12-2017

Do you see what I mean? The role of visual speech information in lexical representations

Ryan Andrew Cannistraci
University of Tennessee, rcannist@vols.utk.edu

To the Graduate Council:

I am submitting herewith a thesis written by Ryan Andrew Cannistraci entitled "Do you see what I mean? The role of visual speech information in lexical representations." I have examined the final electronic copy of this thesis for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Master of Arts, with a major in Psychology.

Jessica Hay, Major Professor

We have read this thesis and recommend its acceptance:

Aaron Buss, Devin Casenhiser, Daniela Corbetta

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Do you see what I mean? The role of visual speech information in lexical representations

A Thesis Presented for the

Master of Arts

Degree

The University of Tennessee, Knoxville

Ryan Andrew Cannistraci

December 2017

Acknowledgements

I first want to thank my advisor, Jessica Hay, for her support and guidance throughout this project. I thank my committee members: Aaron Buss, Daniela Corbetta, and Devin Casenhiser for their input on the design and execution of this project. Additionally, I want to thank Caglar Tas, Shannon Ross-Sheehy, Greg Reynolds, and James Todd, as well as my fellow graduate students, lab managers, and undergraduate research assistants in the Infant Language and Perceptual Learning Lab for their helpful suggestions on the design, coding, analyses, and writing presented here. I want to especially thank Johanna Lohman for recording the speech stimulus videos used in the work presented here. Most importantly, I want to thank my parents, Dan and Elizabeth Cannistraci, for their unwavering love and support.

Abstract

Human speech is necessarily multimodal and audiovisual redundancies in speech may play a vital role in speech perception across the lifespan. The majority of previous studies have focused particularly on how language is learned from auditory input, but the way in which audiovisual speech information is perceived and comprehended remains less well understood. Here, I examine how audiovisual and visual-only speech information is represented for known words, and if intersensory processing efficiency ability predicts the strength of the lexical representation. To explore the relationship between intersensory processing ability (indexed by matching temporally synchronous auditory and visual stimulation) and the strength of lexical representations, adult subjects participated in an audiovisual word recognition task and the *Intersensory Processing Efficiency Protocol* (IPEP). Participants were able to reliably identify a correct referent object across manipulations of modality (audiovisual vs visual-only) and pronunciation (correctly vs mispronounced). Correlational analyses did not reveal any relationship between processing efficiency and visual speech information in lexical representations. However, the results presented here suggest that adults' lexical representations robustly include visual speech information and that visual speech information is sublexically processed during speech perception.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

A fundamental concept central to theories of language processing is that language knowledge is represented in the lexicon. Traditionally, language processing theories posit that phonemes, the smallest sound units of speech perception, are used to access lexical representations (Studdert-Kennedy, 1976). These lexical representations have been theorized to encompass phonological, morpho-syntactic, and semantic information that can be flexibly used to process and comprehend speech (Marslen-Wilson, 1992; Marslen-Wilson, Brown, & Tyler, 1988). This work is important because speech must be perceived and comprehended in real time, where the smallest differences in acoustics can change the meaning of a word, phrase, or sentence. Because phonological representations trigger access to lexical representations, previous research with adult participants has assessed the specificity of how phonological information is stored (for a review, see Kazanina, Bowers, & Idsardi, 2017).

While studying lexical representations through auditory speech processing has proven to be fruitful, we know much less about how visual speech information is represented. Speech is inherently multimodal (Rosenblum, 2008; Campbell, 2008), since the visible facial movements used to articulate speech are redundant to the speech sounds a speaker produces. A growing body of research suggests that infants and adults are sensitive to the redundancies of audiovisual speech (e.g. Lewkowicz, 2010; McGurk & MacDonald, 1976). In fact, audiovisually redundant information can augment adult speech comprehension in noisy environments (Sumby & Pollack, 1954) as well as facilitate infant cognitive development (Bahrick & Lickliter, 2000; 2002; 2004). However, little is known about how infants and adults represent visual speech information. Thus,

in this study, I explore the intersection between individual differences in the ability to process audiovisual speech information and the strength of visual speech representations in the lexicon.

Lexical Representations

Classically, lexical representations have been studied in the auditory domain with adult populations using a neighborhood density paradigm. Stemming from early theoretical work (Studdert-Kennedy, 1976) and computational models (Gaskell & Marslen-Wilson, 1997; Massaro, 1989; McClelland & Elman, 1986; Norris, 1994), Vitevitch & Luce (1999) studied the process of accessing lexical representations from phonological information. Vitevitch & Luce (1999) proposed that two factors are in play when accessing lexical representations—probabilistic phonotactics and neighborhood density. Probabilistic phonotactics are described as the relative frequency of segments in typically occurring words. For example /-en/ is highly probabilistic (i.e. occurs often in English words; e.g. “pen”) while /-rm/ occurs less often (e.g. “worm”). Neighborhood density is described as the number of words that are phonologically similar to a given word. For example, the word *cat* is estimated to have 45 phonological neighbors in English (Vaden, Hickok, & Halpin, 2009). That is, 45 legitimate English words can be created by adding (*scat*), subtracting (*at*), or changing (*pat, cot, cap*) one of the three phonemes in the word *cat*. Vitevich & Luce (1999) found that larger neighborhood densities slowed lexical retrieved, and similarly, words with high probabilistic phonotactics also slowed lexical retrieval. These results suggest that similar-sounding phonotactic sequences create competition at the lexical level, evidenced by the slowed reaction times due to greater neighborhood densities and higher probabilistic phonotactics.

In addition to the evidence in Vitevich & Luce (1999), it has further been suggested that adults parse phonemes, morphemes, and lexical items simultaneously in speech perception.

Vitevitch (2003) replicated the findings of Vitevich & Luce (1999), and again found that words with high phonotactic probabilities and large neighborhood densities elicited slower reaction times than words with low phonotactic probabilities and sparsely populated neighborhood densities. Vitevich (2003) then extended this paradigm to pseudowords and found the same effect, that adults used sublexical representations to process pseudowords. Pseudowords with high probabilistic sequences in English were processed more efficiently than pseudowords with low probabilistic phonotactics. Pitt & Samuel (2006) studied lexical and sublexical retrieval by systematically varying word length, and found that longer words had more robust lexical activation, evidenced by quicker reaction times in a response task. Pitt & Samuel (2006) offer that while the longer words had more phonemes, sublexical processing limited the number of potential neighbors and led to better recognition. However, the short words had a much greater neighborhood density, which in turn slowed reaction times. The neighborhood density and probabilistic phonotactics literature provides evidence that speech comprehension entails a number of online, moment-to-moment strategies to process linguistic information in the auditory domain.

Developmentalists have also been interested in questions about lexical representations, particularly how they are formed through learning processes. As it turns out, studying lexical representations in the infant literature has been more difficult than studying adult lexical representations (for a review, see Newman, 2008). The earliest work on infants' lexical representations focused on learning minimal pair words—words that differ by only a single phoneme (e.g. bin and din). Stager & Werker (1997) demonstrated that 14-month-old infants have a difficult time mapping minimal pair words to novel objects. Using the Switch Task (Werker et al., 1998), infants were habituated to two novel label-object pairs and tested on their

ability to differentiate Same Trials, in which the original label-object pairs are maintained, from Switch Trials, in which the Object A is paired with Label B or vice versa. Violations of the label-object pairings led to dishabituation if the labels were phonologically distinctive (e.g., *lif* and *neem*), but went unnoticed when the two labels were similar sounding minimal pairs (e.g. bih and dih). However, by 17 months of age infants were able to attend to small phonetic differences and successfully map minimal pair words to novel objects (Fennell & Werker, 2004). These results suggest that lexical representations are weakly represented early in the learning process. However, in a word recognition study, 14-month-old infants are able to accurately map familiar minimal pair words (i.e. ball and doll) to referent objects, which suggests that infants have stronger representations of words they have real-world experience with at 14-months of age (Fennell & Werker, 2003).

In addition to studying lexical representations using minimal pair associative learning tasks, a second line of research has aimed to understand the specificity of lexical representations in early development using mispronunciation paradigms. For example, Swingley & Aslin (2000) tested 18-23-month-old's lexical representations by comparing looking accuracy and reaction time for correctly pronounced vs mispronounced commonly known words (e.g. doggy, baby, etc.). Swingley & Aslin (2000) reasoned that if young children have well-specified representations of known words, the mispronunciations should alter their ability to match the phonological form of the word to a referent picture. Conversely, if the children had less-well specified lexical representations of the known words, mispronunciations should not have an effect of looking accuracy or reaction time. Swingley & Aslin (2000) demonstrated that infants' reaction times are affected by mispronunciations of known words, which suggests that mispronunciations *impair*, but do not *inhibit* recognition of familiar words for young children. In

follow up studies Swingley and colleagues (Swingley, 2003; Swingley & Aslin, 2002; 2007) found that mispronunciation effects are not influenced by phonologically similar words, strengthening the hypothesis that auditory lexical representations are specifically defined and robust.

Taken together, the adult literature and developmental work on lexical representations provides a clear picture that auditory lexical representations become robust and specific early in language development, and subsequently persist through adulthood. While the literature has focused on the content and online processing of auditory information in lexical representations, there is also a growing body of literature to suggest that visual speech information may be included in lexical representations (Woollams, 2015). Support for this idea comes from the theoretical stance that human's sensory environments are richly intersensory (Barrett, 2011; Campbell, 2008; Gibson, 1966, 2014; James, 1890; Rosenblum, 2008; Sumby & Pollack, 1954). Further, speech is inherently multimodal (Campbell, 2008; Rosenblum, 2008) and includes language specific auditory and visual speech information. Visual speech information can be defined as the information of the visible facial movements used to produce speech, and these visible movements are redundant to the auditory stream that is produced. Even though linguistic information is redundant across auditory and visual modalities, much less is known about how visual speech information is represented in the lexicon.

Audiovisual Speech

Sensitivity to the audiovisual nature of speech has been demonstrated incredibly early in infancy. As early as two months of age, human infants are able to link aurally presented vowel sounds to facial movements by attending to a correctly articulating face as opposed to an incorrectly articulating face (Patterson & Werker, 2003). Four-month-olds can detect audiovisual

asynchrony in speech perception tasks (Lewkowicz, 2010) and 5-month-old infants preferentially attend to congruent, rather than incongruent audiovisual speech (Kuhl & Meltzoff, 1984). At six-months, visual articulatory information enhances phoneme discrimination, suggesting that audiovisual redundancies may augment the learning of phonetic boundaries in infancy (Teinonen, Aslin, Alku, & Csibra, 2008). Further, Hollich, Newman, & Jusczyk (2005) demonstrated that 7.5-month-old infants are able to selectively attend to a speech stream when a distractor stream is present, if congruent visual information is available. In addition to being sensitive to the audiovisual redundancies in speech, infants are able to use visual information alone to discriminate their native language from an unknown language (Weikum et al., 2007). Four-, 6-, and 8-month-old infants were habituated to visual utterances in a single language (e.g. English) and were then tested on their ability to distinguish an utterance in the same language as habituation (e.g. English) from a second utterance in a different language (e.g. French). Both 4- and 6-month-olds looked for a longer duration to the switch trial (foreign language) than the same trial (native language), suggesting the infants were able to discriminate between their native language and a foreign language based on visual information alone.

The infant literature shows that humans are sensitive to the audiovisual nature of speech early in the lifespan. This sensitivity to audiovisual speech continues to strengthen across the lifespan. One striking example of audiovisual sensitivity in speech perception is the McGurk Effect. In McGurk & MacDonald's (1976) seminal work, they inadvertently violated the typically redundant nature of audiovisual speech information while dubbing audio and video recordings, which then created illusory percepts of audiovisual speech. The canonical example of the McGurk effect is composed of the syllables /ba/ and /ga/. These two syllables are articulated differently and elicit distinctive facial movements. For example, /ba/ is bilabial, meaning the

consonant closure happens at the lips. Conversely, /ga/ is velar, meaning the consonant closure occurs as the back of the tongue makes contact with the velum (top of the mouth). Thus, the visual speech information for /b/ versus /g/ sounds is distinct. In McGurk and MacDonald's seminal study, pre-school children and adult participants reported experiencing the syllable /da/ when viewing an audiovisual stimulus composed of an auditory /ba/ and visual /ga/. This emergent /da/ percept was not actually present in either the auditory or visual signal (e.g. a fused percept). Conversely, when the /ba/ and /ga/ phonemes are switched in dubbing (e.g. visual /ba/ is dubbed with an auditory /ga/), the fused /da/ was reported with lesser frequency, but combination percepts composed of both the auditory and visual domains were also reported (e.g., /gabga/, /bagba/, /baga/, or /gaba/). Since McGurk and MacDonald (1976), there are been many attempts to study the generalizability of the McGurk Effect across other stimulus combinations (e.g., Desjardins & Werker, 2004; MacDonald & McGurk, 1978; Rosenblum, Schmuckler, & Johnson, 1997). The McGurk effect has since been robustly replicated in numerous cross-linguistic adult studies (e.g., Bovo et al., 2009; Munhall, Gribble, Sacco, & Ward, 1996; Sekiyama, 1997; Sekiyama, Soshi, & Sakamoto, 2014) and in the infant literature (e.g., Burnham & Dodd, 1996, 2004; Desjardins & Werker, 2004; Rosenblum, Schmuckler, & Johnson, 1997). Though the McGurk illusion is synthetically induced, this extensive literature may be telling of how humans represent audiovisual speech information (for an opposing view, see Alsius, Paré, & Munhall, 2017).

Lexical Access in Audiovisual and Visual-only Speech

While the infant literature provides evidence of sensitivity to audiovisual speech early in the lifespan, limitations of our methods constrain our understanding of how infants may be able to use visual speech information in speech perception for lexical access. However, we can

address questions of functional use (i.e. comprehension) in adult participants. Previous research suggests that visual speech information facilitates speech comprehension (i.e. lexical access) in noisy environments (Hollich, Newman, & Jusczyk, 2005; Sumby & Pollack, 1954) and visual speech information in a priming task has also been shown to facilitate accurate lexical retrieval (Fort et al., 2013). In addition to the empirical evidence that the McGurk Effect functions on a perceptual level, there is further evidence to suggest that audiovisual integration in McGurk-type percepts can be used to trigger lexical access. Brancazio (2004) tested lexical influences on McGurk Effect perceptions in adults. Participants viewed a speaker on a computer screen articulating a word and simultaneously heard a temporally-synced and length-matched auditory word stream—simulating the audiovisual percept of a word. Half of the trials were audiovisually congruent and the other half of the trials were audiovisually incongruent in the auditory and visual domains, designed to create McGurk effect-like stimuli. The participants were asked to type into a computer prompt their perception of the word initial sound and rate the goodness of their word initial perception (i.e. if their perception seemed like a good consonant in English, or a nonsensical combination of sounds). McGurk effect percepts were most frequent when the resulting percept was a real word and when the auditory signal was not a real word (e.g. auditory "besk" visual "gesk" to produce "desk"). However, when the auditory stimuli used to create a McGurk illusion was a real word (e.g. auditory "beg" visual "geg" to produce "deg"), the McGurk-fusion percept ("deg") was reported less often, and the goodness rating for this type of trials was significantly lower than goodness rating for trials that produced real words. Barutchu et al. (2008) replicate these findings for word-initial audiovisual discrepancies, but also show that word-final discrepancies that should create McGurk effect percepts occur less frequently, as

top-down lexical knowledge of the lexical representation has been triggered from the preceding sounds of the word.

In fact, visual speech information alone is sufficient to access lexical representations. In an experiment to assess the specificity of visual speech information in lexical representations, Tye-Murray, Sommers, & Spehar (2007) present evidence that visual lexical neighborhoods affect audiovisual speech perception in adults. In a recognition task, participants viewed three trial types: audiovisual presentation, auditory only presentation, and visual only presentation. The authors defined visual neighborhood density conservatively by only using words that differ by the first phoneme. During word recognition, participants showed better word recognition for words with lower auditory and visual neighborhood densities as compared to words that had higher neighborhood densities. Thus, these results suggest visual speech information can influence lexical retrieval and that visual speech information is accessed in a similar manner as auditory speech information, evidenced by words with higher visual neighborhood densities being more difficult to access.

Extending the findings of Tye-Murray, Sommers, & Spehar (2007), a recent study by Havy, Foround, Fais, & Werker (2017) aimed to explore if 18-month-olds and adults were able to learn a new word form by solely visual speech information. Infants and adults were both successful in learning new acoustic forms in the auditory modality and able to generalize word recognition to visual-only word forms at test (i.e. the visual articulations of the words learned). However, only adults were able to successfully learn new word forms from the visual speech information alone. These results are quite interesting, especially for the infants as they were able to generalize their representation of words they had learned in the auditory modality to the visual modality, even though they had not been provided with redundant audiovisual cues. This

suggests that both auditory and visual information is stored in infants' phonological representation. Thus, while the infants were unable to learn from the visual information alone, the ability to recognize new word forms in a visual only condition in test provides insight to the strength of audiovisual coupling in lexical representations. In addition, the adult results suggest that people with more experience to audiovisual redundancies (as a function of age) may represent visual speech information in a more robust manner, evidenced by the adults learning of the new word forms based on the visual speech information alone.

The current study

The literature discussed here provides evidence that visual speech information is lexically represented, and further, can influence audiovisual speech perception. Evidence from Tye-Murray Sommers, & Spehar (2007) suggests that the visual speech representations are stable, specific, and able to be used for lexical access. Havy et al. (2017) further provide evidence that adults are able to learn new word forms solely from visual speech information, albeit in a tightly controlled experimental task. The first aim of this study is to further uncover the specificity of visual-speech representations of known words and potential individual differences in audiovisual processing using a mispronunciation paradigm. This replication-and-extension of Swingley & Aslin (2000) will examine how visual speech information is represented in the lexicon and used for speech comprehension in audiovisual and visual-only domains.

After assessing the specificity of audiovisual and visual-only lexical representations, the second aim of this study is to address the relationship between multimodal processing and the nature of the lexical representations. Visual speech perception is prone to large individual differences (Havy et al., 2017; Stevenson, Zemtsov, & Wallace, 2012). It is possible that individual differences in intersensory processing (here, the efficiency of mapping of auditory and

visual information) may be related to the strength of visual information stored in one's lexical representations. To disambiguate individual differences in multimodal processing, I used a new protocol, the *Intersensory Processing Efficiency Protocol* (IPEP; Bahrick, 2017). In the IPEP, participants must bind a stream of auditory information to one specific visual event out of an array. The correct, temporally matching and synchronous visual event is presented alongside five potential distractor videos, thus assessing the participant's efficiency in matching the redundant auditory and visual information. Efficiency in intersensory processing is measured by the proportion of looking time to the correct visual stimulus in comparison to the distractors as well as the participant's reaction time to find the correct target. To my knowledge, this work presents the first exploration of possible commonalities that underlie general audiovisual matching and potential links to individual differences in how lexical representations are structured.

In the audiovisual word recognition task, I predict that the quickest reaction time and most accurate looking behavior will be evidenced in audiovisual, correctly pronounced trials. If participants are able to sublexically process audiovisual speech information on mispronounced audiovisual trials, I expect to observe slight deficits in accuracy and reaction time performance, though well above chance. This pattern of results would conceptually replicate and extend a host of studies on mispronunciations and sublexical processing in audiovisual speech stimuli, as opposed to auditory-only stimuli (Swingley, 2000; Vitevich & Luce, 1999; Vitevich et al., 1997). Further, if visual speech information is included in lexical representations, adults should quickly and accurately identify the correct target for visual-only, correctly pronounced trials (consistent with Havy et al., 2017; Tye-Murray et al., 2007). Further, if participants can sublexically process visual speech information, they should be able to find the correct target for visual-only, mispronounced trials. Vitevich & Luce (1999) and Vitevich et al. (1997) provide evidence that

sublexical processing occurs in the auditory domain, but this is the first study to my knowledge to test if sublexical processing occurs for visual-only speech information.

Additionally, this is the first study to my knowledge that tests for individual differences in the relationship between multimodal (i.e. audiovisual) processing and the content of lexical representations. There is evidence that robust individual differences are present for both multimodal matching (Stevenson, Zemtsov, & Wallace, 2012) and visual-only speech perception (Havy et al., 2017). The extant literature on aging and speech perception suggests that the link between auditory and visual speech information strengthens across the lifespan, both in illusory audiovisual effects like the McGurk effect (Sekiyama, Soshi, & Sakamoto, 2014) and in natural speech perception (Winneke & Phillips, 2011). This audiovisual strengthening effect seen across the lifespan may influence an individual's attention to the visual information of the mouth in speech perception. It is possible that participants who are better at audiovisual mapping in the real world have more robust representations of the links between auditory-only and visual-only information. Thus, I predict that greater intersensory processing ability, as measured by faster reaction times and greater accuracy in the IPEP, will be correlated to faster and more accurate lexical retrieval.

Chapter 2

Materials and Methods

Participants

Thirty-eight undergraduate students (22 females, 16 males; 18-25 years-old) participated in the current study. Participants were recruited through the University of Tennessee, Knoxville's SONA participant database and by word of mouth. A demographic survey, administered after informed consent and completion of the study, insured that all participants were native monolingual speakers of American English with normal hearing and normal/corrected-to-normal vision. All participants reported they were able to clearly see and hear all stimuli during the entirety of both experimental tasks.

Apparatus

The participants sat in a chair approximately 60 centimeters from the computer monitor that displayed the stimuli. Auditory stimuli were played from fixed speakers, located directly behind the computer monitor. A Tobii x60 eye-tracker was mounted below the display monitor to track each participant's eye-gaze data. Each testing session began with a 5-point calibration phase to ensure the participant's corneal reflection was accurately picked up by the machine throughout the bounds of the screen (Dautriche, Swingley, & Christophe, 2015). The audiovisual word recognition task was run exclusively through Tobii Studio and while the IPEP also used Tobii Studio to record eye-gaze data, stimulus presentation was run using a custom-designed Matlab protocol (Bahrick, 2017).

Audiovisual word recognition task: Description

The audiovisual word recognition task was designed to measure each participant's lexical retrieval of known words, across various stimulus manipulations. There were two independent

variables: modality (audiovisual and visual-only) and pronunciation (correctly pronounced and mispronounced). The dependent variables were accuracy and reaction time to locate the correct referent object, in the presence of one potential distractor object.

Audiovisual word recognition task: Stimuli

Each trial included digital photographs of familiar referent objects and videos of a monolingual American English speaker articulating a carrier phrase and a target word (e.g., Where's the [baby]? Can you find it?). Referent objects and target words were as follows: *baby*, *doggy, kitty, ducky, shoe, car, ball*, and *fish* (see Table 1). Diminutive word forms were used in the hopes of comparing adults' performance in this task to infants' performance in later studies. The diminutive word forms were chosen to facilitate infants' interest in the task in later tests. While the diminutive word forms may seem puerile for the adult participants, the initial articulation of each diminutive is constant with the canonical, adult word form (i.e. *cat* and *kitty* both are initially articulated with the voiceless, velar, stop consonant). The words *horse* and *monkey* were used during warmup trails.

A brief pause between the carrier phrase and the target word ensured that there was no co-articulation leading into the target word. Auditory stimuli were played at approximately 65dB. The digital photographs used as referents were normalized for size and saliency. The videos of the speaker articulating the carrier phrase and target word were recorded using a Nikon D3300 DX camera (DSLR Kit) with an 18-55mm f/3.5-5.6G VR II Auto Focus-S DX NIKKOR Zoom Lens. Audio of the speaker's utterances were recorded in Praat (Boersma & Weenink, 1996) using a Blue Snowball USB Microphone and were RMS matched in Adobe Audition®. The audio and video recordings were synced and cropped in iMovie 7.1. The digital photographs of the referent objects and speaker videos were imported into Motion 5 where they were fit into a

proprietary template to ensure the same display settings for each trial.

In the audiovisual word recognition task, there were four different trial types that correspond to the modality and pronunciation of the target word that was presented in each trial (see Table 2). Thus, there were audiovisual correctly pronounced, audiovisually mispronounced, visual-only correctly pronounced, and visual-only mispronounced trial types. Visual-only trials are named as such because the only linguistic information available was through the visual domain (i.e. the lip movements of articulating the target word). In the visual-only trials, the target words were transformed into pink noise using Adobe Audition® (see Table 2). The newly-created audio file with pink noise replacing the target word was then dubbed back onto the original video in iMovie. This ensured that on visual-only trials, participants would be presented with a continuous stream of auditory stimulation in the same acoustic register of human speech, but that they would receive no auditory linguistic information. There were two counterbalanced orders and specific target words were presented in only one modality across the task. This was done to eliminate any possible transfer effects from experience with audiovisual stimuli informing subsequent visual-only perception. Thus, in order 1, *baby*, *ducky*, *ball*, and *fish* were always presented audiovisually, while *doggy*, *kitty*, *car*, and *shoe* were always presented visually—and vice versa for order 2. Participants were randomly assigned to order 1(n=18) or order 2 (n=20).

For each trial, the speaker was located at the top-center of the display and the two referent objects were located at the bottom-left and bottom-right corners of the screen (see Figure 1). At trial onset, the speaker's face and both referent objects were present on the screen. The participants viewed this static image for approximately two seconds before the speaker articulated the carrier phrase and target word. Each trial was designed so that the onset of the

target word began exactly 4 seconds after the onset of the trial. Once the talker finished producing the target word, her face disappeared for the remainder of the trail, leaving only the two referent objects on the monitor. Gaze data for looking accuracy and reaction time were analyzed in a 2 second window following the onset of the target word (see Figure 2).

Audiovisual word recognition task: Procedure

The task began with two correctly pronounced warmup trials that were presented audiovisually. These trials were excluded from final analysis. Following the warmup trials, there were four blocks of 10 trials. Each block presented the target words in the same modality for all 10 trials. In order to eliminate any task-demand effects since participants transitioned between audiovisual speech perception in the audiovisual blocks to lip reading for the visual-only blocks, I included 2 correctly pronounced warmup trials at the beginning of each block. Thus, out of the 42 total trials, 32 trials were included in the analyses (see Figure 3). Of the eight trials in each block that were included in the final analyses, there were four correctly pronounced trials and four mispronounced trials in a pseudorandomized order. The correct target object was counterbalanced to appear at the left or right target location 50% of the time.

Audiovisual word recognition task: Measures

Reaction time

In this study, we used a modified *looking-while-listening paradigm* (Swingley, Fernald, & Pinto, 1999). In the traditional looking-while-listening paradigm, two referent objects are on the screen at all times, and thus reaction times can only be measured when a participant is fixated on the distractor object when they hear the object label, and subsequently shift their gaze to the correct referent object. The procedure used here is slightly different—reaction time was measured from fixations on the speaker's face to the correct target object. For any given trial to

be included in the data analyses, the participant must have been looking at the articulating face at the onset of the target word. This is vital for visual-only trials, where the only linguistic information available was conveyed through the visual modality. To be included in analyses, each participant needed usable reaction time data for at least 3 trials per condition. Only eye-movements to the correct target object initiated between 300 and 2000 milliseconds (ms) after target word onset were included in the reaction time analyses. Any eye movements that occurred before 300 ms could have been initiated prior to word onset, thus not indicative of the participant's lexical processing (Swingley, Fernald, & Pinto, 1999).

Accuracy

I assessed two looking accuracy metrics, the proportion of correct looking during the target window and the accuracy of the first look after word-onset. The analysis window for looking accuracy was from 300 ms to 2300 ms after the onset of the target word (Figure 2). The looking accuracy measure were calculated as a proportion of looking duration to the correct target, divided by looking duration to the incorrect target. This method of calculating looking accuracy is typically used in developmental work, and will be used here since the current task is based on a developmental task using similar analyses. However, since the adult participants tested here have a great deal of experience with all of the words they were tested on and may be less motivated to maintain their gaze on a labeled object, I also calculated a potentially more sensitive measure of looking accuracy—the accuracy of the first eye movement to either the correct or incorrect referent object after hearing the target word. The accuracy of the first look after word onset was simply coded as correct or incorrect for each trial, and then averaged for each stimulus condition.

Intersensory Processing Efficiency Protocol: Description

The Intersensory Processing Efficiency Protocol (IPEP) is a novel protocol designed by Lorraine Bahrick and colleagues to assess individual differences in processing speed and accuracy for multimodal events (Bahrick, 2017). The IPEP is an audiovisual search task that displays a 2x3 matrix of dynamic events (see Figure 4). The IPEP required participants to visually locate which visual event out of the 2x3 matrix matched a simultaneously played audio track. Thus, there was one sound-synchronized target and 5 asynchronous distractors.

Intersensory Processing Efficiency Protocol: Stimuli

Within the IPEP protocol, there were two different conditions – social and nonsocial. The dynamic visual events in the social trials were women reciting a children's story. The visual events in the nonsocial trials were solid objects making contact with a hard surface in an erratic, arrhythmic manner, produces knocking sounds. In each condition, the audio track for each trial matched one of the visual events—either in speech (social) or rhythmic pattern (nonsocial).

Intersensory Processing Efficiency Protocol: Procedure

Following the audiovisual word recognition task, all participants then did the IPEP task. The participants were already seated in the testing area and had given informed consent to participate in the study. Each participant was recalibrated using a 5-point calibration, standard to Tobii Studio. They were instructed to direct their gaze to the visual event that matched the auditory stream they heard. The IPEP had a total of 48 trials, each of which last for 8 seconds. The 48 trials were broken into four blocks (two social, two nonsocial), which were counterbalanced for order and target location across participants. Stimulus presentation was run using custom-designed Matlab software that interfaced with the eye-tracking software (Tobii Studio) to track each participant's eye gaze data.

Intersensory Processing Efficiency Protocol: Measures

Gaze data during the IPEP was recorded by a Tobii x60 eye-tracker. From the eye-tracking data, I was interested in measures of looking accuracy, measured by duration of fixation to the synchronous, correct target. This was calculated as a proportion—the duration of looking to the correct target divided by duration of looking to the distractors. Additionally, I extracted reaction time to fixate to the correct target, measured in milliseconds.

Chapter 3

Results

Audiovisual Word Recognition Task Analyses

A four-way mixed-model ANOVA did not reveal any main effects or interactions of sex or order. Additionally, item analyses revealed no significant differences between test items, so all analyses were collapsed across these variables. To examine the effects of modality (audiovisual vs visual-only) and pronunciation (correct vs mispronounced) on looking accuracy across the analysis window in the audiovisual word recognition task, I performed a 2x2 repeated measures ANOVA. The ANOVA revealed significant main effects of modality $F(1,37)=18.85$, $p<.001$, partial $\eta^2=.497$, power=1.000, and pronunciation $F(1,37)=89.71$, $p<.001$, partial $\eta^2=.520$, power=1.000, as well as a modality by pronunciation interaction $F(1,37)=15.80$, $p<.001$, partial $\eta^2=.261$, power=.941 (Figure 5). These effects demonstrate that participants were more accurate within the analysis window on audiovisual trials compared to visual-only trials and also more accurate on correctly pronounced trials compared to mispronounced trials. While group means for both pronunciations in the audiovisual modality and for visual-only correctly pronounced were at ceiling, the group mean accuracy for visual-only, mispronounced trials was much lower (M=62.76%). However, a single-sample t-test revealed that performance for visual-only mispronounce is reliably above chance $t(37)=3.88$, $p<.001$, d=.629, which demonstrates that while at a performance deficit, mispronounced visual-only speech information can successfully be used for lexical retrieval.

A second 2x2 repeated measures ANOVA was performed to compare the accuracy of first look after target word onset across modality and pronunciation manipulations. The ANOVA revealed significant main effects of modality $F(1,37)=8.547$, $p<.05$, partial $\eta^2=.188$, power=.812,

and pronunciation $F(1,37)=22.785$, $p<.001$, partial $\eta^2=.381$, power=.996. However, there was no significant modality by pronunciation interaction. Interestingly, group means for each trial type increased when only considering first look accuracy. This was especially the case for visual-only, mispronounced trials (M=81.94%), which exhibited an almost 20% increase in accuracy compared to the proportion of correct looking measures (Table 3).

A third 2x2 repeated measures ANOVA was performed to compare reaction times across modality and pronunciation. The ANOVA revealed significant main effects of modality $F(1,37)=36.50$, $p<.001$, partial $\eta^2=.497$, power=1.000, pronunciation $F(1,37)=40.12$, $p<.001$, partial $\eta^2=.520$, power=1.000 and a modality by pronunciation interaction $F(1,37)=13.10$, $p<.001$, partial $\eta^2=.261$, power=.941 (see Figure 6). These main effects demonstrate that participants had faster reaction times for audiovisual compared to visual-only trials, as well as faster reaction times for correctly pronounced vs mispronounced trials. Additionally, the interaction term suggests that the additive effects of the visual-only modality and word mispronunciation contributed to the slowest reaction times for the visual-only, mispronounced trials (see Table 4).

Intersensory Processing Efficiency Protocol and Correlational Analyses

Paired-samples t-tests revealed that participants had faster reaction times $t(37)=4.47$, $p<.001$, d=.731, and were more accurate $t(37)=5.86$, $p<.001$, d=1.048 for social trials, as compared to non-social trials. While descriptive, these group differences were not the main focus of the IPEP analyses because the social and nonsocial trials are fundamentally different from one another. Rather, a correlational analysis between looking accuracy and reaction time across both the audiovisual word recognition task and the IPEP were of great interest, in an attempt to discern any relationships between individual differences in audiovisual processing ability and

content of one's lexical representations (Table 5). The correlational analysis revealed that reaction time for audiovisual mispronounced trials were weakly correlated with non-social accuracy measures on the IPEP (r=.344, p=.035). While significant, these results are difficult to interpret because there are no systematic associations between processing ability, measured by the IPEP, and accuracy or reaction time metrics from the audiovisual word recognition task. In fact, the positive correlation reported suggests that longer reaction times for audiovisually mispronounced trials are related to better accuracy in non-social trials on the IPEP—which is contrary to my prediction that better audiovisual processing would facilitate faster lexical retrieval. In addition, there were no significant correlations between looking accuracy across the analysis window or first look accuracy and any of the IPEP measures.

Chapter 4

General Discussion and Conclusions

Discussion

The first prediction tested in this study was that visual speech information for known words is robustly represented in the lexicon and is able to be sublexically processed. To test this prediction, adult participants' looking accuracy and reaction time was compared across four trial types. These analyses present evidence that adults robustly access the correct lexical target for both correctly pronounced and mispronounced audiovisual words. In addition to these findings, that conceptually replicate and extend previous literature (Swingley, 2000; Vitevich & Luce, 1999; Vitevich et al., 1997), I also present evidence that adults readily access lexical representations when presented with visual-only speech information for both correctly pronounced and mispronounced target words. While looking accuracy is not as strong for visual-only mispronunciation trials and reaction times are significantly slower, adult participants are still able to identify the correct target word above the level of chance.

These results suggest that visual speech information is robustly represented in the lexicon. Further, the results also suggest that visual speech information is able to be sublexically processed. This is evidenced by the adults' ability to correctly identify the appropriate referent object when only visual speech information was present and when this visual-only speech information was mispronounced. The work from Vitevich and colleagues (1997, 1999) provides evidence that adults can readily sublexically process auditory speech information, and the results presented here extend our knowledge of sublexical processing into the domain of visual speech perception. The results here not only provide evidence that visual information is represented for

the facial movements across the entirety of a lexical representation, but also that the representations of visual speech information can be used in speech perception.

A second prediction was that individual differences in audiovisual processing are related to the manner in which visual speech information is represented in the lexicon. Audiovisual processing efficiency was measured using the IPEP (Bahrick, 2017) and the outcome metrics of looking accuracy and reaction time were not significantly correlated to looking accuracy or reaction time measures on the audiovisual word recognition task. There were a few spurious correlations (Table 5) but these correlations do not occur in a systematic pattern of associations between lexical representations and audiovisual processing efficiency measures. These results suggest that there is not an obvious relationship between the audiovisual composition of the lexicon and intersensory processing ability in the two laboratory tasks presented here.

Study Limitations

One potential limitation of this study is that this is the first use of the IPEP in conjunction with measures of lexical retrieval. I used the IPEP to ascertain the relationship of individual differences in audiovisual processing and content of lexical representations. The rational for doing so was based on evidence that while adults can use visual-only speech information for lexical retrieval, there are also large individual differences in accuracy and efficiency (Stevenson, Zemtsov, & Wallace, 2012; Havy et al., 2017). It may, in fact, be the case that the IPEP is not a sensitive enough measure to partial out individual differences of intersensory processing that relate to the composition of lexical representations. While the IPEP is able to measure individual differences in general intersensory processing, these individual differences may not extend into the processing of visual only speech information. It may also be the case that accuracy measures and reaction time measures are not monotonically related. Participants may

be using different visual scanning strategies to correctly identify the targets across the tasks. Some individuals may wait for a longer duration prior to making an eye movement to be sure they are directing a fixation to the correct object. If this is the case, reaction time would not accurately reflect processing efficiency for individuals who use a delayed scanning pattern in the audiovisual word recognition task, and not the IPEP. Thus, the correlational analyses presented here may not accurately assess the relationship between intersensory processing and lexical retrieval.

Future Directions

In future studies, ascertaining each participant's looking preference patterns or looking phenotypes may be helpful to identify potential relationships between looking strategies in addition to intersensory processing and lexical retrieval. Processing efficiency in a single modality (audition or vision) may be a better predictor of the relationship between processing efficiency and the content of lexical representations. A second option for future work is to scale up the difficulty of the audiovisual word recognition task. One way to do so would be to add more referent picture options (for example, four referent pictures) that systematically overlap in auditory or visual neighborhood densities, or sublexical phonotactic probabilities. In doing so, the multiple influences of auditory neighborhood density, visual neighborhood density, and probabilistic phonotactics on lexical retrieval could be assessed using a within subjects design, while also scaling-up task difficulty. Lastly, task difficulty could further be increased by making the experimental task a true visual search task by only presenting the referent object pictures after the target word is articulated. In its current manifestation, the task allows participants to view the referent objects before hearing the word, which lessens the difficulty of scanning behavior to find the correct target object. If the referent objects stay occluded until after the

target word is articulated, each participant would have to keep the perceived word form active in their working memory while visually scanning for the correct referent target.

Additionally, follow up studies could use a randomized trial order so that visual-only and audiovisual trials are mixed throughout the procedure. It is possible that the blocked design facilitated performance on visual-only trials. In recent work with infants, task performance was actually facilitated when stimuli were reliably presented in a single modality (Bahrick, Lickliter, & Castellanos, 2013). It stands to reason that a similar effect could manifest for the adults in our blocked design. In future studies using a similar procedure, this empirical question can be answered by examining if randomized trial sequences attenuates performance on visual-only trials.

Conclusions

Although I did not observe any meaningful relationships between audiovisual processing efficiency and the content of adults' lexical representations, I did find an interesting pattern of results in regard to performance on visual-only, mispronounced trials. While visual-only, mispronounced trials were characterized by observed performance deficits compared to the other conditions, the participants were able to reliably identify the correct target above chance levels. This novel pattern of results suggests that adults can sublexically process visual speech information and accurately retrieve corresponding lexical items. Future research aimed to increase task difficulty by altering trial order or availability of referent objects can further interrogate the findings presented here.

List of References

Alsius, A., Paré, M., & Munhall, K. G. (2017). Forty years after hearing lips and seeing voices: the McGurk Effect revisited. *Brill Online Books and Journals*.

Bahrick, L.E. (2017). Intersensory Processing Efficiency Protocol (IPEP). Databrary. Retrieved July 28, 2017 from http://doi.org/10.17910/B7.336.

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, 36(2), 190.

Bahrick, L. E., & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. *Advances in Child Development and Behavior*, *30*, 153-189.

Bahrick, L. E., & Lickliter, R. (2004). Infants' perception of rhythm and tempo in unimodal and multimodal stimulation: A developmental test of the intersensory redundancy hypothesis. *Cognitive, Affective, & Behavioral Neuroscience*, 4(2), 137-147.

Bahrick, L. E., Lickliter, R., & Castellanos, I. (2013). The development of face perception in infancy: intersensory interference and unimodal visual facilitation. *Developmental Psychology, 49*(10), 1919.

Barrett, L. (2011). Beyond the brain: How body and environment shape animal and human minds. Princeton University Press.

Barutchu, A., Crewther, S. G., Kiely, P., Murphy, M. J., & Crewther, D. P. (2008). When/b/ill with/g/ill becomes/d/ill: Evidence for a lexical effect in audiovisual speech perception. *European Journal of Cognitive Psychology, 20*(1), 1-11.

Boersma, P., & Weenink, D. (1996). Praat: A system for doing phonetics by computer [Computer program] (Rep. No. 132). Amsterdam, the Netherlands: University of Amsterdam, Institute of Phonetic Sciences.

Bovo, R., Ciorba, A., Prosser, S., & Martini, A. (2009). The McGurk phenomenon in Italian listeners. *Acta Otorhinolaryngologica Italica, 29*(4), 203.

Brancazio, L. (2004). Lexical influences in audiovisual speech perception. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(3), 445.

Burnham, D., & Dodd, B. (1996). Auditory-visual speech perception as a direct process: The McGurk effect in infants and across languages. *Speechreading by Humans and Machines* (103-114). Springer Berlin Heidelberg.

Burnham, D., & Dodd, B. (2004). Auditory–visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*(4), 204-220.

Campbell, R. (2008). The processing of audio-visual speech: Empirical and neural bases. *Philosophical Transactions of the Royal Society, 363*, 1001-1010.

Dautriche, I., Swingley, D., & Christophe, A. (2015). Learning novel phonological neighbors: Syntactic category matters. *Cognition, 143*, 77-86.

Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology*, *45*(4), 187-203.

Fennell, C. T., & Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech*, *46*(2-3), 245-264.

Fennell, C. T., & Werker, J. F. (2004). Infant attention to phonetic detail: Knowledge and familiarity effects. *In Proceedings of the 28th annual Boston University conference on language development* (Vol. 1, pp. 165-176). Somerville, MA: Cascadilla Press.

Fort, M., Kandel, S., Chipot, J., Savariaux, C., Granjon, L., & Spinelli, E. (2013). Seeing the initial articulatory gestures of a word triggers lexical access. *Language and Cognitive Processes, 28*(8), 1207-1223.

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*(5-6), 613-656.

Gibson, J. J. (1966). The senses considered as perceptual systems. Houghton Mifflin, Boston.

Gibson, J. J. (2014). The ecological approach to visual perception: classic edition. Psychology Press.

Havy, M., Foroud, A., Fais, L., & Werker, J. F. (2017). The Role of Auditory and Visual Speech in Word Learning at 18 Months and in Adulthood. *Child Development*.

Hollich, G., Newman, R. S., & Jusczyk, P. W. (2005). Infants' use of synchronized visual information to separate streams of speech. *Child Development, 76*(3), 598-613.

James, W. (1890) The principles of psychology (Vol. II). New York: Holt.

Kazanina, N., Bowers, J. S., & Idsardi, W. (2017). Phonemes: Lexical access and beyond. *Psychonomic Bulletin & Review*, 1-26.

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development*, *7*(3), 361-381.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental Science, 9*(2), F13-F21.

Lewkowicz, D. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology, 46*(1), 66-77.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception & Psychophysics*, *24*(3), 253-257.

Marslen-Wilson, W. (Ed.). (1992). Lexical Representation and Process. MIT Press.

Marslen-Wilson, W., Brown, C. M., & Tyler, L. K. (1988). Lexical representations in spoken language comprehension. *Language and Cognitive Processes, 3*(1), 1-16.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology, 21*(3), 398-421.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18*(1), 1-86.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746-748.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58*(3), 351-362.

Newman, R. S. (2008). The level of detail in infants' word learning. *Current Directions in Psychological Science, 17*(3), 229-232.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition, 52*(3), 189-234.

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*(2), 191-196.

Pitt, M. A., & Samuel, A. G. (2006). Word length and lexical activation: Longer is better. *Journal of Experimental Psychology: Human Perception and Performance, 32*(5), 1120.

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, *17*(6), 405-409.

Rosenblum, L. D., Schmuckler M.A., and Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics* 59(3), 347-357.

Sekiyama, K. (1997). Cultural and linguistic factors in audiovisual speech processing: The McGurk effect in Chinese subjects. *Perception & Psychophysics, 59*(1), 73-80.

Sekiyama, K., Soshi, T., & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: a heightened McGurk effect in older adults. *Frontiers in Psychology, 5*.

Stager, C.L. & Werker, C.F. (1997). Infants listen for more phonetic detail in speech perception than in word learning tasks. *Nature, 388*, 381-382.

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.

Stevenson, R. A., Zemtsov, R. K., & Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Journal of Experimental Psychology: Human Perception and Performance, 38*(6), 1517.

Swingley, D. (2003). Phonetic detail in the developing lexicon. *Language and Speech, 46*(2-3), 265-294.

Swingley, D. (2005). 11 – month – olds' knowledge of how familiar words sound. *Developmental Science, 8*(5), 432 – 443.

Swingley, D. & Aslin, R.N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition, 76*, 147 – 166.

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, *13*(5), 480-484.

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology, 54*(2), 99-132.

Swingley, D., Pinto, J. P., & Fernald, A. (1999). Continuous processing in word recognition at 24 months. *Cognition, 71*(2), 73-108.

Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, *108*(3), 850-855.

Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification*, *11*(4), 233-241.

Vaden, K.I., Hickok, G.S., & Halpin, H.R. (2009). Irvine Phonotactic Online Dictionary, Version 1.4. [Data file]. Available from http://www.iphod.com.

Vitevitch, M. S. (2003). The influence of sublexical and lexical representations on the processing of spoken words in English. *Clinical Linguistics & Phonetics, 17*(6), 487-499.

Vitevitch, M. S., Luce, P. A., Pisoni, D. B., & Auer, E. T. (1999). Phonotactics, neighborhood activation, and lexical access for spoken words. *Brain and Language, 68*(1), 306-311.

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science, 316*(5828), 1159-1159.

Werker, J. F., & Fennell, C. T. (2004). Listening to sounds versus listening to words: Early steps in word learning. In D. G. Hall & S. Waxman (Eds.), *Weaving a Lexicon* (79–109). Cambridge, MA: MIT Press.

Werker, J. F., & Tees, R. C. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development, 7*(1), 49-63.

Winneke, A. H., & Phillips, N. A. (2011). Does audiovisual speech offer a fountain of youth for old ears? An event-related brain potential study of age differences in audiovisual speech perception. *Psychology and* Aging*, 26*(2), 427.

Woollams, A. M. (2015). Lexical is as lexical does: computational approaches to lexical representation. *Language, Cognition and Neuroscience, 30*(4), 395-408.

Appendices

Table 1. Target objects, correct pronunciations, and mispronunciations.

Target Object	**Correct Pronunciation**	**Mispronunciation**
Baby	Baby	Gaby
Dog	Doggy	Boggy
Duck	Ducky	Bucky
Cat	Kitty	Pitty
Car	Car	Par
Fish	Fish	Shish
Shoe	Shoe	Foo
Ball	Ball	Gall

Table 2. Stimuli Combinations.

Modality	**Pronunciation**	**Example**
Audiovisual	Correctly Pronounced	**Audio:** *Doggy* **Visual:** *Doggy*
Audiovisual	Mispronounced	**Audio:** *Boggy* **Visual:** *Boggy*
Visual-Only	Correctly Pronounced	**Audio:** *Pink Noise* **Visual:** *Doggy*
Visual-Only	Mispronounced	**Audio:** *Pink Noise* **Visual:** *Boggy*

Table 3. Table of means for looking accuracy

	Proportion Correct Looking		First Look	
	Mean	SD	Mean	SD
Audiovisual CP	.9493	.121	1.00	0
Audiovisual MP	.8207	.173	.9414	.145
Visual-only CP	.9262	.124	.9717	.122
Visual-only MP	.6267	.203	.8193	.219

Table 4. Table of mean reaction times per trial type (in milliseconds).

	Reaction Time	
	Mean	Std. Error
Audiovisual CP	600.67	35.11
Audiovisual MP	728.25	37.65
Visual-only CP	680.16	36.55
Visual-only MP	1020.42	51.25

Table 5. Regression analysis within and between conditions on the audiovisual word recognition task and the IPEP.

	AV CP RT	AV MP RT	VO CP RT	VO MP RT	AV CP ACC	AV MP ACC	VO CP ACC	VO MP ACC	IPEP RT	IPEP Soc RT	IPEP NonSoc RT	IPEP ACC	IPEP Soc ACC
AV CP RT													
AV MP RT	.45**												
VO CP RT	.42**	.31											
VO MP RT	.32	.49**	.26										
AV CP ACC	-.01	.31	.27	.38*									
AV MP ACC	.03	.29	.11	.20	.40*								
VO CP ACC	.15	.20	-.06	.09	.23	.24							
VO MP ACC	-.32	.06	-.22	-.02	.11	.18	.28						
IPEP RT	-.12	-.04	.06	-.22	-.23	.17	-.12	-.04					
IPEP Soc RT	-.15	-.07	.16	-.09	-.21	.15	-.08	.10	.69**				
IPEP NonSoc RT	-.07	-.01	-.04	-.23	-.14	.12	-.11	-.12	.84**	.19			
IPEP ACC	.05	.34*	.30	.30	.44**	.09	.31	.10	-.27	-.12	-.25		
IPEP Soc ACC	.06	.32	.29	.28	.35*	.07	.22	.03	-.13	-.13	-.05	.92**	
IPEP NonSoc ACC	.01	.33*	.26	.27	.46**	.10	.34*	.17	-.32*	-.07	-.37*	.96**	.77**

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

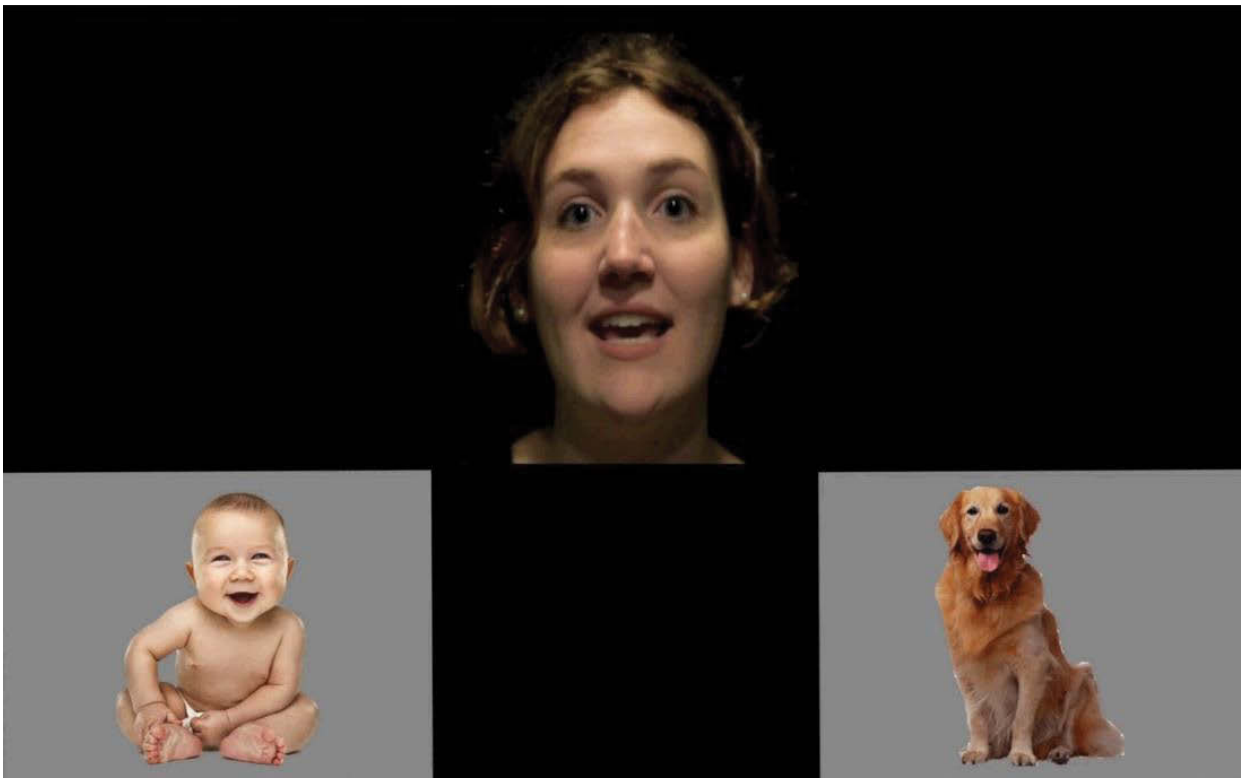

Figure 1. A screen shot of stimulus presentation during the audiovisual word recognition task

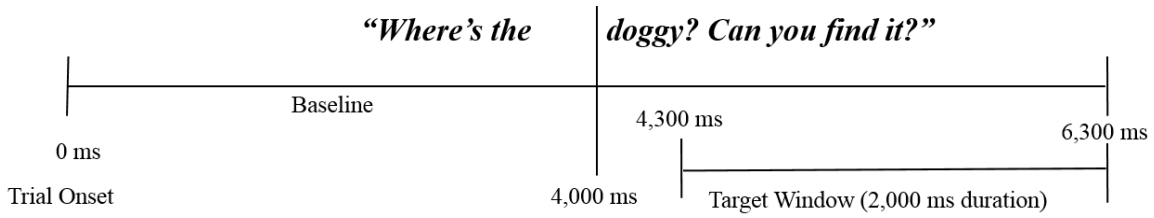

Figure 2. A timeline of the trial presentation in milliseconds, denoting baseline and target phases for eye gaze analyses.

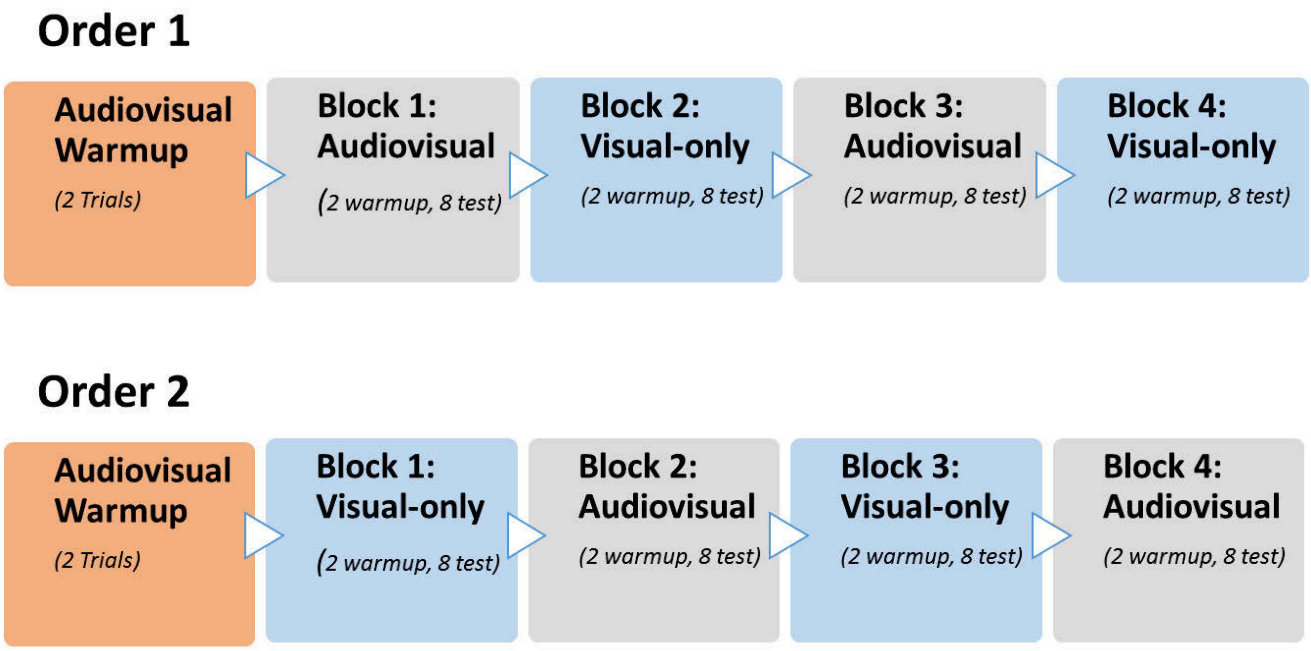

Figure 3. Trial layout for the audiovisual word recognition task for order 1 and order 2.

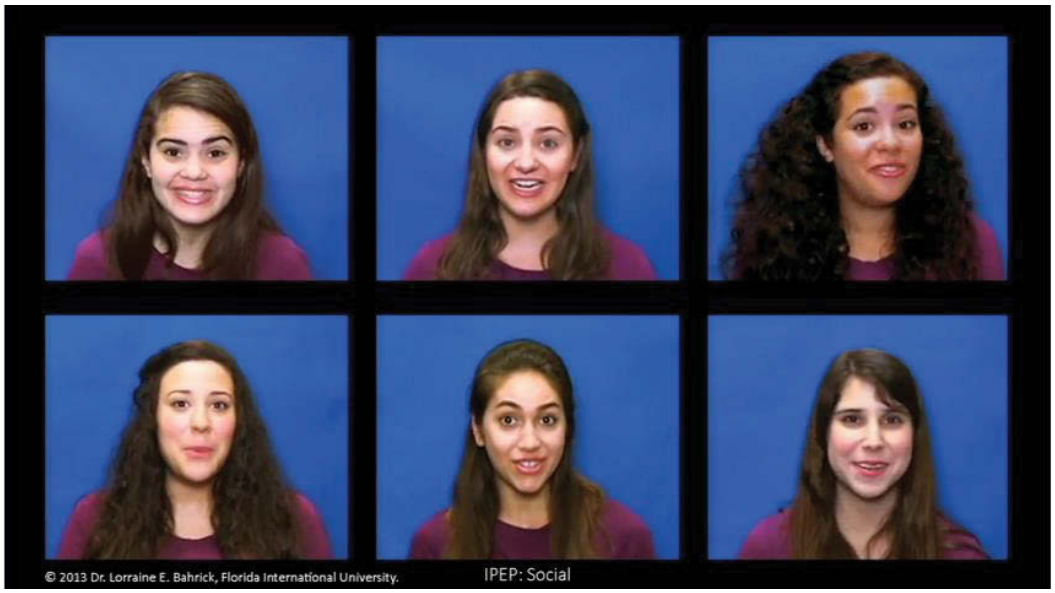

Figure 4. A screen shot of a social trial on the IPEP.

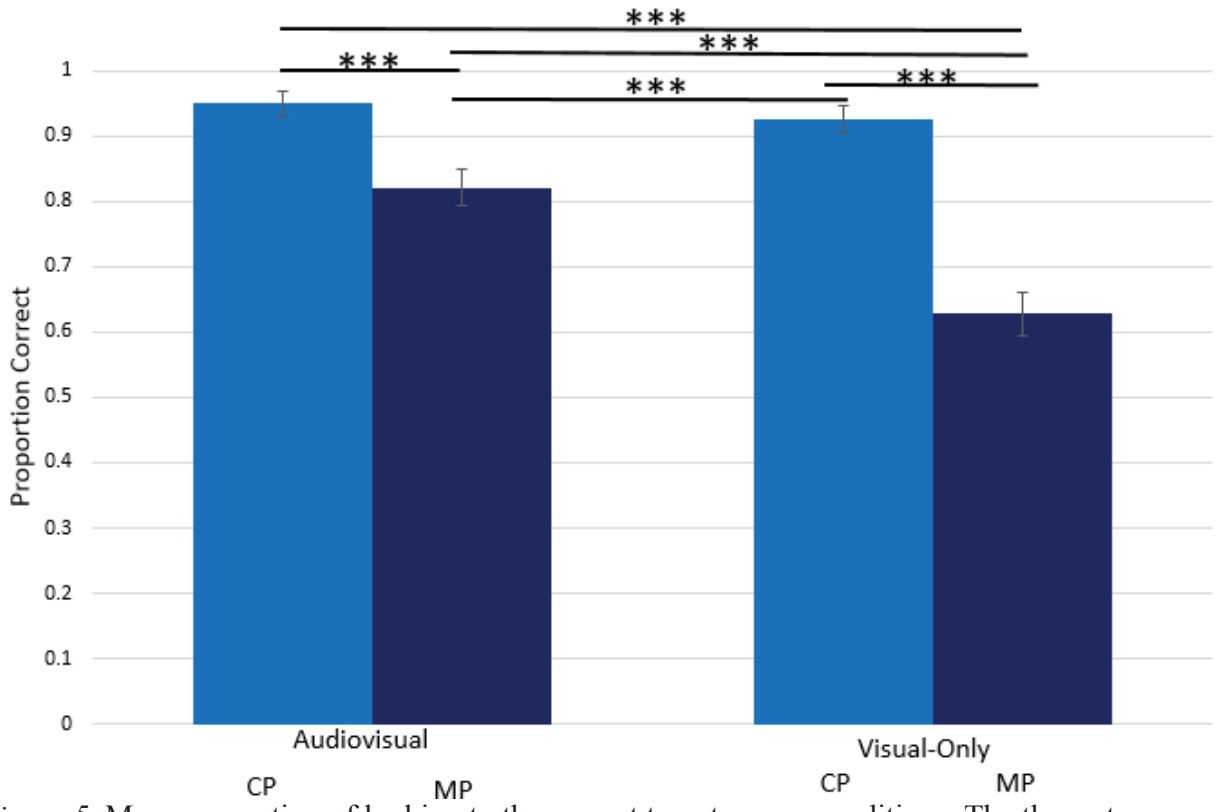

Figure 5. Mean proportion of looking to the correct target across conditions. The three stars (***) denote significant pairwise comparisons at $p < .001$.

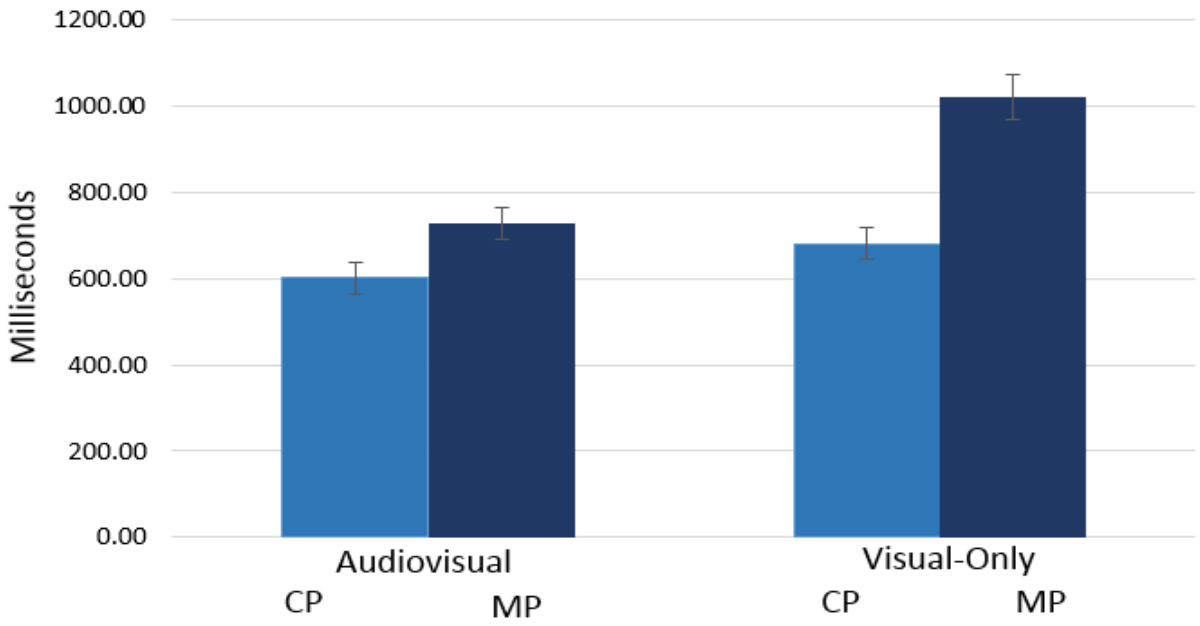

Figure 6. Reaction time to fixate on the correct target across conditions. All pairwise comparisons are significant ($p < .001$).

Vita

Ryan Andrew Cannistraci was born in Hershey, PA to Dan and Elizabeth Cannistraci. He is the second child of four siblings; Daniel, Stephen, and Lauren. He graduated from Cedar Crest High School in Lebanon, PA in 2010. He attended Millersville University and obtained a Bachelor of Arts degree in Psychology in 2014 under the mentorship of Dr. Shaun Cook and Dr. Shawn Gallagher. Following graduation, he accepted a graduate teaching assistantship at the University of Tennessee, Knoxville in the Experimental psychology program to work with Dr. Jessica Hay. Ryan is currently a fourth year graduate student and a graduate teaching associate. He will continue to pursue his Ph.D. in Experimental psychology at UTK under Dr. Hay's supervision. His research interests focus on the influence of audiovisual speech information for speech perception and early language learning using converging approaches—including computational modeling methods and neuroimaging techniques in addition to eye-tracking and behavioral measures.