and multiple talker (blue) learning conditions. Data is averaged across trials. Dashed horizontal lines represent the onset of the target word (0 ms) and the analysis time window (200 ms and 1800ms). Ribbons around lines indicate ± 1 SE.

There was no main effect of talker condition ($b = .022$, $F(1,49.42) = .165$, $p = .68$) or test difficulty ($b = .013$, $F(1,74) = .231$, $p = .63$) nor an interaction between talker condition and test difficulty ($b = .059$, $F(1,74) = 2.20$, $p = .14$). Performance was similar between the multiple talker ($M = .85$, $SD = .032$) and single talker learning groups ($M = 0.88$, $SD = .018$, Fig. 2). Moreover, when test trials were partitioned by difficulty level, performance was still similar between the two groups (Easy: Single: $M = .91$, $SD = .07$; Multiple: $M = .87$, $SD = .15$; Hard: Single: $M = .86$, $SD = .10$; Multiple: $M = .85$, $SD = .16$). Thus, even under adverse listening

situations, talker variability neither facilitated nor hampered robust representation of lexical categories.

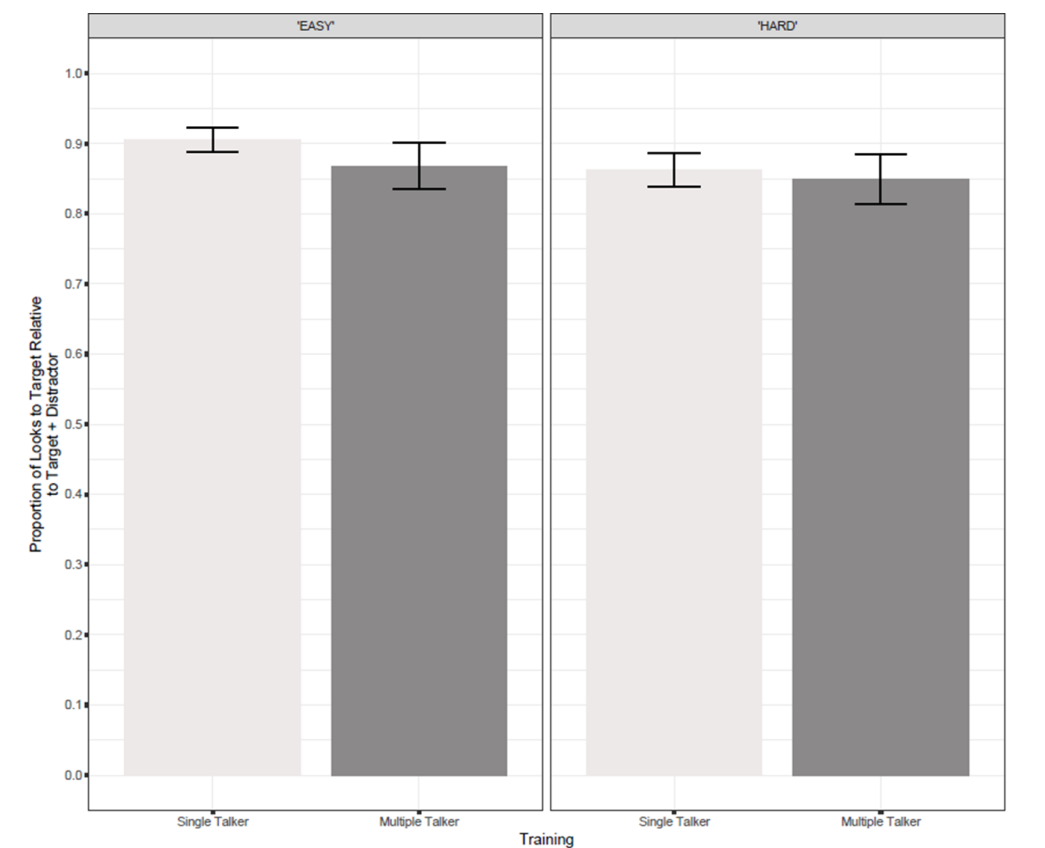


Figure 2. **Mean Accuracy by Talker Condition.** Proportion of looks to target relative to target and distractor during the critical time window for the single talker (light gray bar) and multiple talker (dark gray bar) group. Data represents mean accuracy on easy (left panel) and hard (right panel) test trials following each talker condition. Error bars represent ± 1 SE.

Audiovisual speech on word learning in noise

Next, we examined whether audiovisual speech would improve word learning in noise. We hypothesized that audiovisual speech cues would enhance word learning in noise by providing complementary information to the auditory signal. As seen in Figure 3, viewing a talking face rendered quicker and increased proportion of looks to the target object than viewing

a static image of the talker. To analyze the data, we averaged the proportion of looks to the target object (mean accuracy) within the same critical window described in the previous section.

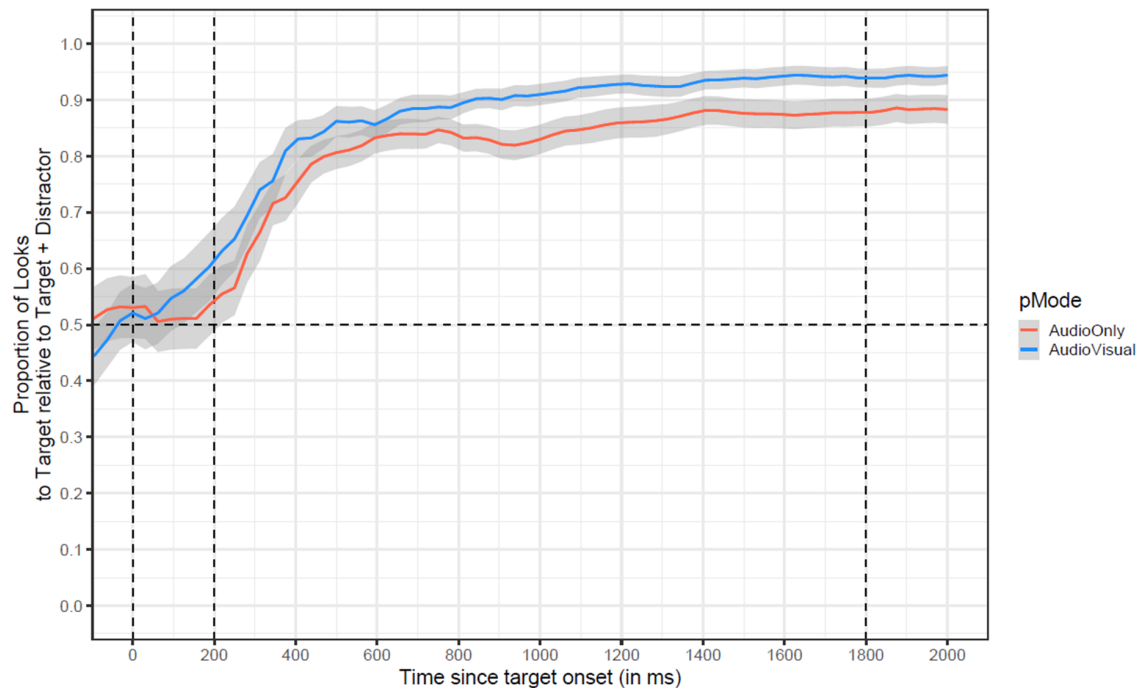


Figure 3. Time course of Fixations by Presentation mode. Time course of fixations to target object following the audio only (red) and audiovisual (blue) learning phases. Data is averaged across trials. Dashed horizontal lines represent the onset of the target word (0 ms) and the analysis time window (200 ms and 1800ms). Ribbons around lines indicate ± 1 SE.

There was a significant effect of presentation mode ($b = .069$, $F(1,71) = 4.13$, $p < .05$) but no significant interaction between presentation mode and test difficulty ($b = .00207$, $F(1,75) = .0054$, $p = .94$). Overall, listeners learned words significantly better when also seeing rather than solely hearing the talker speak during the learning phase (Audio: $M = .84$, $SD = .15$; AudioVisual: $M = .90$; $SD = .11$, Fig. 4). Note that while listeners were exposed to the words with audio only or audiovisual input, they were tested with solely auditory input. Thus, this result suggests that audiovisual input allows listeners to form robust representations of lexical items that they can then generalize to a novel speaker.

Although there was not a significant interaction of presentation mode and test difficulty, mean accuracy was highest for both test trial types following an audiovisual learning phase (Easy: Audio: $M = .85$, $SD = .17$; Audiovisual: $M = .92$, $SD = .092$; Hard: Audio: $M = .82$, $SD = .18$; Audiovisual: $M = .88$, $SD = .14$, Fig. 4). To examine the effects of presentation mode on test difficulty, we ran separate analyses on easy and hard trials. For each trial type, we implemented a model that included presentation mode as a fixed effect and by-subject intercept as a random effect. For both easy and hard test trials, performance significantly improved with audiovisual compared to audio only input (Easy: $b = .061$, $F(1,38) = 7.10$, $p < .05$; Hard: $b = .059$, $F(1,38) = 4.66$, $p < .05$). Thus, audiovisual input allowed listeners to distinguish words, regardless of the degree of perceptual similarity.

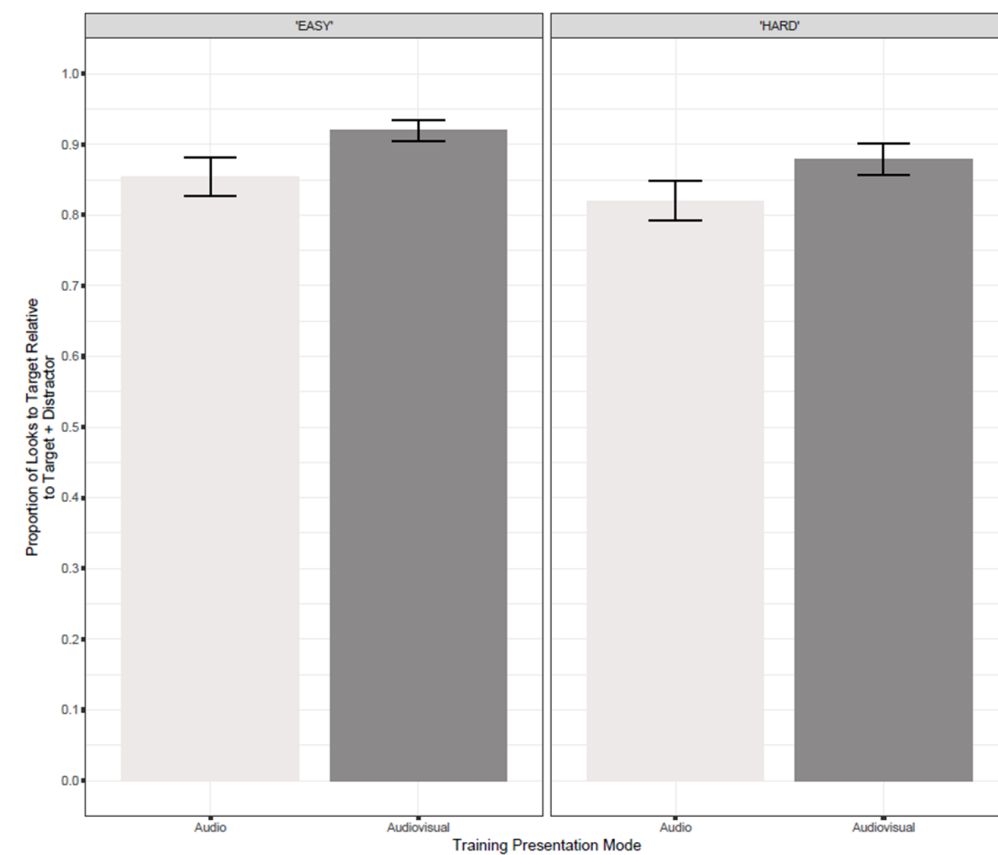


Figure 4. **Mean accuracy by learning presentation mode and test difficulty.** Proportion of looks to target relative to total looks to the target and distractor objects during the critical time window for the

audio only (light gray bar) and audiovisual (dark gray bar) group. Data represents mean accuracy on easy (left panel) and hard (right panel) test trials following each talker condition. Error bars represent ± 1 SE.

Combination of talker variability and audiovisual speech on word learning in noise

Finally, we examined whether word learning in noise would be enhanced when both talker variability and audiovisual speech are presented concurrently rather than independently. As seen in Figure 5, participants showed greater proportion of looks to the target after learning with audiovisual than auditory input, regardless of talker condition. To analyze the data, we averaged the proportion of looks to the target object within the critical window.

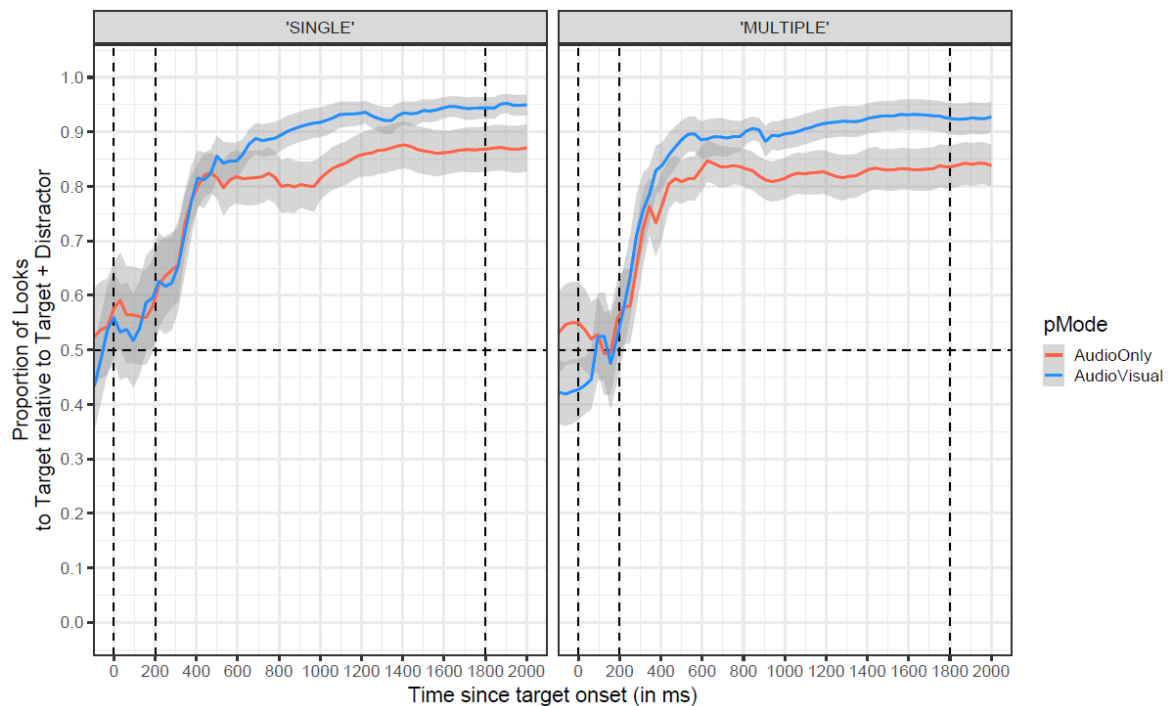


Figure 5. Time course of Fixations by Talker Condition and Presentation Mode. Proportion of looks to the target relative to total looks to target and distractor for the single talker (left panel) and multiple talker group (right panel) as a function of presentation mode (audio only: red lines audiovisual: blue lines). Dashed horizontal lines represent the onset of the target word (0 ms) and the analysis time window (200 ms and 1800ms). Ribbons around lines indicate ± 1 SE.

There was no significant interaction between talker variability and audiovisual speech, suggesting that learning is not enhanced when both cues are presented simultaneously (Single: Audio: $M = .87$, $SD = .08$; Audiovisual: $M = .90$, $SD = .09$; Multiple: Audio: $M = .81$, $SD = .2$; Audiovisual: $M = .90$, $SD = .12$). Although the interaction was not significant, mean accuracy differed between the presentation modes for the multiple talker group, but not the single talker group (Fig. 6). To investigate this finding further, we ran separate analyses for each training condition using linear mixed model, approximated with a Kenward-Roger degrees of freedom. Mean accuracy was regressed on presentation mode, test difficulty, and an interaction between presentation mode and test difficulty. The model also included a by-subject random intercept and by-subject random slope for presentation mode. For the single talker group, there was a significant effect of test difficulty ($b = .072$, $F(1,34) = 8.44$, $p < .01$). Overall, individuals in the single talker group performed higher on easy test trials than hard test trials. We did not find an effect of presentation mode ($b = .047$, $F(1, 38) = 2.82$, $p = 0.10$) nor an interaction effect of presentation mode and test difficulty ($b = -.028$, $F(1,34) = .65$, $p = .42$). For the multiple talker group, we found a trend towards significance for presentation mode ($b = .06$, $F(1,36) = 3.03$, $p = .09$), but no main effect of test difficulty ($b = .013$, $F(1,40) = .187$, $p = .67$) nor an interaction effect of presentation mode and test difficulty ($b = .028$, $F(1,40) = .437$, $p = .51$). These results suggest the presence of visual speech cues enhances learning when the talker varies across trial, but not when the talker remains the same.

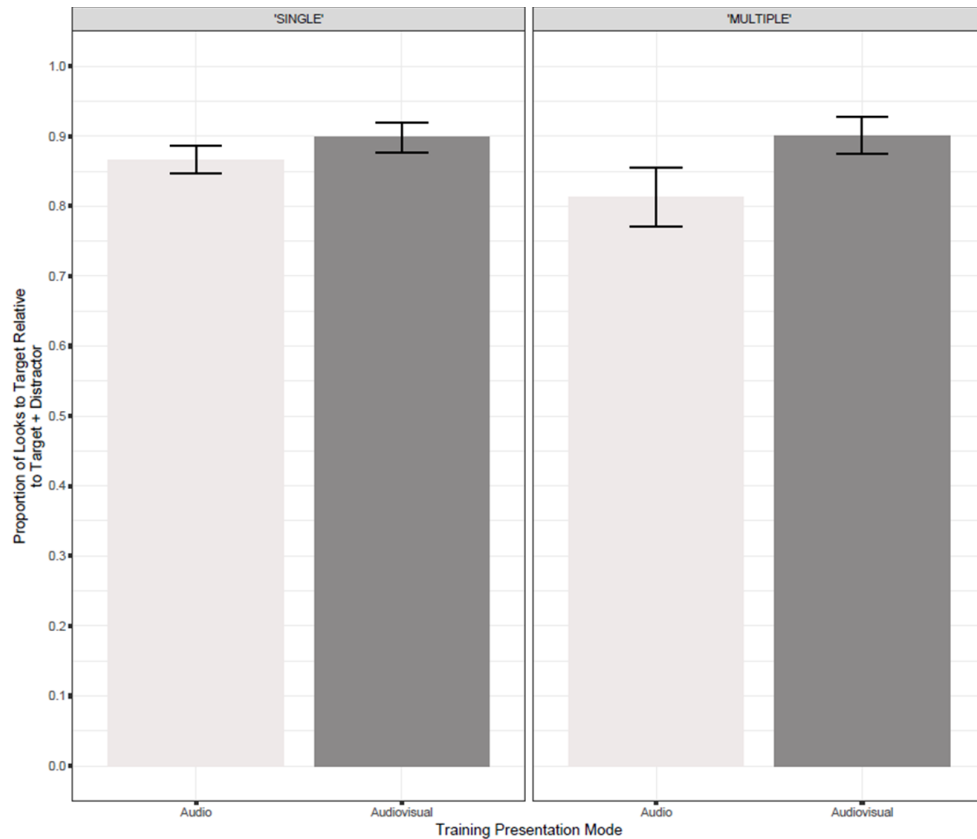


Figure 6. **Mean Accuracy by Talker Condition and Presentation Mode.** Proportion of looks to the target relative to total looks to target and distractor during the critical time period for the single talker (left panel) and multiple talker group (right panel) as a function of presentation mode (audio only: light gray bars; audiovisual: dark gray bars). Error bars represent ± 1 SE.

Exploratory analyses of individual differences in audiovisual benefit

Several studies have found that the benefit of audiovisual speech input varies between participants (Alsus et al., 2016; Lorin, Pisoni, & Kirk, 2001; Stevenson, Zemstov & Wallace, 2012). Deviating from the pre-registered report, we examined the individual differences in audiovisual benefit on word learning in noise. For each participant, we calculated the relative benefit in accuracy due to the presence of the audiovisual signal (i.e., visual gain) using the following formula: $[(\text{audiovisual score} - \text{auditory-only score}) / (1 - \text{auditory-only score})]$. This calculation is consistent with the standard measurement of visual gain (Alsus et al., 2016; Grant,

2002; Lorin, Pisoni, & Kirk, 2001; Sumby & Pollack, 1954). A positive score reflects an enhancement in learning due to the presence of visual input, whereas a negative score reflects a reduction in learning with audiovisual input. For nine participants, we were unable to calculate their visual gain because they performed at ceiling on both the audio only and audiovisual conditions. As seen in Fig. 7, the majority of participants (~78%) showed a visual gain. In contrast, seven participants did not show a visual score, as reflected by a negative score. Altogether, these results suggests that for most participants, audiovisual input enhances word learning in noise.

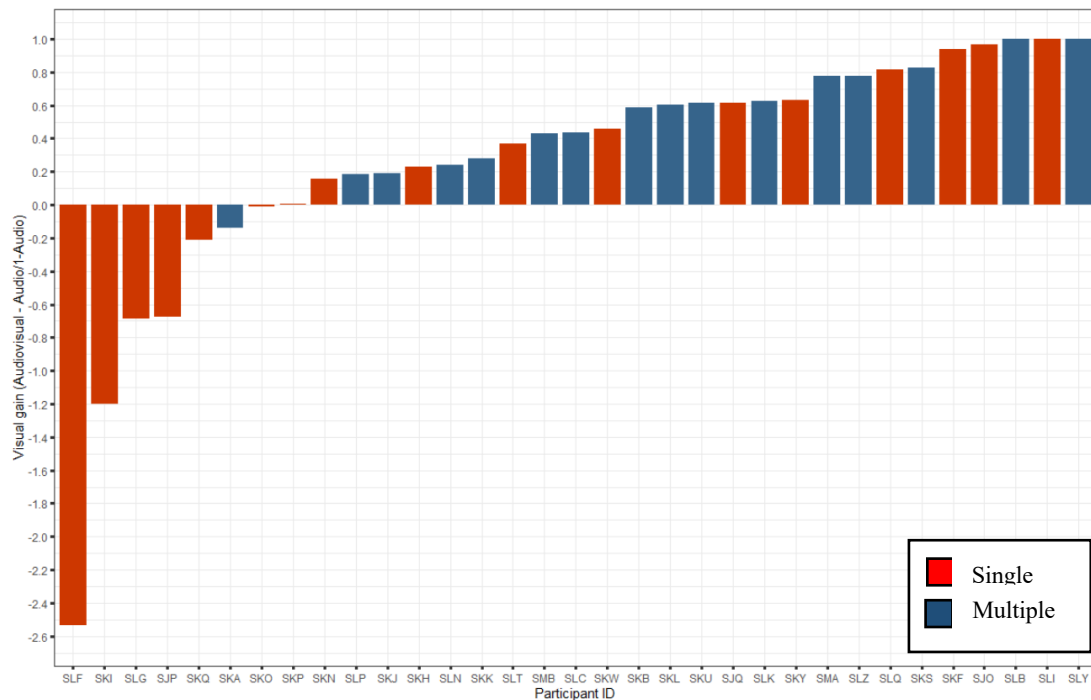


Fig 7. Individual Differences in Visual Gain. Relative benefit in accuracy due to the presence of visual cues for each participant. Blue bars represent participants assigned to the multiple talker group while red bars represent those assigned to the single talker group.

Interestingly, most of the participants who showed a visual gain were assigned to the multiple talker condition. To examine whether talker condition influenced the relative audiovisual benefit, we regressed visual gain scores on training condition (model: visual gain ~

Training + (1|Subject)). There was a trend towards significance due to training. On average, visual gain was higher for individuals in the multiple talker group ($M = .53$) compared to those in the single talker group ($M = .05$). However, the sample size was too small to draw any strong conclusions from these analyses.

CONCLUSION

The primary goal of this study was to examine the independent and combined influence of talker variability and visual speech on word learning in noise. Our results revealed that audiovisual speech cues, but not talker variability, enhanced word learning in noise for NH listeners. Participants looked more and quicker to the target object following an audiovisual learning phase than an auditory only learning phase. Moreover, seeing the talker speak during learning allowed listeners to distinguish perceptually similar and distinct words during test.

Previous studies have found that talker variability supports learning by highlighting the contrastive acoustic dimension of the words to be learned (e.g., Gomez, 20002; Lively, Logan, & Pisoni, 1992; Perry et al., 2010; Quam et al., 2017; Rost & McMurray, 2009; 2010). However, in our study, we found that, overall, talker variability neither helped nor harmed word learning in noise. Listeners were able to develop robust representations of the words, regardless of the talker condition they were assigned to. Consistent with previous studies (Davis et al., 2015; Hartman, Saffran, & Litovsky, 2022, under review), our finding suggests that for adults, talker variability is not required for developing robust representations, even if the acoustic input is impoverished.

There are at least two possible reasons for this null effect. One reason may be the fact that varying the talker across trials makes it difficult for listeners to keep track of the target speaker. Indeed, studies on auditory scene analysis have demonstrated that recurring structures aid in sound source segregation ((Bregman, 1990; McDermott et al., 2011). Thus, in our study, the

single talker condition might have supported learning by allowing listeners to parse out the target speaker from the background noise. Additionally, it is possible that the presence of noise might have prevented listeners from utilizing acoustic variability to develop robust lexical categories. For example, noise distorts the acoustic signal by reducing the audibility of acoustic cues (Bidelman et al., 2019; Parikh & Loizou, 2005). This distortion can be extremely detrimental if listeners can no longer perceive the relevant acoustic dimension that contrasts the words to be learned or if the acoustic signal becomes less variable across talkers. The current data does not provide insight into the acoustic cues that were reliably transmitted through the signal, despite the presence of noise. Such an analysis could provide insight as to why talker variability does not enhance word learning in noise. Nonetheless, these findings reveal the extent to which acoustic variability supports word learning.

Turning to the audiovisual results, we found performance significantly improved following the audiovisual learning phase compared to the audio only condition. This finding is consistent with previous studies that have demonstrated the benefit of audiovisual input on speech perception (Bidelman, Sigley, & Lewis, 2019; Grant & Seitz, 2000; Sumby & Pollack, 1954; Tye-Murray, Sommers, & Speher, 2007). In our study, the presence of visual cues allowed listeners to develop robust lexical categories. This result suggest that audiovisual cues sharpen phonetic categories, particularly under adverse listening conditions.

It is worth noting that both the audio and audiovisual learning conditions contained visual information about the talker. However, while the audio only condition presented static visual cues (e.g., a still image of the talker), the audiovisual learning condition presented dynamic cues. Our results indicate that merely seeing a static face does not help listeners to perceive words better. Rather, seeing articulatory movements helps listeners to acquire spoken words. These

dynamic cues not only provide the talker's identity but also signal the onset of the acoustic signal (Chandrasekaran et al., 2009). It is possible that the presence of a two-talker, mixed gender babble made it difficult for listeners to use static cues. In this situation, listeners could not have used a visual image of the talker to predict who to listen to. Future studies should examine whether static visual cue augment learning in noise when target and masker differ by gender.

We also found that dynamic visual cues boosted performance for the multiple talker group, but not the single talker group. In fact, the presence of visual cues did not afford much benefit to the single talker, as evidenced by the similarity in accuracy scores between the two presentation modes. Moreover, our individual differences analysis revealed that more individuals in the multiple talker group showed a visual gain than those in the single talker group. Due to our small final sample size, we did not observe an interaction between talker variability and presentation mode. However, it is possible that talker variability hampers learning in noise when only acoustic input is available. Such a finding would be consistent with the idea of talker normalization (Magnuson & Nusbaum, 2007; Martin et al., 1989; Mullennix et al., 1998; Nygaard & Pisoni, 1998; Wong et al., 2004); that is, talker variability incurs a perceptual cost. It might also explain why audiovisual visual speech cues boost performance more for the multiple talker group, but not the single talker group. By providing redundant information, audiovisual speech cues might reduce the perceptual cost of talker variability.

Finally, there was no additive effect of talker variability and audiovisual speech when both cues co-occurred, indicating that the presence of both cues does not enhance learning in noise. One possible explanation for this null result may be due to the fact that talker variability and audiovisual speech cues may serve as opposing forces in noisy learning situations. Whereas talker variability may impede learning in noise, audiovisual speech supports it. This

interpretation would suggest that talker variability is not helpful beyond audiovisual cues. It may also explain why talker variability did not enhance word learning for CI listeners, a population that is constantly exposed to an impoverished acoustic signal (Hartman, Saffran, & Litovsky, 2022, under review). In that study, learners were presented with acoustic variability and visual speech cues. Thus, listeners may rely on audiovisual cues to develop richly defined lexical categories when both cues are available.

In terms of applicability of this work to realistic everyday situation, the study was designed to address a commonly found scenario, which is that typical learning environments are noisy and contain both acoustic variability and visual speech cues. Our findings suggest, that in a noisy environment, listeners may not be able to capitalize on talker variability to infer the structure of phonetic categories. However, they may be able to use audiovisual speech cues to acquire newly spoken words.

Chapter 5. General Discussion

The goal of this dissertation was to examine the impact of talker variability and visual speech on word learning in adverse listening conditions. In this dissertation, I focused on two sources of acoustic degradation: listening with a cochlear implant (CI; Chapter 3) and listening with background noise (Chapter 4). In Chapter 3, I sought to address the following questions: 1) does talker variability improve word learning for CI listeners and 2) how do CI listeners direct their gaze while viewing a talker speak? In Chapter 4, I sought to examine the impact of talker variability and visual speech on learning when each cue is presented independently or in combination. These questions were addressed using a novel word learning task. Adults with TH and with CIs were taught novel word object pairings spoken by a single talker or multiple talkers. In Chapter 3, these pairings were taught in quiet to older adults with TH and with CIs. In Chapter 4, the same word object pairs were taught to young TH adults in the presence of noise. Immediately following learning, adults were tested on the word-object pairings with a novel talker.

The results showed that talker variability had no effect on learning in quiet for CI or MH listeners. That is, listeners were able to successfully acquire new words, regardless of the talker condition. Moreover, while learning words, CI listeners focus more on the talker's mouth than NH listeners. The same results were found in Chapter 4 when noise was added to the background. However, performance did improve when young NH adults were presented with audiovisual speech cues during learning than with audio only cues. Furthermore, there was no additive effect of talker variability and audiovisual cues on word learning in noise. The next sections will discuss the implications of each finding.

The Role of Talker Variability

Previous research has demonstrated that talker variability is beneficial for learning. Both children and second-language learning adults are able to successfully acquire words spoken by multiple talkers than a single talker. Through variability, learners are able to tune into the relevant acoustic dimension that contrast the words to be learned, while ignoring the irrelevant ones. To date, no study has assessed whether the benefits associated with talker variability could extend to a listening condition in which the acoustic input is degraded. The studies described in this dissertation were designed to bridge this gap. Similar to the previous studies, the multiple talker condition consisted of natural speech tokens that varied by many acoustic properties simultaneously.

In the current studies, talker variability had no impact on learning while listening with degraded or non-degraded (for the NH listeners in Chapter 3) auditory input. Both CI and NH listeners were able to successfully acquire newly learned word, regardless of learning from the same talker or different talkers. These findings add to a set of studies that also found no benefit of talker variability on learning (Bulgarelli & Bergelson, 2022; Bulgarelli & Weiss, 2021, Davis et al., 2015). Specifically, for adults, talker variability does not enhance learning of an artificial grammar (Bulgarelli & Weiss, 2021) nor of novel words (Davis et al., 2015).

As mentioned in the previous chapters and in the last two studies cited above, one reason for this null effect may be due to the fact that talker variability is only helpful for establishing new phonetic categories. For example, the benefit of talker variability has been found for infants and second language learners who need to acquire the contrastive features of their native and second language, respectively. However, the adults in our studies may have already acquired these contrastive features of their native language, which may have facilitated learning and encoding of novel words.

This null effect may also be due to the design of the experiment. Rather than yoking a pair of words, every word within a set served as a distractor label for the remaining words. This design allowed for the construction of easy and hard test trials. However, it may have also allowed listeners to learn the words as the test progressed through process of elimination. Future studies should yoke the word pairs to prevent learning during test phase.

Despite the lack of benefit associated with talker variability, listeners in both experiments performed remarkably well on the test trials. Accuracy was significantly above chance, indicating that participants were able to establish novel word-object associations. For the CI group, performance reached approximately 90% and 80% for the easy and hard test trials, respectively. These findings deviate from those of prior studies observing poor performance in CI listeners on word learning tasks. For example, Havy et al (2013) found that CI children failed to distinguish between similar sounding words, despite receiving audiovisual input during learning. One reason for the discrepancy between the current CI study and prior studies may be due to the age group tested. Whereas prior studies tested children, the current study tested adults. It is possible that children failed to learn similar sounding words because they have less language experience. Indeed, TH children sometimes fail to distinguish similar-sounding words (Werker & Tess, 1984). By adulthood, CI listeners may have acquired sufficient linguistic experience to distinguish between words that differ by a single phonetic feature.

Additionally, in the current study, young TH adults learned words successfully, despite the presence of background noise during learning (Chapter 4). Compared to the findings in Morini and Newman (2021), my study found that performance reached above 80% following the audio only condition. There are several differences between these two studies. First, the stimuli in Morini and Newman (2021) consisted of trisyllabic nonwords, whereas the current study

consisted of bisyllabic nonwords. Encoding and storing trisyllabic words may be more challenging because of the finite amount of cognitive resources available. Second, Morini and Newman presented white noise in the background, whereas the noise stimuli in the current study consisted of two-talker babble. Unlike two-talker babble, white noise does not contain spectrotemporal dips in the signal, preventing listeners from glimpsing acoustic information in the target signal. Thus, factors, such as age, masker type, and word length can influence spoken language acquisition in adverse listening conditions.

The Role of Visual Speech on Learning

In regards to the role of visual speech on learning, this dissertation found two main findings: 1) while learning words, CI listeners focused more on the talker's mouth than NH listeners, and 2) word learning in noise significantly improved for NH adults in the presence of audiovisual compared to audio only cues. In the next sections, the implications of each finding will be further discussed.

Visual Attention to Talking Faces

A talker's face provides highly informative cues that support learning. For example, the mouth region is the primary source of phonetic information and moves in alignment with the acoustic signal. The eye region conveys information about the affective state and identity of the talker. Given that the face contains a rich source of information, listeners must utilize a strategy to efficiently gather information. Several studies have examined how NH adults direct their gaze while viewing a talking face. However, few studies have explored how CI listeners direct their visual attention to a talking face. The studies described in this dissertation were designed to bridge this gap.

Consistent with previous research, our finding that CI listeners direct their gaze to the talker's mouth reflect listeners' reliance on visual speech cues (Rouger et al., 2007, 2008; Tremblay et al., 2010). As mentioned in Chapter 3, there are two possible reasons that explain CI listeners' gaze behavior. One reason is due to the impoverished acoustic input that listeners receive. By attending to the mouth, listeners receive redundant audiovisual cues to disambiguate the signal. Additionally, CI listeners' attention to the mouth might be a remnant of their period of auditory deprivation. During this period, the mouth region might have served as the only source of linguistic input. Thus, our results suggest that CI listeners utilize a gaze strategy in which they can efficiently extract visual speech information.

In the current study, the majority of CI listeners fixated to the talker's mouth more than 90% of the time. Consistent with Rouger et al (2007), this result suggests that any duration of deafness propels listeners to rely on visual speech cues. However, our data does not provide any insight into the efficiency of audiovisual integration in CI listeners. Even though any length of auditory deprivation may propel listeners to rely on visual cues, the duration or onset of deafness may impact how well listeners can integrate audiovisual information.

One limitation of this study is that words occurred in isolation. In typical face-to-face communication, words occur in a sentential context, which can convey additional information, such as prosodic cues. As mentioned in Chapter 2, prosodic information is distributed across the face. Thus, future studies should examine how CI listeners scan a talking face when additional visual information is available at other facial regions.

The Role of Visual Speech on Learning in Noise

Prior studies have demonstrated that the presence of visual cues enhances speech perception. Seeing a talker's face allows listeners to capture redundant audiovisual speech cues available at the mouth. Audiovisual cues have been shown to disambiguate the acoustic signal, particularly in adverse listening conditions, such as listening in noise. The current dissertation sought to expand on this line of research by explore the role of visual speech on learning. In the Study 2 (Chapter 4), the presence of visual cues allowed listeners to develop robust lexical categories, suggesting that audiovisual cues sharpen phonetic categories. Specifically, dynamic visual cues helped listeners to acquire spoken words than static visual cues. Moreover, the benefit associated with visual cues was more prominent for participants in the multiple talker group than in the single talker group. Whereas visual cues boosted performance for the multiple talker group, the presence of these cues did not afford much benefit to the single talker group.

Finally, word learning did not improve when talker variability and audiovisual speech were presented simultaneously than when audiovisual speech was presented alone. These results indicate that the presence of both cues does not enhance learning in noise. As suggested in both studies, it is possible that listeners may rely on audiovisual cues to support learning. This interpretation would suggest that talker variability is not helpful beyond audiovisual cues.

Conclusion

This dissertation demonstrated that talker variability is not required for robust representation and generalization of newly acquired word in adults. The current studies also found that visual speech allows listeners to develop robust representations of newly acquired words. However, the nature of these categories still remain an open question. From this dissertation, it is clear that listeners were able to develop rich categories that generalized to surface-level (e.g. non-phonemic) changes. However, it is unclear whether these categories were

defined enough to reject mispronunciations. Anecdotally, one of the CI participants ask if the label for the “gita” object was pronounced “gita” or “kita” following the learning phase, suggesting a fragile lexical representation. Thus, future work is needed to explore the role of talker variability in building well-defined categories.

References

- Ayneto, A., & Sebastian-Galles, N. (2017). The influence of bilingualism on the preference for the mouth region of dynamic faces. *Developmental Science*, 20(1). <https://doi.org/10.1111/DESC.12446>
- Barenholtz, E., Mavica, L., & Lewkowicz, D. J. (2016). Language familiarity modulates relative attention to the eyes and mouth of a talker. *Cognition*, 147, 100–105. <https://doi.org/10.1016/J.COGNITION.2015.11.013>
- Basirat, A., Brunellière, A., & Hartsuiker, R. (2018). The Role of Audiovisual Speech in the Early Stages of Lexical Processing as Revealed by the ERP Word Repetition Effect. *Language Learning*, 68. <https://doi.org/10.1111/lang.12265>
- Bergeson-dana, T., Bergeson, T. R., Pisoni, D. B., & Davis, R. A. O. (2005). *Development of Audiovisual Comprehension Skills in Prelingually Deaf Children With Cochlear Implants*. May. <https://doi.org/10.1097/00003446-200504000-00004>
- Besle, J., Fischer, C., Bidet-Caulet, A., Lecaigard, F., Bertrand, O., & Giard, M. H. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: Intracranial recordings in humans. *Journal of Neuroscience*, 28(52), 14301–14310. <https://doi.org/10.1523/JNEUROSCI.2875-08.2008>
- Bidelman, G. M., Sigley, L., & Lewis, G. A. (2019). Acoustic noise and vision differentially warp the auditory categorization of speech. *The Journal of the Acoustical Society of America*, 146(1), 60. <https://doi.org/10.1121/1.5114822>
- Birulés, J., Bosch, L., Brieke, R., Pons, F., & Lewkowicz, D. J. (2019). Inside bilingualism: Language background modulates selective attention to a talker's mouth. *Developmental Science*, 22(3). <https://doi.org/10.1111/DESC.12755>
- Blais, C., Fiset, D., Roy, C., Régimbald, C. S., & Gosselin, F. (2017). Eye fixation patterns for categorizing static and dynamic facial expressions. *Emotion (Washington, D.C.)*, 17(7), 1107–1119. <https://doi.org/10.1037/EMO0000283>
- Blaiser, K. M., Nelson, P. B., & Kohnert, K. (2014). Effect of Repeated Exposures on Word Learning in Quiet and Noise: [Http://Dx.Doi.Org/10.1177/1525740114554483](http://Dx.Doi.Org/10.1177/1525740114554483), 37(1), 25–35. <https://doi.org/10.1177/1525740114554483>
- Bregman, A. S. (1990). *Auditory Scene Analysis*. <https://doi.org/10.7551/MITPRESS/1486.001.0001>
- Brooke, N. M., & Summerfield, Q. (1983). Analysis, synthesis, and perception of visible articulatory movements. *Journal of Phonetics*, 11(1), 63–76. [https://doi.org/10.1016/s0095-4470\(19\)30777-6](https://doi.org/10.1016/s0095-4470(19)30777-6)

- Brooks, R., & Meltzoff, A. N. (2005). The development of gaze following and its relation to language. In *Developmental Science* (Vol. 8, Issue 6).
- Brunellière, A., Sánchez-García, C., Ikumi, N., & Soto-Faraco, S. (2013). Visual information constrains early and late stages of spoken-word recognition in sentence context. *International Journal of Psychophysiology*, 89(1), 136–147.
<https://doi.org/10.1016/j.ijpsycho.2013.06.016>
- Brungart, D. S., Rabinowitz, W. M., & Durlach, N. I. (2000). Evaluation of response methods for the localization of nearby objects. *Perception & Psychophysics*, 62(1), 48–65.
<https://doi.org/10.3758/BF03212060>
- Buchan, J. N., Pare', M. P., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience*, 2(1), 1–13.
<https://doi.org/10.1080/17470910601043644>
- Bulgarelli, F., & Bergelson, E. (2022). Talker variability shapes early word representations in English-learning 8-month-olds. *Infancy*, 27(2), 341–368.
<https://doi.org/10.1111/inf.12452>
- Bulgarelli, F., & Weiss, D. J. (2021). Desirable Difficulties in Language Learning? How Talker Variability Impacts Artificial Grammar Learning. *Language Learning*, 71(4), 1085–1121. <https://doi.org/10.1111/LANG.12464>
- Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C. R., McGuire, P. K., Woodruff, P. W. R., Iversen, S. D., & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *Science*, 276(5312), 593–596.
<https://doi.org/10.1126/science.276.5312.593>
- Calvert, G. A., Campbell, R., & Brammer, M. J. (2000). *Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex.*
- Chandrasekaran, C., Trubanova, A., Stillitano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. *PLoS Computational Biology*, 5(7).
<https://doi.org/10.1371/JOURNAL.PCBI.1000436>
- Cvejic, E., Kim, J., & Davis, C. (2012). Recognizing prosody across modalities, face areas and speakers: Examining perceivers' sensitivity to variable realizations of visual prosody. *Cognition*, 122(3), 442–453. <https://doi.org/10.1016/J.COGNITION.2011.11.013>
- Davidson, L. S., Geers, A. E., & Nicholas, J. G. (2014a). The effects of audibility and novel word learning ability on vocabulary level in children with cochlear implants. *Cochlear Implants International*, 15(4), 211–221.
<https://doi.org/10.1179/1754762813Y.0000000051>

- Davidson, L. S., Geers, A. E., & Nicholas, J. G. (2014b). The effects of audibility and novel word learning ability on vocabulary level in children with cochlear implants. *Cochlear Implants International*, 15(4), 211–221.
<https://doi.org/10.1179/1754762813Y.0000000051>
- Dick, A. S., Solodkin, A., & Small, S. L. (2010). Neural development of networks for audiovisual speech comprehension. *Brain and Language*, 114(2), 101–114.
<https://doi.org/10.1016/j.bandl.2009.08.005>
- Durlach, N. I., Mason, C. R., Kidd, G., Arbogast, T. L., Colburn, H. S., & Shinn-Cunningham, B. G. (2003). Note on informational masking. *The Journal of the Acoustical Society of America*, 113(6), 2984. <https://doi.org/10.1121/1.1570435>
- Fort, M., Ayneto-Gimeno, A., Escrichs, A., & Sebastian-Galles, N. (2018). Impact of Bilingualism on Infants' Ability to Learn From Talking and Nontalking Faces. *Language Learning*, 68, 31–57. <https://doi.org/10.1111/LANG.12273>
- Gildersleeve-Neumann, C. E., & Wright, K. L. (2010). English Speech Acquisition in 3- to 5-Year-Old Children Learning Russian and English. *Language, Speech, and Hearing Services in Schools*, 41(4), 429–442. [https://doi.org/10.1044/0161-1461\(2009/09-0059\)](https://doi.org/10.1044/0161-1461(2009/09-0059))
- Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *J Acoust Soc Am*, 103(5 Pt 1), 2677–2690.
<https://doi.org/10.1121/1.422788>
- Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: effect of location and type of interferer. *The Journal of the Acoustical Society of America*, 115(2), 833–843. <https://doi.org/10.1121/1.1639908>
- Heald, S. L. M., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience*, 8(MAR), 35.
<https://doi.org/10.3389/FNSYS.2014.00035/BIBTEX>
- Hernández-Gutiérrez, D., Abdel Rahman, R., Martín-Loeches, M., Muñoz, F., Schacht, A., & Sommer, W. (2018). Does dynamic information about the speaker's face contribute to semantic speech processing? ERP evidence. *Cortex*, 104, 12–25.
<https://doi.org/10.1016/j.cortex.2018.03.031>
- Hillairet de Boisferon, A., Tift, A. H., Minar, N. J., & Lewkowicz, D. J. (2017). Selective attention to a talker's mouth in infancy: role of audiovisual temporal synchrony and

- linguistic experience. *Developmental Science*, 20(3).
<https://doi.org/10.1111/DESC.12381>
- Hillaiet 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. *Journal of Experimental Child Psychology*, 172, 189–200.
<https://doi.org/10.1016/j.jecp.2018.03.009>
- Hisanaga, S., Sekiyama, K., Igasaki, T., & Murayama, N. (2016). Language/Culture Modulates Brain and Gaze Processes in Audiovisual Speech Perception. *Nature Publishing Group*. <https://doi.org/10.1038/srep35265>
- Houston, D. M., Stewart, J., Moberly, A., Hollich, G., & Miyamoto, R. T. (2012a). Word learning in deaf children with cochlear implants: effects of early auditory experience. *Developmental Science*, 15(3), 448–461. <https://doi.org/10.1111/j.1467-7687.2012.01140.x>
- Houston, D. M., Stewart, J., Moberly, A., Hollich, G., & Miyamoto, R. T. (2012b). Word learning in deaf children with cochlear implants: effects of early auditory experience. *Developmental Science*, 15(3), 448–461. <https://doi.org/10.1111/j.1467-7687.2012.01140.x>
- IJsseldijk, F. J. (1992). Speechreading performance under different conditions of video image, repetition, and speech rate. *Journal of Speech and Hearing Research*, 35(2), 466–471. <https://doi.org/10.1044/jshr.3502.466>
- Jiang, J., Dai, B., Peng, D., Zhu, C., Liu, L., & Lu, C. (2012). Neural synchronization during face-to-face communication. *Journal of Neuroscience*, 32(45), 16064–16069.
<https://doi.org/10.1523/JNEUROSCI.2926-12.2012>
- Jones, G. L., & Litovsky, R. Y. (2011). A cocktail party model of spatial release from masking by both noise and speech interferers. *The Journal of the Acoustical Society of America*, 130(3), 1463–1474. <https://doi.org/10.1121/1.3613928>
- Jordan, T. R., & Thomas, S. M. (2011). When half a face is as good as a whole: Effects of simple substantial occlusion on visual and audiovisual speech perception. *Attention, Perception, and Psychophysics*, 73(7), 2270–2285. <https://doi.org/10.3758/s13414-011-0152-4>
- Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and Lexical Effects on Audiovisual Word Recognition by Adults With Cochlear Implants. *Journal of Speech Language and Hearing Research*, 46(2), 390. [https://doi.org/10.1044/1092-4388\(2003/032\)](https://doi.org/10.1044/1092-4388(2003/032))

- Kaplan, E., & Jesse, A. (2019). *Fixating the eyes of a speaker provides sufficient visual information to modulate early auditory processing*.
<https://doi.org/10.1016/j.biopsycho.2019.107724>
- Kidd, G., Mason, C. R., Swaminathan, J., Roverud, E., Clayton, K. K., & Best, V. (2016). Determining the energetic and informational components of speech-on-speech masking. *The Journal of the Acoustical Society of America*, 140(1), 132–144.
<https://doi.org/10.1121/1.4954748>
- Kim, J., Cvejic, E., & Davis, C. (2014). Tracking eyebrows and head gestures associated with spoken prosody. *Speech Communication*, 57, 317–330.
<https://doi.org/10.1016/J.SPECOM.2013.06.003>
- Kirk, K., & Pisoni, D. (2002). Audiovisual integration of speech by children and adults with cochlear implants. *Seventh International Conference On*, 1689–1692.
http://www.isca-speech.org/archive/icslp_2002/i02_1689.html
- Kubicek, C., Hillairet De Boisferon, A., Dupierri, E., Lè Ne Loevenbruck, H., Gervain, J., & Schwarzer, G. (2013). Face-scanning behavior to silently-talking faces in 12-month-old infants: The impact of pre-exposed auditory speech. *International Journal of Behavioral Development*, 37(2), 106–110.
<https://doi.org/10.1177/0165025412473016>
- Lansing, C. R., & McConkie, G. W. (1999). Attention to Facial Regions in Segmental and Prosodic Visual Speech Perception Tasks. *Journal of Speech, Language, and Hearing Research*, 42, 526–539.
- Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception and Psychophysics*, 65(4), 536–552. <https://doi.org/10.3758/BF03194581>
- Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences of the United States of America*, 109(5), 1431–1436.
<https://doi.org/10.1073/PNAS.1114783109>
- 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. *The Journal of the Acoustical Society of America*, 94(3), 1242–1255. <https://doi.org/10.1121/1.408177>
- Magnuson, J. S., & Nusbaum, H. C. (2007). Acoustic Differences, Listener Expectations, and the Perceptual Accommodation of Talker Variability. *Journal of Experimental*

- Psychology: Human Perception and Performance*, 33(2), 391–409.
<https://doi.org/10.1037/0096-1523.33.2.391>
- Marassa, L. K., & Lansing, C. R. (1995). Visual word recognition in two facial motion conditions: Full-face versus lips-plus-mandible. *Journal of Speech and Hearing Research*, 38(6), 1387–1394. <https://doi.org/10.1044/jshr.3806.1387>
- Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, 11(3), F9. <https://doi.org/10.1111/J.1467-7687.2008.00671>
- Martin, C. S., Mullennix, J. W., Pisoni, D. B., & Summers, W. v. (1989). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(4), 676–684. <https://doi.org/10.1037/0278-7393.15.4.676>
- Mcdaniel, J., Camarata, S., & Yoder, P. (2018). Comparing Auditory-Only and Audiovisual Word Learning for Children With Hearing Loss. *Journal of Deaf Studies and Deaf Education*, 382–398. <https://doi.org/10.1093/deafed/eny016>
- McDermott, J. H., Wroblewski, D., & Oxenham, A. J. (2011). Recovering sound sources from embedded repetition. *Proceedings of the National Academy of Sciences of the United States of America*, 108(3), 1188–1193.
https://doi.org/10.1073/PNAS.1004765108/SUPPL_FILE/PNAS.201004765SI.PDF
- McMillan, B. T. M., & Saffran, J. R. (2016). Learning in Complex Environments: The Effects of Background Speech on Early Word Learning. *Child Development*, 87(6), 1841–1855.
<https://doi.org/10.1111/CDEV.12559>
- Miller, S., Zhang, Y., & Nelson, P. (2016). Neural correlates of phonetic learning in postlingually deafened cochlear implant listeners. *Ear and Hearing*, 37(5), 514–528.
<https://doi.org/10.1097/AUD.0000000000000287>
- Misurelli, S. M., & Litovsky, R. Y. (2015). Spatial release from masking in children with bilateral cochlear implants and with normal hearing: Effect of target-interferer similarity. *The Journal of the Acoustical Society of America*, 138(1), 319.
<https://doi.org/10.1121/1.4922777>
- Morales, M., Mundy, P., Delgado, C. E. F., Yale, M., Messinger, D., Neal, R., & Schwartz, H. K. (2000). Responding to Joint Attention Across the 6- Through 24-Month Age Period and Early Language Acquisition. *Journal of Applied Developmental Psychology*, 21(3), 283–298. [https://doi.org/10.1016/S0193-3973\(99\)00040-4](https://doi.org/10.1016/S0193-3973(99)00040-4)

- Morales, M., Mundy, P., Delgado, C. E. F., Yale, M., Neal, R., & Schwartz, H. K. (2000). *Gaze following, temperament, and language development in 6-month-olds: A replication and extension*.
- Morales, M., Mundy, P., & Rojas, J. (1998). Following the direction of gaze and language development in 6-month-olds. *Infant Behavior and Development*, 21(2), 373–377. [https://doi.org/10.1016/S0163-6383\(98\)90014-5](https://doi.org/10.1016/S0163-6383(98)90014-5)
- Morini, G., & Newman, R. S. (2020). Monolingual and Bilingual Word Recognition and Word Learning in Background Noise. *Language and Speech*, 63(2), 381–403. <https://doi.org/10.1177/0023830919846158>
- Morin-Lessard, E., Poulin-Dubois, D., Segalowitz, N., & Byers-Heinlein, K. (2019). Selective attention to the mouth of talking faces in monolinguals and bilinguals aged 5 months to 5 years. *Developmental Psychology*, 55(8), 1640–1655. <https://doi.org/10.1037/dev0000750>
- Mullennix, J. W., Pisoni, D. B., & Martin, C. (1998). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, 85(1), 365. <https://doi.org/10.1121/1.397688>
- Mundy, P. (1998). Individual differences in joint attention skill development in the second year. *Infant Behavior and Development*, 21(3), 469–482. [https://doi.org/10.1016/S0163-6383\(98\)90020-0](https://doi.org/10.1016/S0163-6383(98)90020-0)
- Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & Psychophysics*, 60(3), 355–376. <https://doi.org/10.3758/BF03206860>
- Pare, M., Richler, R. C., ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics*, 65(4), 553–567.
- Parikh, G., & Loizou, P. C. (2005). The influence of noise on vowel and consonant cues Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing Spectral and temporal cues for phoneme recognition in noise The influence of noise on vowel and consonant cues. *Citation: The Journal of the Acoustical Society of America*, 118, 1609. <https://doi.org/10.1121/1.2118407>
- Paris, T., Kim, J., & Davis, C. (2016). Using EEG and stimulus context to probe the modelling of auditory-visual speech. *Cortex*, 75, 220–230. <https://doi.org/10.1016/j.cortex.2015.03.010>

- Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. In *Cortex* (Vol. 68, pp. 169–181).
<https://doi.org/10.1016/j.cortex.2015.03.006>
- Pekkola, J., Ojanen, V., Autti, T., Jääskeläinen, I. P., Möttönen, R., & Sams, M. (2006). Attention to visual speech gestures enhances hemodynamic activity in the left planum temporale. *Human Brain Mapping*, 27(6), 471–477.
<https://doi.org/10.1002/HBM.20190>
- 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 and Hearing*, 39(5), 980–991.
<https://doi.org/10.1097/AUD.0000000000000560>
- Pons, F., Bosch, L., & Lewkowicz, D. J. (1982). Twelve-month-old infants' attention to the eyes of a talking face is associated with communication and social skills. & Sebastián-Gallés. <https://doi.org/10.1016/j.infbeh.2018.12.003>
- Pons, F., Bosch, L., & Lewkowicz, D. J. (2015). Bilingualism Modulates Infants' Selective Attention to the Mouth of a Talking Face. *Psychological Science*, 26(4), 490–498.
<https://doi.org/10.1177/0956797614568320>
- Rigler, H., Farris-Trimble, A., Greiner, L., Walker, J., Tomblin, J. B., & McMurray, B. (2015). The slow developmental timecourse of real-time spoken word recognition. *Developmental Psychology*, 51(12), 1690. <https://doi.org/10.1037/DEV0000044>
- Riley, K. G., & McGregor, K. K. (2012). Noise hampers children's expressive word learning. *Language, Speech, and Hearing Services in Schools*, 43(3), 325–337.
[https://doi.org/10.1044/0161-1461\(2012/11-0053\)](https://doi.org/10.1044/0161-1461(2012/11-0053))
- Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, 12(2), 339–349.
<https://doi.org/10.1111/j.1467-7687.2008.00786.x>
- Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, 15(6), 608–635.
<https://doi.org/10.1111/j.1532-7078.2010.00033.x>
- Smith, N. A., Gibilisco, C. R., Meisinger, R. E., & Hankey, M. (2013). Asymmetry in infants' selective attention to facial features during visual processing of infant-directed speech. *Frontiers in Psychology*, 4(SEP), 601.
<https://doi.org/10.3389/fpsyg.2013.00601>

- Srinivasan, R. J., & Massaro, D. W. (2003). Perceiving Prosody from the Face and Voice: Distinguishing Statements from Echoic Questions in English*. *Language and Speech*, 46(1), 1–22.
- Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, 26(2), 212–215.
<https://doi.org/10.1121/1.1907309>
- 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(4), 534–553. <https://doi.org/10.1111/j.1532-7078.2012.00135.x>
- Tenenbaum, E. J., Sobel, D. M., Sheinkopf, S. J., Malle, B. F., & Morgan, J. L. (2015). Attention to the mouth and gaze following in infancy predict language development. *Journal of Child Language*, 42(6), 1173–1190.
<https://doi.org/10.1017/S0305000914000725>
- Thomas, S. M., & Jordan, T. R. (2004). Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, 30(5), 873–888. <https://doi.org/10.1037/0096-1523.30.5.873>
- Tsang, T., Atagi, N., & Johnson, S. P. (2018). Selective attention to the mouth is associated with expressive language skills in monolingual and bilingual infants. *Journal of Experimental Child Psychology*, 169, 93–109.
<https://doi.org/10.1016/j.jecp.2018.01.002>
- Tun, P. A., Wingfield, A., & Stine, E. A. (1991). Speech-processing capacity in young and older adults: a dual-task study. *Psychology and Aging*, 6(1), 3–9.
<https://doi.org/10.1037/0882-7974.6.1.3>
- Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and Visual Lexical Neighborhoods in Audiovisual Speech Perception. *Trends in Amplification*, 11(4), 233–241. <https://doi.org/10.1177/1084713807307409>
- van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences*, 102(4), 1181–1186. <https://doi.org/10.1073/pnas.0408949102>
- Vatikiotis-Bateson, E., Eigsti, I.-M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics*, 60(6), 926–940. <https://doi.org/10.3758/BF03211929>

- Walker, E. A., & McGregor, K. K. (2013). Word Learning Processes in Children With Cochlear Implants. *Journal of Speech Language and Hearing Research*, 56(2), 375. [https://doi.org/10.1044/1092-4388\(2012/11-0343\)](https://doi.org/10.1044/1092-4388(2012/11-0343))
- Weatherhead, D., Arredondo, M. M., Garcia, L. N., & Werker, J. F. (2021). The Role of Audiovisual Speech in Fast-Mapping and Novel Word Retention in Monolingual and Bilingual 24-Month-Olds. *Brain Sciences* 2021, Vol. 11, Page 114, 11(1), 114. <https://doi.org/10.3390/BRAINSCI11010114>
- Wong, P. C. M., Nusbaum, H. C., & Small, S. L. (2004). Neural Bases of Talker Normalization. *Journal of Cognitive Neuroscience*, 16(7), 1173–1184. <https://doi.org/10.1162/0898929041920522>
- Yi, A., Wong, W., & Moshe, E. (2013). Gaze Patterns and Audiovisual Speech Enhancement. *Journal of Speech, Language, and Hearing Research*, 56(2), 471–480. [https://doi.org/10.1044/1092-4388\(2012/10-0288\)](https://doi.org/10.1044/1092-4388(2012/10-0288))