VISUAL INFLUENCES ON SPEECH PERCEPTION IN INFANCY

by

Donald Kyle Danielson

B.A., Duke University, 2009

M.Sc., University of Alberta, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Psychology)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

August 2016

Abstract

The perception of speech involves the integration of both heard and seen signals. Increasing evidence indicates that even young infants are sensitive to the correspondence between these sensory signals, and adding visual information to the auditory speech signal can change infants' perception. Nonetheless, important questions remain regarding the nature of and limits to early audiovisual speech perception. In the first set of experiments in this thesis, I use a novel eyetracking method to investigate whether English-learning six-, nine-, and 11-month-olds detect content correspondence in auditory and visual information when perceiving non-native speech. Six- and nine-month-olds, prior to and in the midst of perceptual attunement, switch their face-scanning patterns in response to incongruent speech, evidence that infants at these ages detect audiovisual incongruence even in non-native speech. I then probe whether this familiarization, to congruent or incongruent speech, affects infants' perception such that auditory-only phonetic discrimination of the non-native sounds is changed. I find that familiarization to incongruent speech changes—but does not entirely disrupt—six-month-olds' auditory discrimination. Nine- and 11-month-olds, in the midst and at the end of perceptual attunement, do not discriminate the non-native sounds regardless of familiarization condition. In the second set of experiments, I test how temporal information and phonetic content information may both contribute to an infant's use of auditory and visual information in the perception of speech. I familiarize six-month-olds to audiovisual Hindi speech sounds in which the auditory and visual signals of the speech are incongruent in content and, in two conditions, are also temporally asynchronous. I hypothesize that, when presented with temporally synchronous, incongruent stimuli, infants rely on either the auditory or the visual information in the signal and use that information to categorize the speech event. Further, I predict that the addition of a temporal offset to this incongruent speech changes infants’ use of the auditory and visual information. Although the main results of this latter study are inconclusive, post-hoc analyses suggest that when visual information is presented first or synchronously with auditory information, as is the case in the environment, infants exhibit a moderate matching preference for auditory information at test.

Preface

This dissertation is my own original work, and I was the senior author on all collaborative projects (Chapters 2-3).

Chapter 1. I am the primary author of this chapter, with intellectual contributions and comments from Janet F. Werker, PhD (supervisor).

Chapter 2. I am the primary author of this chapter. The research questions and study design were decided in collaboration with Alison J. Bruderer, PhD, Padmapriya Kandhadai, PhD, Eric Vatikiotis-Bateson, PhD (supervisory committee member), and Janet F. Werker, PhD (supervisor), all of whom also provided intellectual contributions to the current report. I collected all data in collaboration with research assistants in my lab, and I conducted all analyses. This research is covered under UBC Ethics Certificate B95-0023/H95-80023 (UBC Behavioural Research Ethics Board). This chapter in its entirety has been submitted, but not yet accepted, for publication.

Chapter 3. I am the primary author of this chapter. The research questions and study design were decided in collaboration with Cassie Tam (BA student), Padmapriya Kandhadai, PhD, and Janet F. Werker, PhD (supervisor), all of whom also provided intellectual contributions to the current report. Cassie Tam and I collected all data. I conducted all analyses. This research is covered under UBC Ethics Certificate B95-0023/H95-80023 (UBC Behavioural Research Ethics Board).

Chapter 4. I am the primary author of this chapter, with intellectual contributions and comments from Janet F. Werker, PhD (supervisor).

Table of Contents

List of Figures

List of Symbols

ɑ	open back unrounded vowel
ɑ:	open back unrounded vowel (lengthened)
b	voiced bilabial stop
d̪	voiced dental stop
ɖ	voiced retroflex stop
d	voiced alveolar stop
g	voiced velar stop
η^{2}_{P}	partial eta-squared
l	alveolar lateral approximant
p	voiceless bilabial stop
ɹ	postalveolar approximant
v	voiced labiodental fricative

Acknowledgements

There are a great number of people without whose advice and guidance I could not have completed this project. First and foremost, I wish to thank my supervisor, Dr. Janet Werker, for her constant support, wisdom, and patience. Thank you for teaching me and believing in me, even when I did not. You have made me a better researcher, scholar, and teacher and—through your words and actions—have taught me how to be successful and maintain a healthy work/life balance at the same time. I am so grateful to have had you as a supervisor. I also especially thank Dr. Eric Vatikiotis-Bateson and Dr. Padmapriya Kandhadai, for their patient and thorough assistance with the drafts of the manuscripts comprising this dissertation. My heartfelt thanks go also to my collaborators, colleagues, and friends in the Infant Studies Centre, especially to Dr. Alison Bruderer, Cassie Tam, and Savannah Nijeboer, for their tireless contributions to the projects presented in this thesis. I gratefully acknowledge UBC and the Acoustical Society of America for their financial support of my degree program, and the National Institutes of Health (U.S.) for their grant to Janet Werker that allowed these projects and many others to materialize. Of course, when working with infants, no list of acknowledgements is complete without recognition of the families upon whose participation we are dependent: thank you for taking a day (and often more than one!) out of your precious parental leave time to come to UBC and participate in my studies.

On a personal note, I thank my friends, most especially Matt, Ian, Lizzy, Annie, Katy, and Josh. Your friendships have sustained me through the highs and lows of completing a Ph.D. program. Finally, I thank my partner and best friend Luke for his love, patience, and

kindness, and for taking a chance and moving to Vancouver so that I could embark on this journey. I am so very lucky to have you in my life.

Chapter 1: Introduction

1.1 Overview

The neonate is faced with a new world of objects and events, many of which simultaneously provide information to more than one of the child's developing sensory systems. These multisensory experiences, from the seen and felt touch of a parent's hand to the heard and seen whirr of hospital machinery to the heard, seen, and felt cries that the infant produces herself, envelop a young child in a complex sea of information from the first moments after birth. The complexity of the new environment far outpaces that of the infant's previous environment, the uterus, where the child's sensory systems had begun to develop and had begun to detect auditory, vibrotactile, and olfactory/gustatory information. Now, with eyes open and ears clearing of fluid, the infant begins the years-long process of making sense of this environment full of bright, loud, heavy, and smelly stimuli.

The infant does not begin this process empty-handed, however. Increasingly robust evidence indicates that the human infant is born prepared to learn (or even to expect) correspondences between multisensory signals when they are provided by a single source. Some of the most pervasive work on this type of multisensory processing in early infancy has been conducted in the domain of speech perception, where the signals provided in the auditory and visual modalities are tightly coupled in the timing of their onsets and offsets and the peaks and troughs in their respective amplitudes. Not only do the multimodal signals of speech correlate in these low-level, *modality-general* properties of their waveforms, but they also share more specific characteristics. The sight of a human mouth open in vocalization produces a specific type of sound (a vowel), one that differs dramatically from the sound produced when the mouth aperture is restricted by an oral articulator (e.g., the teeth, lips, tongue, palate), resulting in a consonant. And the specificity extends beyond this simple vowel-consonant dichotomy: the relative size of the mouth aperture is highly correlated with the acoustic properties of specific vowels, as the place and manner of vocal tract closure is associated with the acoustic properties of consonants. In fact, the correspondence between the modality-specific characteristics of auditory and visual speech is so tightly coupled that computational models can, supplied only with visual information from the speaking face, predict a great deal of the variance in the acoustic signal (Vatikiotis-Bateson & Munhall, 2015).

As adults with years of experience in our native language(s), we are sensitive to these visual-acoustic correspondences in speech, and we notice inconsistencies in audiovisual congruence. Most of us are familiar with the nuisance of watching poorly dubbed film; even the slightest of temporal asynchronies or structural incongruencies between the audio that we hear and the video that we see are noticeable to most adult perceivers of speech (although, after sustained exposure to asynchronous speech, adults become less sensitive to temporal inconsistencies (Navarra et al., 2005)). But, as more deeply reviewed later in this introductory chapter, adults' simultaneous use of auditory and visual information when processing speech is not limited to the rather artificial example of video recorded speech. Even normally hearing adults rely heavily on visual information to comprehend auditory speech events, and auditory speech perception is improved when paired with corresponding visual speech in both noisy and ideal listening conditions (Sumby & Pollack, 1954; Grant & Seitz, 2000; Remez, 2005; Ross et al., 2006). Moreover, adults' auditory speech perception can be changed by the imposition of a visual signal, a phenomenon most clearly demonstrated in the famous McGurk effect (McGurk & Macdonald, 1976).

That adults rely so heavily on both auditory and visual information when processing speech, and that they have such stringent expectations regarding the correspondences between the two, may not be surprising given the years of exposure to audiovisual speech to which the typically developing adult has been exposed. Although there are exceptions (talking on the phone, having one's back turned to one's speaking partner, listening to speech in the dark or watching speech from across a crowded, noisy room), speech perception is normally both auditory and visual. Adults, then, have been exposed to the specific pairings of auditory and visual speech since birth, and so their sensitivity to temporal asynchrony in speech (at thresholds hovering around 100 ms) (Dixon & Spitz, 1980; Munhall, Gribble, Sacco, & Ward, 1996; van Wassenhove, Grant, & Poeppel, 2007) and content incongruence (for example, English-speaking adults can spot the difference between a visual /d/ and a visual /b/ in the articulatory configuration of a speaker's face) could simply be driven by experience.

Strikingly, however, young, inexperienced infants also exhibit robust sensitivity to both temporal and structural correspondences in audiovisual speech. Infants detect the match between heard and seen speech events in their native language (Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999), doing so as early as two months (Patterson & Werker, 2003). Infants' auditory speech discrimination, like that of adults, is improved when they have access to the corresponding visual signal (Teinonen. Aslin, Alku, & Csibra, 2008; Ter Schure, Junge, & Boersma, 2016), and the imposition of an incongruent visual signal can change an infant's perception of an auditory speech event (Burnham & Dodd, 2004; Desjardins & Werker, 2004; Rosenblum, Schmuckler, & Johnson, 1997). But young infants are also capable of feats in audiovisual speech perception that adults are not. For approximately the first half of the first year of life, infants match the heard and seen speech signals of languages with which they are unfamiliar (Pons, Lewkowicz, Soto-Faraco, & Sebastian-Galles, 2009; Kubicek et al., 2014), and can discriminate between two different languages when shown visual-only displays (Weikum et al., 2007). During this same period, they can even match the heard and seen calls of non-human primates (Lewkowicz & Ghazanfar, 2006). Adults do not exhibit such sensitivity, discriminating visually between two languages only when at least one of them is familiar (Soto-Faraco et al., 2007), and matching heard and seen signals from an unfamiliar language only at chance levels (Pons et al., 2009).

These and other remarkable abilities of the young infant, reviewed in depth below, have reshaped researchers' understanding of the early months of speech perception development. Although speech perception has historically been considered an auditory endeavour with support from vision and other sensory systems, these and other recent advances in the field have led to the re-characterization of infants' speech perception as a multisensory process. Moreover, as our understanding of infants' audiovisual speech perception advances and we discover the surprising breadth of infants' early perceptual capabilities, we have come to understand better the role that experience with a specific language may (or may not) play in the formation of the early speech percept.

Nonetheless, our understanding of the development of infants' audiovisual speech perception is itself continuing to develop rapidly. As the evidence accumulates that speech perception relies on both auditory and visual information from early in development, important questions arise around our previously held beliefs about speech perception, many of which were formed after extensive testing in the auditory domain alone. Moreover, as we come to better understand the complexity and specificity of infants' abilities to match heard and seen speech and to use visual speech to improve auditory perception, we find ourselves asking *how* it is that infants—with so little experience—develop such a rich connection between the heard and seen signals of spoken language.

In this thesis, I report the results of two sets of experiments, each addressing a set of interrelated hypotheses. In the first experimental set, I probe whether our understanding of a relatively well-understood phenomenon of early language development, perceptual attunement, is modified when we consider carefully the audiovisual nature of speech perception in infancy. I first test whether infants at three ages, before, during, and after perceptual attunement, are sensitive to incongruence in the auditory and visual signals of non-native speech, and predict that such a sensitivity, if shown, weakens developmentally throughout the first year of life. In that same experimental set, I then probe whether familiarizing infants to audiovisual speech boosts their discrimination of non-native auditory contrasts, predicting that visual information improves phonetic discrimination for nine-month-old infants undergoing perceptual attunement. In my second experimental set, I probe infants' integration of auditory and visual signals in speech more deeply, testing whether—when the two provide conflicting information—infants rely more heavily on information from one of these modalities over the other. Finally, I test the interaction between infants' use of content and temporal information in the audiovisual signal of speech, predicting that the addition of a temporal offset between the auditory and visual signals shifts infants' use of the auditory or visual information to make phonetic distinctions.

1.2 Multisensory perception

The environment is rich with objects and events that simultaneously produce auditory and visual signals, among others. On a rainy morning in the city, I see the texture and the size of the raindrops as they fall. I hear them as they fall on me and on the surrounding surfaces of the street, the trees, and the vehicles. I use different sensory organs to detect these auditory and visual signals, and each has its unique, modality-specific characteristics. Without taking some poetic license, one cannot easily describe the colour of an auditory event or the pitch of a visual one (unless one has synesthesia (Spector & Maurer, 2009)). The transparency of a raindrop has no auditory equivalent, and the sound of its contact with the ground cannot be described visually. But because these signals are produced at the same source, they also share modality-general properties, such as their temporal onsets and offsets, their rhythmicity, their intensity, and their spatial location, among others (Lewkowicz & Turkewitz, 1980). Temporally, I detect the heard and seen signals of the rain simultaneously. The rhythm of the falling rain is reflected in both sensory modalities, and the intensities of the auditory and visual signals are highly correlated in a particular time window.

Humans are sensitive to these modality-general correspondences between seen and heard stimuli, and—perhaps because of these highly reliable correspondences—the integration of signals from multiple modalities into one unified percept is argued to be a "fundamental characteristic of the brain" (Stein, 1998, p. 124). Humans match auditory and visual signals produced at the same source based on their rhythms, intensities, spatial locations, *et cetera*, and when simultaneously provided with redundant sensory information from more than one modality, processing of a stimulus is faster and more accurate than when information is presented in one modality alone. For example, when instructed to detect a target visual event in a series of distractor events, adults' performance is improved when the target event is paired with a high-pitch tone embedded in a background of low-pitch auditory stimuli presented with the distractor events (Vroomen & de Gelder, 2000). Likewise, when subjects are instructed to locate a visual signal to the right or left of a centrally-located fixation point, their reaction time is reduced (improved) when the visual stimulus is accompanied by an auditory event presented on the same side (Posner, Nissen, & Klein, 1976). Learning is also improved when stimuli are presented multimodally rather than unimodally. Adult subjects better remember visual pictures when they are presented concurrently with sound pairings than when they are presented in silence (Murray et al., 2004). Even infants learn to discriminate visual-only rhythms better after familiarization with synchronous, multisensory (audiovisual) familiarization than after unisensory familiarization alone (Bahrick & Lickliter, 2000).

However, when cross-modal signals are non-redundant (i.e., when they provide conflicting information), typically because of an experimental manipulation, processing is slower and less accurate (Kim, Seitz, & Shams, 2008). Conflicting, incongruent cross-modal information can also *change* the perception of an event. When auditory tones are paired with visual light of differing durations, the length of the auditory tone affects the perceived duration of the visual light (Walker & Scott, 1981). Similarly, when blinking visual lights and intermittent auditory tones or clicks are presented simultaneously but at different rates, the rate of presentation of the auditory stimulus influences the perception of the rate of light blinks (Gebhard & Mowbray, 1959; Shipley, 1964; Welch, DuttonHurt, & Warren, 1986). Cross-modal conflicts can change the spatial localization of an event as well, a phenomenon exemplified by the *ventriloquist effect* (Howard & Templeton, 1966). The pairing of a synchronous visual event (in this case, a puppet's mouth moving) with an auditory event (a ventriloquist speaking) induces a change in source localization, with observers localizing the speech to the puppet instead of to the performer. This effect has been demonstrated outside its specific domain of origin. When subjects are presented with an auditory beep and a temporally synchronous but spatially displaced flashing light, their localization of the beep is displaced to be closer to the location of the light (Bertelson & Aschersleben, 1998).

The human adult is able to identify the sensations detected by the various sensory organs, and can use language to discuss the modality-specific aspects of multisensory events. Perhaps unsurprisingly, then, human perceptual function was historically characterized as a set of modular systems operating independently of one another (Shimojo & Shams, 2001). However, the prevalent and robust facilitatory, disruptive, and illusory effects of cross-modal congruence and incongruence briefly reviewed here have led to the re-evaluation of the perceptual systems, such that they are now seen as multisensory in nature. Perception of auditory events are altered by the imposition of visual information, whether that information is congruent or incongruent, and the integration of visual and auditory signals appears to be automatic and obligatory.

1.3 The case of speech

Even the most cursory read of the vast literature on multisensory perception in humans will include a great deal of reference to one domain in which information is presented across multiple modalities: speech. Although speech perception, as with other perceptual domains, has historically been considered a unisensory process, data from the last century has revised this view considerably (see Campbell, 2008, for a detailed historical review). The perception of human speech is now considered to be robustly multisensory, as rich acoustic, visual, and vibrotactile information is produced by the vocal tract and is transmitted to the perceiver's corresponding sensory receptors. As noted above, although there are instances in which the signal of only one speech modality is detectable, the vast majority of speech perception involves the detection and processing of dynamic information from more than one modality. When watching and listening to another person speak, the modality-general properties of the visible signals from the speaker's moving articulators (e.g., rhythm, duration, intensity, temporal onsets and offsets) correlate highly with these same modality-general properties of the acoustic signal, and together they provide tightly coupled information to the perceiver. Subjects are quite sensitive to the temporal synchrony of speech events, noticing asynchrony even when auditory information precedes visual information by as little as 60-130 ms (Munhall, Gribble, Sacco, & Ward, 1996; van Wassenhove, Grant, & Poeppel, 2007; Dixon & Spitz, 1980).

In addition to the modality-general properties that the auditory and visual signals of speech share (e.g., rhythm, intensity, temporal onsets and offsets), there also exists a tight relationship between the *modality-specific* signals of speech. Certain visual articulatory gestures correspond to specific acoustic realizations, and *vice versa*, and these correspondences are tightly coupled. At a gross level, the sight of the mouth opening, wherein the vocal articulators (e.g., the tongue, lips, teeth, palate) minimally constrict the opening of the vocal tract, is associated with the acoustic structure of vowels, comprised of resonant frequency bands. The sight of the mouth closing, wherein one or more of the aforementioned articulators restricts the movement of air in the vocal tract, is associated with the rather different and varied acoustic properties of consonants. But, as noted above, audiovisual links in speech events are more specific than this simple consonant-vowel dichotomy. For example, the extent of mouth aperture affects the formants of vowels (the ranges of frequencies with high spectral peaks that give vowels their individual characteristics). Close vowels such as /i/ have a lower first formant than do open vowels such as /ɑ/. Likewise, the visual attributes accompanying specific vocal closures associated with consonants correlate highly with those consonants' acoustic characteristics. Not only does the voiced bilabial stop /b/ differ from the voiced velar stop /g/ in its visual characteristics (with lips closing in a canonical production of /b/ and tongue body meeting velum in canonical productions of /g/), but the acoustic characteristics of the two sounds covary with the changes in visual information. As noted above, even an algorithm without linguistic experience can predict a great deal of the variance in the acoustic signal of speech when provided with the visual signal (see Vatikiotis-Bateson & Munhall, 2015, for a review).

As with non-speech events, there are strong cross-modal interactions in the perception of the visual and auditory signals of speech. In one of the most cited examples of this interaction, Sumby and Pollack (1954) demonstrated that adults' auditory perception of speech in noise is highly improved by the imposition of a matching visual signal. The addition of visual information allows adults to perceive speech in a certain level of noise (4-6 dB) with accuracy akin to their perception of speech in silence (MacLeod & Summerfield, 1987). Later investigations into this phenomenon revealed that visual information not only improves auditory speech perception when conditions are noisy, but also under ideal listening conditions in which the signal-to-noise ratio is high (Remez, 2005). Moreover, even hearing adults are adept speechreaders, and are able, to varying degrees, to interpret silent visual speech (Summerfield, 1992), even when fixating to regions of the speaker's head and face other than the mouth (Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). Visual speech information also aids in deciphering auditory speech that is less familiar, such as that produced in a second language. For example, Spanish-dominant Spanish-Catalan bilinguals discriminate a Catalan auditory contrast when auditory tokens are paired with visual information, but not when they are presented without a visual display (Navarra & Soto-Faraco, 2005).

In day-to-day speech interactions, complementary information provided by the auditory and visual signals of speech aids in its comprehension. However, one of the most striking demonstrations of the interaction between auditory and visual information in the speech percept comes from the unnatural imposition of *conflicting* auditory and visual speech information. As with non-speech stimuli, the artificial pairing of incongruent auditory and visual signals results in slower processing, and may also—depending on the modality-specific characteristics of the two signals—result in an illusory percept. The best known of these illusory percepts was discovered by McGurk and Macdonald (1976), and is known eponymously as the McGurk effect. In the classic demonstration of the McGurk effect, an auditory-only token of the consonant-vowel (CV) syllable /bɑ/ is paired with a silent visual token of the CV syllable /gɑ/. In response, English-speaking adults report perceiving the illusory, intermediate syllable /dɑ/. This effect has been repeatedly replicated (quite recently by van Wassenhove, Grant, & Poeppel, 2007; Keil, Muller, Ihssen, & Weisz, 2012; Setti, Burke, Kenny, & Newell, 2013), and appears to be robust to certain attempts to mitigate it. Even when adults are instructed to ignore the visual information in a McGurk-like paradigm, their auditory perception is nonetheless influenced by the visual signal (Buchan & Munhall, 2011). Crucially, perception of a McGurk-like illusion breaks down when the auditory and visual signals are not presented temporally simultaneously (within about 60 ms of one another) (Munhall et al., 1996), further indicating that both content information (in this case, the acoustic and visual characteristics of the visual /gɑ/ and auditory /bɑ/ syllables) as well as modality-general information (in this case, temporal synchrony) contribute to the adult's perception of audiovisual speech.

The evidence reviewed here indicates that, at least for the adult, speech perception is multisensory, with vision and audition (among other sensory modalities, including haptic perception) obligatorily implicated in the process (Rosenblum, 2005). As Ghazanfar and Takahashi eloquently argue in their review of the evolution of speech, "audiovisual (or 'multisensory') speech is really the primary mode of speech perception and not a capacity that was simply piggy-backed onto auditory speech perception later in the course of our evolution" (2014, p. 545). Although speech perception has historically been investigated using methods that probe auditory-only speech perception, that adults rely heavily on *both* audition and vision when perceiving speech should not be a surprise to the lay reader. After all, the typically developing adult has been exposed to audiovisual speech since birth. To borrow the title of McGurk and Macdonald's landmark (1976) study, adults have had decades of experience "hearing lips and seeing voices". Surely these years of exposure to language would prepare the adult to have stringent expectations not only about the shared modality-general characteristics of speech, common to all multisensory objects and events, but also to the specific acoustic properties of visual phones and the visual properties of auditory ones.

However, it is not in adulthood, or even in later childhood, that evidence of the audiovisual nature of speech perception is first observable. From the earliest ages tested, young infants also exhibit robust audiovisual speech perception. Such findings indicate that the audiovisual nature of speech perception is not solely a learned tendency, but that speech perception is audiovisual from birth. The remainder of this literature review explores the recent advances that have been made in better understanding how infants use both visual and auditory information in the processing of speech, and explores how—as a result of these advances—the pervasive view in early speech perception research has shifted from an auditory-dominant one to one that considers speech perception as a multisensory process from early in ontogeny.

1.4 Audiovisual speech perception in infancy

At birth, infants appear prepared to detect the correspondences between seen and heard speech, attending preferentially to faces over non-face stimuli (Johnson, Dziurawiec, Ellis, & Morton, 1991; Goren, Sarty, & Wu, 1975; Valenza, Simion, Cassia, & Umilta, 1996), and to speaking faces over still faces (Nagy, 2008). From the first moments of post-natal life, infants are thus exposed to audiovisual speech from their caregivers, preparing them to rely on such information for the acquisition of complex language throughout the early years of the lifespan. In the sections that follow, I review the pertinent literature on the development of audiovisual speech perception across the first year of life, and make the case that sensitivity to the correspondences between auditory and visual speech are present without specific experience with seen and heard language in the environment.

1.4.1 Audiovisual matching

Perhaps the best known demonstration of infants' early sensitivity to the correspondence between auditory and visual speech is their ability to match heard and seen vowels (Kuhl & Meltzoff, 1982, 1984; MacKain, Studdert-Kennedy, Spieker, & Stern, 1983; Patterson & Werker, 1999, 2003). In their groundbreaking study, Kuhl and Meltzoff (1982) tested 4.5-month-old infants by showing them side-by-side videos of faces silently articulating the English vowels /ɑ/ and /i/. They simultaneously presented the infants with a stream of auditory tokens consisting of either /ɑ/ or /i/, projected from the centre of the experimental apparatus (between the two side-by-side faces). Infants exhibited longer looking time to the face articulating the matching vowel than they did to the face articulating the incongruent vowel. This result was interpreted as evidence that infants are sensitive to the correspondences between the specific auditory and visual signals of speech, and that—unlike previously thought—such a sensitivity emerges very early in ontogeny when infants are relatively inexperienced with speech. Infants' matching of heard and seen vowels precedes even their matching of a speaker's voice to a face of his or her gender, which does not emerge until approximately eight months of age (Patterson & Werker, 2002).

Nearly two decades after the initial finding exhibiting infants' phonetic matching abilities was published, it was replicated with two-month-old infants (Patterson & Werker, 2003), who also exhibited robust sensitivity to the match between heard and seen speech sounds in their native language. These authors argued that, based on their findings with such young infants, the intermodal matching of the auditory and visual signals of speech was either the object of extremely rapid experience-based acquisition, or was somehow privileged by underlying perceptual biases present at birth.

Following the publication of these studies, additional evidence has been accumulated indicating that young infants not only match heard and seen speech in their native language, but also do so in languages with which they are unfamiliar. Such results are important because they indicate that the intermodal perception of speech is not restricted to sound-sight combinations that have been highlighted in the infant's environment. To construct lexical items (words), each of the world's languages uses a subset of the possible sounds (phones) that the human vocal tract is capable of producing. The sounds that each language uses contrastively are called *phonemes*. For example, in English, the alveolar approximant /ɹ/ (as in 'rag') and the alveolar *lateral* approximant /l/ (as in 'lag') are both used to contrast meaning; they are both phonemic. The production of these two sounds by English speakers follows a bimodal distribution, with the greatest acoustic distance between the two sounds observed in the frequency of the third formant, the values of which do not overlap in canonical English productions (Lotto, Satto, & Diehl, 2004). English-learning infants are therefore exposed to a bimodal distribution of /ɹ/-like and /l/-like audiovisual speech tokens in the environment (for example, whenever a caregiver wishes to produce the words 'rag' or 'lag').

However, in Japanese, for example, such a distinction does not exist; the language does not utilize an alveolar or postalveolar approximant phonemically. While Japanese-learning infants may hear and see /ɹ/-like and /l/-like speech tokens, they are much more rare than they are in English, and are likely distributed widely throughout the speech that the infant perceives. This unsystematic distribution of non-phonemic speech sounds likely results in reduced discrimination of such sounds by older infants and adults (see Maye & Gerken, 2000; Maye, Werker, & Gerken, 2002 for evidence that discrimination is enhanced following presentation of a bimodal, and reduced following presentation of a unimodal auditory distribution of phones). Japanese-speaking adults and Japanese-learning infants, after about 9-10 months of age, no longer exhibit discrimination of the English /ɹ/-/l/ distinction (Goto, 1971; Miyawaki et al. 1975; Best & Strange, 1992; Tsushima et al., 1994).

Nevertheless, it appears as though—at least early in infancy—humans are capable of matching the heard and seen signals even of languages with which they are unfamiliar and which use sound contrasts that are irrelevant in the infant's native language. For example, Japanese-learning 8-month-old infants match the seen and heard signals of consonants, even when the consonant is a non-canonical bilabial trill /B/ that is not present in their language's repertoire of speech sounds (Mugitani, Kobayashi, & Hiraki, 2008). Similarly, Spanish-learning infants match the heard and seen signals of two English speech sounds (the voiced labiodental fricative /v/ and the voiced bilabial stop /b/), even though Spanish does not use these sounds contrastively (Pons et al., 2009). German-learning infants, after familiarization to auditory-only fluent speech from either French or German, look preferentially to one of two side-by-side silent visual displays producing the language that matched what they heard during familiarization (Kubicek et al., 2014). The results of these studies probing infants' detection of audiovisual correspondence in non-native languages provide evidence for the early role of visual information in the perception of speech, even when specific experience is absent. These results advance the proposal that there may be a privileged mapping between auditory speech perception and visual speech perception from birth, and that experience is not required for infants to match the seen and heard content of audiovisual speech.

There are two additional patterns in these latter two studies that merit attention for the purposes of the present endeavour. First, both Pons and colleagues (2009) and Kubicek and colleagues (2014) examined infants' sensitivity to the match between heard and seen speech developmentally across the first year of life. As reviewed above, during the first half of this period, infants are sensitive to the auditory distinctions between non-native speech sounds, a sensitivity that declines later in development. Both of these studies (Pons et al., 2009; Kubicek et al., 2014) replicated this finding, but with audiovisual matching. Infants were sensitive to the match between non-native heard and seen speech, but only until a certain point in development after which they cease to discriminate between auditory-only exemplars from these languages. Pons and colleagues (2009) argue, in light of their studies, that the sensitive period for audiovisual matching in a non-native language is similar to the sensitive period for auditory-only contrasts and that, perhaps, the sensitive period is "pan-sensory" (p. 10598).

Moreover, in both the Pons and colleagues (2009) and Kubicek and colleagues (2014) study, infants matched heard and seen speech in the absence of modality-general temporal cues that might have bootstrapped infants' matching of the signals from the two modalities. Unlike in prior studies probing audiovisual speech matching (e.g., Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999, 2003), in these two recent studies, infants were presented with information from the two modalities sequentially rather than simultaneously. First infants were familiarized to auditory stimuli, and then they were probed for preference to one of two *silent* visual displays. Although in all audiovisual matching procedures attempts are made to control for infants' possible use of certain modality-general cues (e.g., auditory stimuli are presented from a centrally located speaker to prevent infants from using spatial localization as a cue for matching), not all prior studies have controlled for the possibility that infants are simply matching heard and seen speech based on the temporal simultaneity of the onsets and offsets in the two modalities. Both Pons and colleagues (2009) and Kubicek and colleagues (2014) did control for such a possibility, and their nonetheless positive results are strong evidence that infants are capable of relying on more than just modality-general cues (e.g., temporal simultaneity) to match heard and seen speech. As such, it appears as though even young, inexperienced infants are sensitive to the correspondences between the modality-specific information conveyed by auditory and visual speech, matching information from the two modalities based on the content information therein and not just their temporal co-incidence.

Throughout the thesis, I use the term ‘phonetic content information’ to refer to the aspects of the acoustic and visual signals of speech that are encoded in one modality but that have reliable correspondences in the other modality (e.g., the position and extent of mouth aperture in the visual signal is highly correlated with the frequency of the first and second formants in the acoustic signal). I use the term ‘temporal information’ to refer to the simultaneity of onsets and offsets in the acoustic and visual signals, which is typical of natural speech and which may be used as a low-level cue by both infants and adults to match seen and heard speech. However, it is important to note that—in natural speech—the temporal and phonetic qualities of the speech signal are inextricable from one another. The phonetic content of the audiovisual speech signal (e.g., vowel formant heights, frequency of frication, mouth aperture, lip rounding) can be *described* in static terms, but the speech signal nonetheless unfolds across time. The resulting modulations in the acoustic and visual signals are strictly temporally arranged, and if this arrangement is artificially disrupted, speech comprehension may be reduced or entirely negated. The phonetic content of audiovisual speech thus, to some degree, includes temporal information.

1.4.2 The McGurk effect and visual capture in infancy

Additional evidence highlighting the audiovisual nature of infants' speech perception comes from studies that have attempted to replicate the McGurk effect (McGurk & Macdonald, 1976) with infants. To review, the McGurk effect occurs when certain visual phones (in the classic case, the English voiced velar stop /g/) are artificially paired with certain mismatching auditory phones (e.g., the English voiced bilabial stop /b/). Adult English speakers, when attending to the visual stimulus, perceive an illusory third percept, typically the voiced alveolar stop /d/ (or, if frication is present in the auditory production of /b/, the illusory percept is the voiced dental fricative /ð/). This obligatory percept, present even when the adult is instructed to ignore visual information (Buchan & Munhall, 2011) has been taken as evidence that the adult speech percept involves the integration of information from both the auditory and the visual modalities. Burnham and Dodd (2004) attempted to replicate the McGurk phenomenon with young, inexperienced infants, in order to determine whether their auditory perception, too, would be affected by mismatching visual information, or whether such an effect only occurs after years of linguistic experience. Such a question was important, as the original McGurk and Macdonald (1976) study showed that, while young children (at 3-5 and 7-8 years of age) exhibited the McGurk effect, they did so with less consistency than did the adults tested in the same procedure.

Burnham and Dodd (2004) habituated 4.5-month-old infants to McGurk-like stimuli (visual /gɑ/ and auditory /bɑ/), and tested them using three types of dishabituation stimuli: /bɑ/, /dɑ/, and /ðɑ/. Infants that were exposed to McGurk-like stimuli during habituation dishabituated to /bɑ/, but not to /dɑ/ or /ðɑ/, indicating that, during habituation, they too perceived an illusory intermediate percept. These results indicate that, at least with the sound-sight pairings tested, infants integrate the heard and seen signals of speech into one percept when the two are mismatched. A similar study habituated infants to audiovisual tokens of /vɑ/, and then tested for dishabituation with an auditory /bɑ/-visual /vɑ/ combination (Rosenblum, Schmuckler, & Johnson, 1997). When adults are exposed to this latter combination, they exhibit *visual capture*, wherein the percept matches the syllable that is produced visually (/vɑ/) (Rosenblum & Saldaña, 1992, 1996). Infants too appeared to exhibit visual capture, failing to dishabituate to this auditory /bɑ/-visual /vɑ/ combination (thus indicating that, due to visual capture, they continued to perceive /vɑ/, as adults do and as the infants themselves had during habituation with matched auditory /vɑ/-visual /vɑ/ stimuli). Infants did, however, dishabituate to an auditory /dɑ/-visual /vɑ/ combination, for which adults do not exhibit visual capture. Dishabituation to this stimulus indicated that, like that of adults, infants' auditory perception of /dɑ/ is not systematically modified by the imposition of a visual /vɑ/. Crucially, in each of the studies outlined here (as in most studies probing the McGurk effect), the auditory and visual stimuli were presented simultaneously. As noted, when the simultaneity of the auditory and visual signals is disrupted by more than about 60 ms, the McGurk effect breaks down (Munhall, Gribble, Sacco, & Ward, 1996). However, the temporal simultaneity maintained in these infant-directed McGurk-like studies is further evidence that infants rely on more than modality-general temporal information to decipher audiovisual speech. If infants' integration of heard and seen speech were based only on modality-general cues such as temporal simultaneity of speech onsets and offsets, their auditory speech perception should not be affected by the imposition of an incongruent, but temporally synchronous, visual signal. Rather, it appears as though infants are attentive to the modality-specific information inherent in the auditory and visual signals, in addition to their attention to low-level modality-general cues.

Importantly, although the effect of visual capture and the similar McGurk effect appear to be present in infancy, they may be weaker than they are in adulthood. In another study testing visual capture by the visual phone /v/, but with a different vowel (/i/), only male infants appeared to be systematically affected by the imposition of the mismatching visual signal, while female infants did not exhibit evidence of visual capture (Desjardins & Werker, 2004). Taken together, the results of these studies and those probing the McGurk effect in older children (e.g., McGurk & Macdonald, 1976; Dupont, Aubin, & Ménard, 2005; Tremblay et al., 2007; Sekiyama & Burnham, 2008) seem to indicate that while the visual influence on auditory speech perception is present early in infancy, its strength increases developmentally.

1.4.3 Visual facilitation of auditory speech perception

As our understanding of infants' speech perception has evolved from an auditory-centric one to one that considers the role of visual information as a contributor to the early speech percept, new questions have arisen that probe the *importance* of this contributor developmentally as infants acquire perceptual proficiency with the phonemic system of their native language(s). Studies probing infant matching of audiovisual stimuli have demonstrated conclusively that infants are capable of matching auditory and visual speech, and that they appear able to do so based on the rather specific correspondences between the modality-specific information of both audition and vision. Moreover, studies probing the McGurk effect have demonstrated that the imposition of incongruent visual information can change or disrupt the auditory percept. Recently, a few studies have probed whether the imposition of *congruent* visual information onto the auditory speech signal can in fact improve perception of speech sounds. A positive result to this question would indicate not only that infants are sensitive to the correspondences between auditory and visual speech, but also that they may *rely* on visual information during language acquisition. As has been noted earlier in this chapter, infants' learning of non-speech stimuli is improved when presented with multimodal rather than unimodal exemplars (Bahrick & Lickliter, 2000), and adults' speech perception is improved with the addition of a congruent visual signal (Sumby & Pollack, 1954; Ross et al., 2006; Summerfield, 1987; Remez, 2005).

Teinonen and colleages (2008) were the first to test whether the addition of visual information to the auditory speech signal would improve discrimination of the latter. To do so, they used synthetically degraded stimuli corresponding to English /bɑ/ and /dɑ/. The voiced stops /b/ and /d/ are phonemic in English, and in natural English /b/ and /d/ productions are distributed bimodally. That is, productions of /b/ and /d/ vary in natural speech, but cluster around canonical /b/ and /d/ sounds, respectively, which are acoustically dissimilar. Relatively fewer productions of either sound occur in the middle of this distribution (that is, if /b/ and /d/ were produced similarly, there would be more confusion between the words 'bark' and 'dark', among others). Infants are adept at differentiating speech sounds when they have been exposed to the stimuli in a bimodal distribution, which mimics the distribution of phonemically distinct speech sounds in the environment, but perform less well when exposed to the stimuli in a unimodal distribution (Maye, Weiss, & Aslin, 2008). Teinonen and colleagues (2008) familiarized English-learning infants to a unimodal distribution of synthetic utterances of /bɑ/ and /dɑ/, but split their subjects into two different visual conditions. In one condition, all auditory utterances of /bɑ/ and /dɑ/ were paired with one type of visual display, a face producing *either* /bɑ/ or /dɑ/. In the other condition, items from one side of the auditory /bɑ/-/dɑ/ continuum (e.g., the sounds closer to /bɑ/) were paired with a matching visual display (e.g., a visual face producing /bɑ/), while the other auditory items were paired with the opposite visual articulation. At test, infants in the two-category familiarization condition better discriminated auditory-only exemplars of the synthetic /bɑ/ and /dɑ/ stimuli than did infants in the one-category familiarization condition. These results provided the first evidence that visual articulatory information may boost infants' auditory discrimination of native language sounds.

The results presented by Teinonen and colleagues were further extended by a recent study in which the visual effects on auditory discrimination were tested with a non-native language (Ter Schure, Junge, & Boersma, 2016). They trained Dutch-learning infants with unimodal and bimodal distributions of an English-only vowel contrast (/ɛ/ versus /æ/) that is not phonemic in Dutch. Two groups of infants were trained with an auditory-only distribution (one group with a unimodal distribution and one with a bimodal distribution). Two more groups were trained with visual-only distributions, and two with audiovisual distributions. Then infants were tested using a habituation-dishabituation procedure in which discrimination was inferred by longer looking time to 'switch' trials (in which the dishabituation stimulus belonged to a different category than the habituation stimuli) than to 'same' trials (in which the dishabituation and habituation stimuli were the same). Although the effects of distribution and modality did not emerge as significant predictors of infants' overall discrimination at test, follow-up by-condition analyses revealed that only infants who were trained with distributions that were bimodal *and* audiovisual succeeded in exhibiting discrimination at test. These results preliminarily extend those of Teinonen and colleagues (2008), indicating that infants' use of visual information not only aids in auditory discrimination in their own language, but also in languages with which they are unfamiliar.

1.5 Current studies: Shifting our understanding of auditory speech perception development

Although speech perception has been historically conceptualized as an auditory-dominant endeavour (Campbell, 2008), the past few decades of results have shifted our understanding of the speech percept to one that takes into account the important role of visual information in the perception of speech. While the pioneering work of Sumby and Pollack (1954) began to illuminate the audiovisual nature of adult speech perception, it was assumed for many decades that such visual influences on auditory speech perception were the result of learning from linguistic input in the environment. Since the publication of Kuhl and Meltzoff's (1982) study probing the intersensory matching of heard and seen speech, however, this viewpoint has been revised. Like adults, infants are sensitive to the match between heard and seen speech, and they are so from a remarkably young age. Other studies probing infants' intersensory matching in speech have revealed that they are capable of this feat at two months (Patterson & Werker, 2003). Moreover, that infants match heard and seen speech even in non-native languages (e.g., Pons et al., 2009; Kubicek et al., 2014) indicates that infants' detection of the correspondences between seen and heard speech may be privileged from the outset of postnatal life, and may not rely heavily on experience with a specific language. Still more evidence for the robustness of infants' audiovisual speech perception is found in McGurk-like studies, which demonstrate that infants' auditory speech perception is altered by incongruent visual information, though perhaps to a lesser degree than is that of adults (Burnham & Dodds, 2004; Desjardins & Werker, 2004; Rosenblum, Schmuckler, & Johnson, 1997). Finally, a few more recent studies (e.g., Teinonen et al., 2008; Ter Schure, Junge, & Boersma, 2016) have indicated that it may be possible to *boost* infants' discrimination of auditory-only speech sounds with the imposition of congruent visual information. Such results lend support to the notion that visual information in speech perception is not only salient to the developing infant, but may actually aid in the process of language acquisition.

Nevertheless, largely because of the field's history of conceptualizing speech perception as a primarily auditory endeavour, there remain significant questions regarding the development of speech perception in a multisensory context. In this thesis, I ask two main questions and use various behavioural methods with six- to 11-month-old infants to probe them. First, I explore how our understanding of speech perception in infancy, often characterized intentionally or inadvertently as a predominantly auditory process, is modified when we consider carefully the audiovisual nature of speech and of infants' encounters with speech. Second, via a series of experimental manipulations, I attempt to probe more deeply *how* it is that infants detect correspondences in auditory and visual information when processing speech, and whether they rely on *both* phonetic content information as well as temporal information when processing the audiovisual speech signal. I do so by asking whether infants rely more heavily on auditory or visual information when the two provide conflicting content (but consistent temporal) information, and whether the temporal order in which such information is presented modifies this initial bias. To answer both of these main questions, I rely on linguistic stimuli with which the infants I test are unfamiliar, using non-native language to attempt to control for the months of experience with native-language audiovisual speech that infants have accumulated with their native language.

In the first experimental chapter of this thesis, Chapter 2, I probe how a relatively well-understood phenomenon of early speech perception, perceptual attunement, may be modified when we consider speech as an audiovisual signal. In the familiarization phase, I first explore whether (English-learning) infants' sensitivity to the congruence of seen and heard speech—robust in their native language—extends to the auditory and visual components of speech in a non-native language (Hindi), and whether such an ability declines in tandem with auditory perceptual attunement. I test this question using a novel paradigm. Rather than familiarizing infants to auditory speech and testing them using visual speech, as has been done in previous studies (e.g., Pons et al., 2009; Kubicek et al., 2014), I rely on previously established patterns of infants' facial scanning behaviour (Hunnius & Geuze, 2004; Lewkowicz & Hansen-Tift, 2012; Tomalski et al., 2013), and use eyetracking to test explicitly whether infants detect content incongruence in dynamic audiovisual speech and change their face-scanning patterns as a result.

My second question in Chapter 2 probes whether the well-established trajectory of perceptual attunement may be *shifted* by the addition of visual information to the speech signal. I probe this question developmentally with the same three age groups of English-learning infants, by testing their auditory discrimination after the audiovisual familiarization to congruent or incongruent speech described above. As has been reviewed throughout the current chapter, multisensory information appears to promote learning more effectively than does unisensory information. And, as noted, Teinonen and colleagues (2008) and Ter Schure and colleagues (2016) both demonstrated that infants' discrimination of auditory phones may be improved by the addition of visual information. I hypothesize that, by familiarizing infants to congruent audiovisual non-native syllables, their later discrimination of these syllables in an auditory-only test will be improved. I further predict that this effect will be one that specifically depends on the phonetic content congruence between the auditory and visual signals of speech: I posit that familiarization to *incongruent* audiovisual stimuli will not have the same facilitatory effect on auditory discrimination at test.

In Chapter 3, I expand upon the questions addressed in Chapter 2 and further explore the nature of the early audiovisual speech percept by asking *how* it is that infants detect audiovisual correspondence in the seen and heard signals of speech and *what* it is that infants perceive when watching incongruent speech in which the two signals are mismatched. Throughout this current chapter, I have differentiated between the modality-general characteristics shared between auditory and visual speech (e.g., spatial localization, temporal synchrony, rhythm and intensity matching) and the modality-specific characteristics of auditory and visual speech. These latter characteristics, such as the amount of mouth aperture in the visual modality and formant height in the auditory modality, or the specific auditory and visual signals associated with burst release formed bilabially (e.g., /b/, /p/), are encoded in their individual modality-specific signals, but consistently co-occur. Nonetheless, some authors (e.g., Lewkowicz, 2010; Baart et al., 2014) have argued that it is sensitivity to the low-level, modality-general characteristics of auditory and visual speech—not content information within the spectra of the signals themselves—that drives infants' matching of heard and seen signals and allows them to seamlessly integrate the two signals into a unified percept. However, as noted earlier in this chapter, some studies probing infants' audiovisual matching *sequentially* (i.e., without temporal cues) have demonstrated that infants succeed at matching even when modality-general correspondences are absent (e.g., Pons et al., 2009; Kubicek et al., 2014). I thus propose that infants may use *both* temporal and content information in their perception of audiovisual speech, and I probe this interaction by manipulating both sets of cues.

By relying on the same Hindi speech sounds used in Chapter 2 and in other recently reported work (e.g., Bruderer et al., 2015), in Chapter 3 I again attempt to control for the effects of language-specific experience. I then probe whether, when presented with incongruent audiovisual speech in which the auditory and visual signals are isolated from different speech sounds (in this case, Hindi dental and retroflex consonants), infants rely on the visual modality or the auditory modality when processing speech, exhibiting a matching preference for the visual or auditory information at test. I then probe whether the addition of a slight temporal asynchrony modifies this initial bias. Specifically, I predict that the temporal precedence of either the auditory or the visual signal will increase the salience of that signal and cause infants to rely more heavily on it when processing the familiarization stimuli. I propose that the results of the study presented in Chapter 3 may lend support to the notion that infants use both content and temporal information when integrating and processing the auditory and visual signals of speech.

In Chapter 4, I review the results obtained in the two experimental chapters, and argue that infants' abilities in audiovisual speech perception are not the result of experience with specific sound-sight pairings of their native language, but rather that they develop based on endogenous sensory experience and exposure to language in general. Additionally, based on results from both experimental chapters, I posit that infants' detection of audiovisual correspondence in language is not only driven by their use of low-level cues like temporal synchrony, but also by their sensitivity to content congruence, infants' detection of which is as sensitive, and perhaps more so, than that of experienced adults. I conclude by proposing avenues for future investigations to build upon these results, and by suggesting some ways in which these basic findings are applicable to infants' acquisition of language outside the laboratory.

Chapter 2: The organization and reorganization of audiovisual speech perception in the first year of life

2.1 Introduction

For the first six months after birth, infants auditorily discriminate between the speech sounds of many of the world's languages. After about six months of age, evidence of this language-general discrimination begins to decline. Infants' discrimination of sounds in their native language(s) improves, while that of non-native sounds is diminished. The developmental timing of this process, perceptual attunement, is typically quite consistent, leading to the proposal that the middle part of the first year of life is best described as a sensitive or critical period in development. Here, using a novel eyetracking paradigm that relies upon the previously established face-scanning patterns of young infants, we first investigated whether detection of audiovisual congruence in non-native speech follows a similar pattern of perceptual attunement to that of auditory speech sound discrimination. Then we probed whether the temporal trajectory of auditory perceptual attunement can be *modified* by providing infants with richer, audiovisual exemplars of speech prior to testing them auditorily. The results of the present study indicate that the timing of perceptual attunement for detection of congruence in audiovisual speech is similar to that for discrimination of auditory speech contrasts, but may last somewhat longer into ontogeny. Moreover, these results demonstrate that the characteristics of auditory speech perception can be changed by pre-exposure to congruent or incongruent audiovisual speech, but only up to a point in development (about six months of age) when sensitivity to the auditory contrast remains evident. Taken together, these results suggest that our current understanding of perceptual attunement of speech perception can be deepened by considering sensitive periods in a richer, multisensory environment.

2.1.1 Background

Much is known about the organization of speech perception in the first year of life. In the auditory domain, considerable evidence supports the concept of a sensitive period for speech sound discrimination between six and 12 months of age, at the beginning of which infants exhibit sensitivity to the differences between speech sounds in their own and non-native languages, but by the end of which discrimination of non-native sounds has declined. However, evidence also reveals that adults and infants rely on information from sensory modalities outside the auditory domain, including visual and sensorimotor information, when perceiving speech. Nonetheless, with few exceptions, little research has investigated how sensitive period(s) for speech perception may be temporally and/or characteristically different when infants are studied in a multisensory environment rather than a unisensory one. Moreover, the question of whether the addition of information from one modality (e.g., vision) can alter later speech discrimination in another modality (e.g., audition) has not been explored in depth. The current set of studies addresses these issues, and affirms that, even in the young infant, the perception of speech is comprised of inextricably linked information from the auditory and visual modalities.

To enrich and extend current understanding of the development of audiovisual speech perception, we asked two specific questions. First, using sounds with which infants were unfamiliar, we tested whether or not young infants are sensitive to the congruence between the auditory and visual information in the speech signal and, if so, whether their sensitivity is independent of experience with specific sound-sight pairings from the native language. We explored the possibility that such sensitivity, if revealed, might decline in tandem with perceptual attunement, the period in the first year of life when infants' discrimination of non-native speech contrasts declines. Second, we asked whether prefamiliarization to congruent versus incongruent audiovisual speech can alter subsequent auditory-only speech perception, and possibly reveal sensitivity to non-native auditory distinctions beyond the age at which infants typically discriminate non-native sounds.

2.1.2 Perceptual attunement

From a young age, infants auditorily discriminate many of the similar consonant sounds used across the world's languages, regardless of whether such sounds are used to contrast meaning between two words (phonemically) in the language(s) that the child hears. For example, at six to eight months of age, both English- and Hindi-learning infants discriminate between the voiced dental and retroflex consonants of Hindi (/d̪/ and /ɖ/, respectively), though no such phonemic distinction exists in English, and English-speaking adults exhibit no such discrimination (Werker et al., 1981; Werker & Tees, 1984; Werker & Lalonde, 1988). However, by the time they are nine months old, English-learning infants exhibit reduced discrimination of non-native consonantal phonemic distinctions. By 11 months, auditory discrimination of many non-native consonantal phonemes has declined even further, while discrimination of native phonemes has improved (Polka, Colantonio, & Sundara, 2001; Sundara, Polka, & Genesee, 2006; Tsao, Liu, & Kuhl, 2006; Kuhl et al., 2006; Narayan, Werker, & Beddor, 2010).

This pattern of decline in sensitivity to non-native consonant contrasts and improvement in sensitivity to native contrasts across the first year of life is called *perceptual attunement*. Similar findings have emerged for discrimination of tone distinctions (Mattock & Burnham, 2006; Yeung, Chen, & Werker, 2013), and even for the discrimination of handshape distinctions in visual-only sign language (Palmer, Fais, Golinkoff, & Werker, 2012) and for discrimination of articulatory configurations in silent visual-only speech (Weikum et al., 2007; Sebastián-Gallés et al., 2012). The same pattern is seen for perception of vowel distinctions, but may develop earlier than for consonants (Polka & Werker, 1994). The consistency in the timing of this pattern of change, particularly for perception of consonant contrasts, suggests a critical or sensitive period in development between six and 12 months of age, during which the speech input plays an especially important role in changing perceptual sensitivities (Doupe & Kuhl, 1999; Kuhl, 2010; Friederici & Wartenburger, 2010; Werker & Tees, 2005; Maurer & Werker, 2014; Werker & Hensch, 2015).

2.1.3 Audiovisual speech perception

Although the bulk of research in speech perception—and in perceptual attunement—has been conducted by investigating the role of individual modalities, the audiovisual nature of speech perception has nevertheless been well attested in adults. A commonly observed piece of evidence in support of a multisensory view of speech perception is adults' robust ability to speechread: to use visual information from an interlocutor's eyes and mouth to aid in perceiving speech in noise (Sumby & Pollack, 1954; MacLeod & Summerfield, 1987; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998; Grant & Seitz, 2000). Even more evidence comes from the imposition of *incongruent* visual information onto the auditory speech signal. Under certain conditions, when adult listeners are presented with simultaneous auditory and visual signals that conflict with each other (e.g., a visual /bɑ/ and an auditory /gɑ/), an entirely different illusory percept arises (adults report perceiving /dɑ/), a phenomenon known as the McGurk effect (McGurk & Macdonald, 1976; Massaro, Cohen, & Smeele, 1996; Rosenblum & Saldaña, 1992, 1996; *inter alia*).

A growing body of work suggests that speech perception is audiovisual for the infant as well. Infants exhibit the same McGurk effect that adults do (Burnham & Dodd, 2004; Rosenblum, Schmuckler, & Johnson, 1997), although perhaps less strongly (Desjardins & Werker, 2004). Like adults, infants' auditory perception of speech in noise is improved when visual information is added (Hollich, Newman, & Jusczyk, 2005).

Much research on audiovisual processing of speech in infancy has involved cross-modal matching. When shown first a video display of two side-by-side identical faces, one articulating one syllable and the other articulating a different syllable, and are then shown the same video display accompanied by the sound for one of the syllables, infants as young as two months of age look longer to the side articulating the syllable that matches the sound that they hear (Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999, 2002, 2003; MacKain, Studdert-Kennedy, Spieker, & Stern, 1983). This evidence indicates that infants' perception of heard and seen speech is audiovisual from early in life. Moreover, in the first six months of life, infants match audiovisual speech combinations from languages with sounds that are unfamiliar to them (Walton & Bower, 1993; Pons et al., 2009; Kubicek et al., 2014), and even with pairs of non-human animal faces and their vocalizations (Vouloumanos, Druhen, Hauser & Huizink, 2009; Lewkowicz & Ghazanfar, 2006; Lewkowicz, Leo, & Simion, 2010).

Just as the perception of auditory speech attunes in the infant's first year to just those distinctions used in the native language, so too does the matching of the auditory and visual signal. By 11 months of age, infants no longer match heard and seen speech if the stimuli are from a non-native language. For example, six-month-old Spanish-learning infants look longer at a face articulating /bɑ/ (than a face articulating /vɑ/), when hearing the sound /bɑ/, and longer at the face articulating /vɑ/ when hearing the sound /vɑ/, even though Spanish does not use these two sounds contrastively. However, by 11 months of age, Spanish-learning infants no longer match heard and seen /bɑ/ and /vɑ/, whereas infants learning English—in which the distinction is used contrastively—continue to do so (Pons, et al, 2009; but see Kubicek et al., 2014, for possibly contrasting results with 12-month-olds). While this work could be explained solely on the basis of sensitive periods for the attunement of auditory speech perception, Pons and colleagues (2009) argue that their results may also indicate that perceptual attunement is a "pan-sensory" process. Presently, we explore this possibility further by probing whether infants detect (in)congruence in the content of dynamic speech events, even when those events are from an unfamiliar language. If they do, such a finding would provide additional evidence that infants' speech perception is audiovisual, and that it is so independently of infants' experience with a specific language system. Furthermore, the discovery that the decline of such sensitivity to audiovisual congruence follows a different temporal trajectory than does auditory-only speech discrimination could indicate that the sensitive period for speech contrast discrimination is altered when information from more than one modality is taken into account.

In the current study, we operationalize infants' detection of audiovisual (in)congruence by focusing on infants' attention to various areas of a speaker's face while observing speech. While most infants and adults fixate on the eyes of a speaking face (Haith, Bergman, & Moore, 1977; Vatikiotis-Bateson et al, 1998; Cassia, Turati, & Simion, 2004; Hunnius & Geuze, 2004; Merin, Young, Ozonoff, & Rogers, 2006*)*, 8- to 12-month-old infants fixate preferentially on the speaker's mouth, a pattern that is even more pronounced when infants are viewing non-native speech (Lewkowicz & Hansen-Tift, 2012; Kubicek et al., 2013). Lewkowicz and Hansen-Tift (2012) explain this effect by proposing that during the period of perceptual attunement, infants may attend to the visual information provided by the mouth of a speaking face to boost auditory perception and phonetic production. Indeed, children who attend more to their mothers' mouths in early infancy exhibit higher expressive vocabularies in toddlerhood (Young, Merin, Rogers, & Ozonoff, 2009). Moreover, at least one study has demonstrated that infants at six to 12 months of age attend to the mouth region of the face when observing incongruent audiovisual speech in their native language (Tomalski et al., 2013). Taken together, these results suggest that where infants look on the face while they perceive audiovisual speech may provide information about their perception of non-native speech, their sensitivity to audiovisual (in)congruence, and their progress in the developmental trajectory of perceptual attunement.

2.1.4 Visual modification of auditory discrimination

Although prior studies have suggested that infants' speech perception is audiovisual and that infants match auditory and visual content when perceiving speech, it is not known if and how visual information presented in audiovisual speech might change infants' auditory phonetic discrimination before, during, and after perceptual attunement. To the extent that the speech percept is audiovisual for the young infant, the addition of visual articulatory information to the auditory speech signal could alter this discrimination.

Indeed, Teinonen and colleagues (2008) demonstrated that pairing degraded speech sounds with photographs of visual articulations matching those sounds boosted six-month-old infants' ability to discriminate the auditory speech sounds in a later test. Crucially, that study tested infants using native speech sounds with which they were familiar. Also, since infants in their study were presented with only congruent stimuli (albeit minimally so), it remains unclear whether it was necessary that the visually-presented mouth shapes corresponded to the sounds being tested, or whether infants' performance would have been boosted simply by the presence of *any* consistent visual correlate. Another recent study explored a similar question by attempting to boost infants' discrimination of a non-native vowel contrast by pairing visual articulatory information with auditory sounds during a distributional learning paradigm (Ter Schure, Junge, & Boersma, 2016). In that study, infants familiarized to a bimodal audiovisual distribution exhibited a moderate boost to subsequent auditory discrimination. However, as in the study conducted by Teinonen and colleagues (2008), no incongruent visual correlates were tested to determine whether content congruence of the auditory and visual signals was important. Given these results, it is possible that infants' discrimination of similar stimuli in one modality (e.g., audition) can be aided by pairing those items consistently with distinctive stimuli from an additional modality (e.g., vision), even when the link between the auditory and visual items in each pair is arbitrary. Such *acquired distinctiveness* has been shown to boost discrimination of otherwise similar stimuli (Lawrence, 1949; Hall, 1991; Norcross & Spiker, 1957; Reese, 1972), including non-native speech sounds at nine months of age (Yeung & Werker, 2009; see General Discussion).

It is important, however, to note that speech is typically perceived as a dynamic event in which the auditory and visual signals are presented both synchronously and congruently. Thus, it would be informative to determine the extent to which an alteration of infants' auditory discrimination depends on the content congruence between auditory and visual signals. If speech perception is audiovisual from the earliest stages of life, congruent, synchronous visual information could affect subsequent auditory discrimination of these sounds differently than would incongruent information, even when the latter is presented synchronously.

2.1.5 Current study

The current set of studies was thus designed to test two questions. First, we asked whether and how infants detect content congruence in non-native audiovisual speech and whether this sensitivity to congruence declines in tandem with the trajectory of perceptual attunement previously established in auditory perception studies. Our second question probed whether the addition of congruent visual information would alter subsequent auditory discrimination of these same speech contrasts, possibly constituting a shift in the timing of the sensitive period for auditory speech perception. In each of the conditions described here, the Hindi dental-retroflex (/d̪///ɖ/) contrast was utilized. English-learning monolingual infants were sampled from three age populations: at six months, when infants auditorily discriminate the sounds used; at nine months, when perceptual attunement is underway and infants' auditory discrimination abilities of non-native contrasts have begun to decline; and at 11 months, when perceptual attunement for speech sounds has stabilized and infants are expected to fail at discriminating these sounds.

Each of the present manipulations began by familiarizing participants to audiovisual videos of Hindi dental and retroflex syllables. Half of the infants were familiarized to incongruent, temporally aligned audiovisual speech, and the other half was familiarized to congruent, temporally aligned audiovisual speech. To address the first question, infants' familiarization data were analyzed to determine whether, as hypothesized, those familiarized to incongruent speech would exhibit a different pattern of looking to regions of the model's face as compared to those familiarized to congruent speech. A finding of greater looking to the mouth rather than the eyes while watching incongruent audiovisual speech has been demonstrated in infants viewing incongruent speech in their own language (Tomalski et al., 2013). Thus, infants' looking patterns to two anatomical regions of interest (the eyes and the mouth) were measured to determine whether infants in the incongruent familiarization group displayed differential patterns of looking to these facial regions as compared to the infants familiarized to congruent speech.

We predicted that, if detection of audiovisual congruence in unfamiliar speech declines at the same time as does auditory discrimination of unfamiliar sounds, such an effect would be observable in infants before perceptual attunement (at six months of age), attenuated for infants undergoing perceptual attunement (at nine months), and absent once attunement is complete (at 11 months).

To test the second question, following familiarization, infants were tested on discrimination of these same non-native speech sounds auditorily, with no visual information provided. It was predicted that the additional, redundant cross-modal information provided to infants by congruent audiovisual familiarization might enrich their perception of the non-native phonetic contrast, thus boosting subsequent auditory-only discrimination of the contrast for infants undergoing perceptual attunement (at nine months). Moreover, it was hypothesized that *incongruent* audiovisual information would not produce this effect, and might in fact *alter* perception and change the discrimination patterns of the youngest group of infants, who might be more sensitive to incongruence in unfamiliar speech.

2.2 Materials and methods

2.2.1 Sample

Infants were sampled from three different age groups from a database of families recruited from a maternity hospital in Western Canada. Parents of all infants tested reported that their children heard approximately 90-100% English; none heard a language that uses the dental-retroflex contrast phonemically, and none had been diagnosed with an audiological disorder. Infants in the first age group (before perceptual attunement) were six months old (*n* = 32; mean age = 198 days; age range = 182-225 days; 16 females). Infants in the second group (during perceptual attunement) were nine months old (*n* = 32; mean age = 269 days; age range = 256-281 days; 16 females), and infants in the third age group (after perceptual attunement) were 11 months old (*n* = 32; mean age = 329 days; age range = 308-345 days; 16 females). Additional infants were tested and excluded from final data analysis as follows: from the six-month-old sample: two infants due to experimenter error; three infants due to poor eyetracker calibration; 13 infants who did not finish the experiment due to crying or fussiness; from the nine-month-old sample: four infants due to poor eyetracker calibration; eight infants who did not finish the experiment due to crying or fussiness; from the 11-month-old sample: five infants due to poor eyetracker calibration; 11 infants who did not finish the experiment due to crying or fussiness; and three infants due to parental interference during the experiment (e.g., talking, feeding).

2.2.2 Stimuli

One female native speaker of Hindi was recorded to create the stimuli for these experiments. The speaker was video-recorded using a Panasonic AJ-PX270 HD camcorder and a Sennheiser MKH-416 interference tube microphone. During recording, the speaker produced triads of monosyllabic utterances consisting of a target consonant (/d̪/ or /ɖ/) and a vocalic segment (/ɑː/) in infant-directed speech (see Figures 2.1 and 2.2). The speaker was oriented at a 45° angle from the camera, to optimize the viewer's ability to see the orofacial and head motions associated with the two stimulus syllables. For example, the retraction and raising of the tongue tip for the retroflex, /ɖ/, should be produced with the jaw in a lower position and possibly slightly protruded. This may result in less jaw lowering for the following vowel, /ɑː/, compared to that associated with the transition from the dental consonant, /d̪/, to the following vowel, /ɑː/. Another visible difference concerns the tongue tip, which is likely to be visible for the dental consonant, but not for the retroflex.

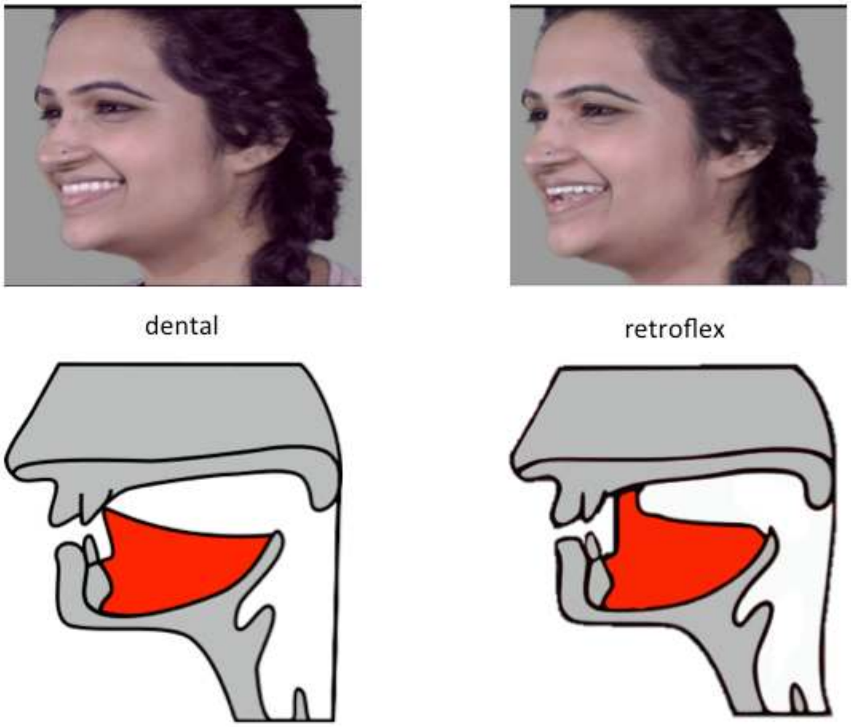

Figure 2.1. Still frames and schematics of model producing dental and retroflex consonants.

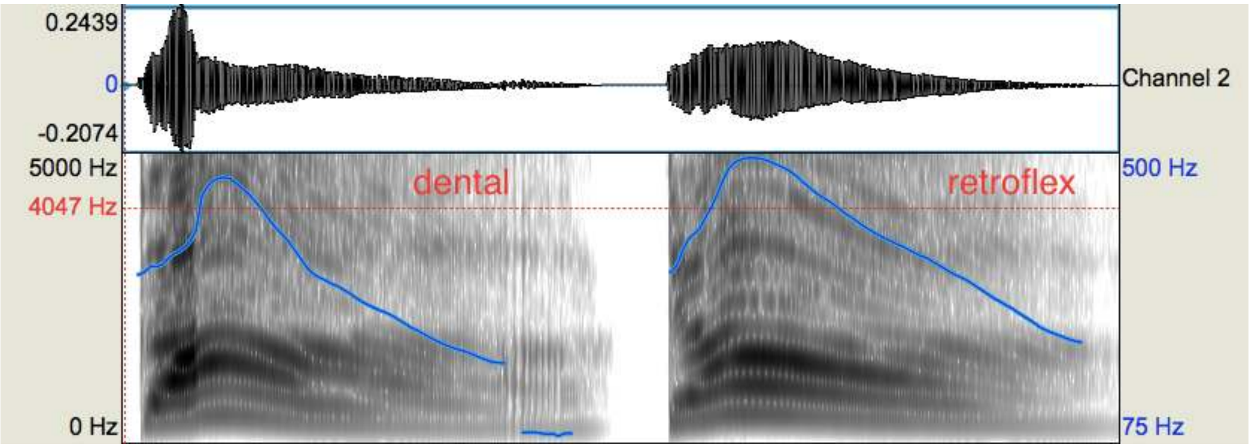

Figure 2.2. Spectrograms of model producing a dental syllable (left) and a retroflex syllable (right).

From this raw material, experimental stimulus items were chosen from among the second items in each triad sequence, in order to control for list intonation effects. Final stimulus tokens were those that had a natural duration between 750 and 1000 ms, and which contained no abnormalities in pitch contour or phonation. Stimuli were then combined to create familiarization sequences and test sequences. Familiarization sequences each consisted of eight audiovisual tokens from the same category (audiovisually congruent /d̪ɑ:/ or /ɖɑ:/, and audiovisually incongruent stimuli with visual /d̪ɑ:/-audio /ɖɑ:/, or visual /ɖɑ:/-audio /d̪ɑ:/). To create incongruent audiovisual stimulus items, visual tracks of stimulus items were spliced with duration-matched auditory tracks from tokens of the opposite phonetic category (auditory /d̪ɑ:/ paired with visual /ɖɑ:/ and auditory /ɖɑ:/ paired with visual /d̪ɑ:/). To ensure that the process of mismatching did not result in asynchronous audiovisual stimuli, consonant burst releases from the original video tokens were aligned with the burst releases of the incongruent, auditory token. The interstimulus interval within the familiarization sequences was 2.2 seconds, and sequences were 20 seconds in total length. Test stimuli were eight-item auditory-only sequences of two types: alternating sequences consisted of tokens from both phonetic categories, while non-alternating sequences consisted of tokens from only one category (Best & Jones, 1998). The interstimulus interval for test sequences was 2.2 seconds and the total length of each test sequence was 20 seconds.

2.2.3 Procedure

All participants were tested in a developmental psychology laboratory at a university in Western Canada. Infants were tested in a dimly lit, sound-attenuated room while sitting on a caregiver's lap. The experimenter and the equipment in the experimental room were hidden from the infant's view by dark curtains. Caregivers, who were asked not to speak to their infants, wore darkened sunglasses to avoid potential interference from their gaze on the eyetracking data, and to prevent their own responses to the stimuli from affecting the responses of the infant.

Infants were seated facing a television screen (101 cm x 57 cm) equipped with a small video camera and a Tobii Technology X60 eyetracker sampling at 60 Hz at a distance of 90 cm from the screen. Stimuli were presented using Psyscope (Cohen et al., 1993). Eyetracker data were recorded using Tobii Studio (Version 1.7.3), and a reference video was recorded with iMovie (Version 9.0.9). Before the study, the eyetracker was calibrated using a five-point visual display with non-linguistic tones to establish each infant's eye gaze characteristics. Prior to familiarization, infants watched an animated waterwheel attention-attractor until they had fixated on the screen. Half of the infants in each age group (*n* = 16) were then familiarized to congruent audiovisual sequences of the Hindi dental and retroflex CV syllables (four dental sequences and four retroflex sequences). The other half of the infants was familiarized to incongruent audiovisual sequences. All stimuli were presented at a mean intensity level of 65 dB, calibrated using a digital sound level meter. Between familiarization trials, infants regained attention to a silent animated ball attention-attractor, and only proceeded to the next familiarization trial after attention was refixated on the screen. The eyetracker provided data indicating where on the screen infants were looking during familiarization, and the duration of fixation to each area of the screen was calculated using Tobii Studio.

After familiarization, all infants were tested using an auditory discrimination task in which they were exposed to eight sequences of auditory test stimuli while watching a still checkerboard. Four of these sequences (*non-alternating* sequences) consisted of test tokens from one phonemic category (/ɖɑ:/ or /d̪ɑ:/), and four sequences (*alternating* sequences) consisted of tokens from both phonemic categories. Trials were separated by the attention-attracting ball, and infants proceeded to the next test trial when they had refixated on the ball. Alternating and non-alternating sequences alternated with one another during the test phase, and counterbalancing ensured that half of the infants began their trials with non-alternating sequences and the other half with alternating sequences. In this procedure, longer looking times to one type of trial (alternating versus non-alternating) indicates discrimination of the sound contrast (Best & Jones, 1998; Yeung & Werker, 2009).

2.3 Results

Familiarization data were analyzed to determine on which anatomical regions of the face the infants fixated during presentation of the audiovisual videos. To code familiarization looking time data, the screen to which infants were fixated was divided into regions of interest (ROI). Although ROIs were defined using static images of the moving faces, they were large enough to cover the entire facial region in question throughout the dynamic audiovisual presentation. One region of interest corresponded to the area surrounding the model's mouth, and the other region of interest to the area surrounding her eyes. Mean differences of looking to the eyes minus the mouth for all ages and conditions are visualized in Figure 2.3. Prior to analyzing familiarization data by age group, a three-way 3 (Age Group) x 2 (Condition) x 2 (Region) mixed-design ANOVA was fitted to an aggregate dataset containing all of the familiarization data from the three age groups. A medium-sized three-way interaction between age group, condition, and region of interest emerged ($F(2,90) = 3.64$, $p = .030$, $\eta^2_P = .07$).

Test data were analyzed to probe auditory discrimination, and specifically to determine whether congruent or incongruent audiovisual familiarization had any effect on discrimination at test, as exhibited by a difference in looking time between alternating and non-alternating stimuli sequences. Of the 768 test trials across the three age groups (32 infants in three groups completed eight test trials each), the eye tracker did not capture looking time data for 15 trials, but no two trials of the same sequence type (alternating/non-alternating) were skipped in an individual infant's dataset. In order to analyze data from all subjects, these 15 points were replaced with each infant's sequence-type-specific mean looking time. Test data can be visualized in Figure 2.4 as differences between looking to alternating over non-alternating trials. Test data were analyzed in pairs of trials. The first pair consisted of the first and second test trials (one alternating and one non-alternating trial); the second pair consisted of the third and fourth test trials, and so on. Prior to analyzing test data by age group, a four-way 3 (Age Group) x 2 (Condition) x 2 (Sequence Type) x 4 (Pair) mixed-design ANOVA revealed a significant effect of age group ($F(2,90) = 3.32$, $p = .041$, $\eta^2_P = .07$) and of pair ($F(3,270) = 33.60$, $p < .001$), $\eta^2_P =$ .27), though no other main effects or interactions emerged as significant. Subsequent test analyses were conducted separately for each age group.

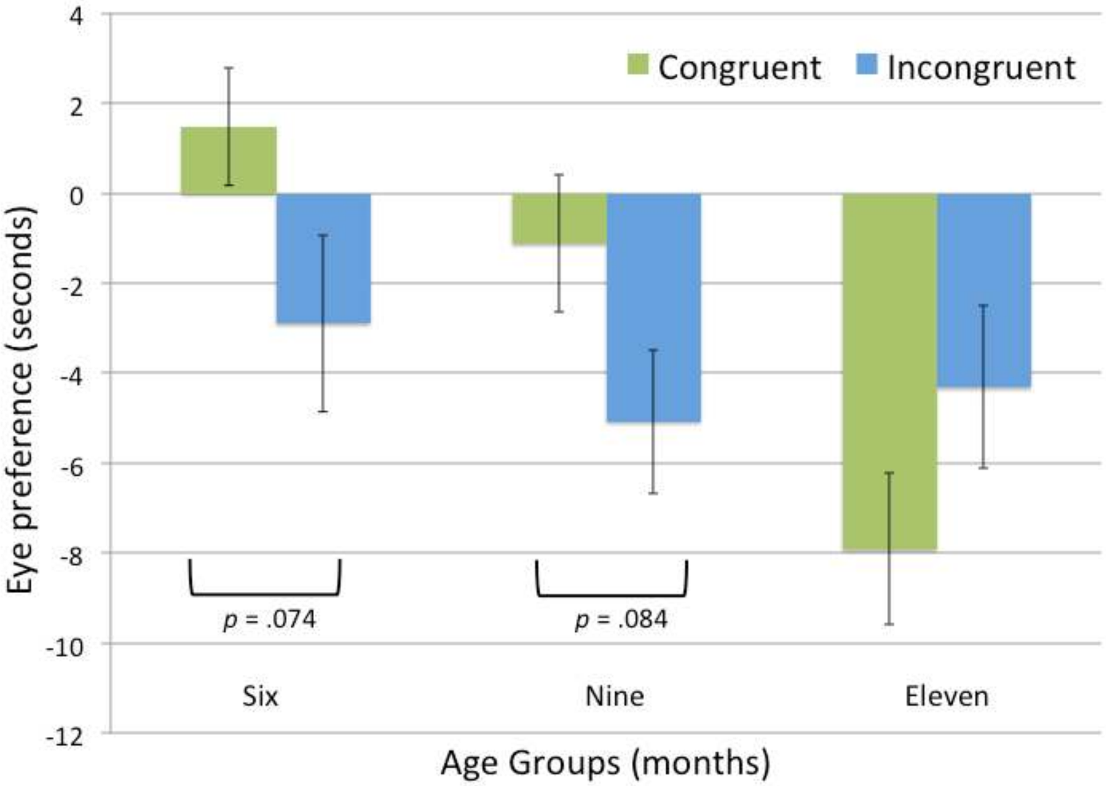

Figure 2.3. Differences in looking time during familiarization to the mouth region of the model's face minus the eye region. Positive scores indicate preference for the eye region. Error bars are +/- one standard error of the mean difference in looking times.

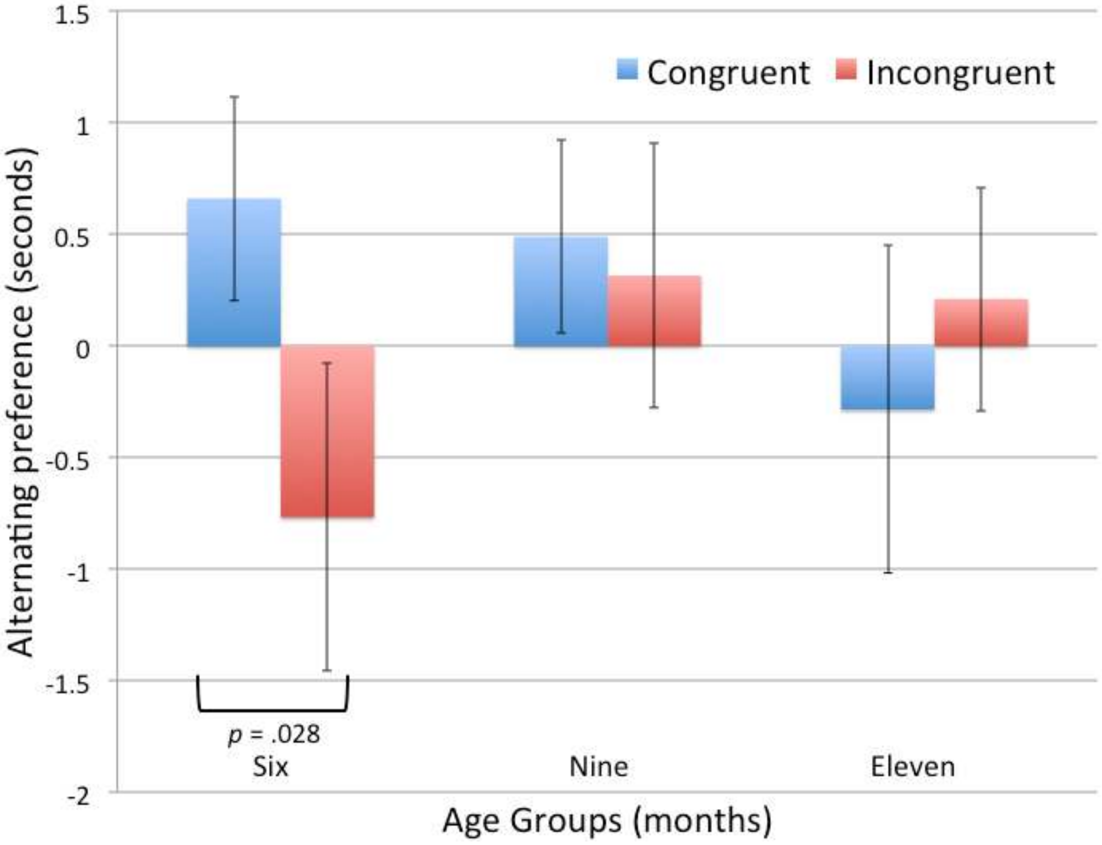

Figure 2.4. Differences in looking time during test to alternating minus non-alternating sequences. Positive scores indicate preference for alternating sequences. Error bars are +/- one standard error of the mean difference in looking times.

2.3.1 Six-month-olds

A two-way 2 (Condition) x 2 (Region of interest) mixed-design ANOVA was performed on the 6-month-olds' familiarization looking time data. There were no main effects of condition ($F(1,30) = .89$, $p = .353$, $\eta^2_P = .03$) or region of interest ($F(1,30) = .36$, $p = .554$, $\eta^2_P = .01$), but a medium-sized interaction between condition and region of interest nearly reached significance ($F(1,30) = 3.43$, $p = .074$, $\eta^2_P = .10$). Infants familiarized to congruent stimuli looked more to the eye region of the model's face (M_{eyes} - M_{mouth} = 1.48 seconds, $SD = .65$) while infants familiarized to incongruent stimuli looked more to the mouth region of her face (M_{eyes} - M_{mouth} = -2.89 seconds, *SD* = .72).

A three-way 2 (Condition) x 2 (Sequence type) x 4 (Pair) mixed-design ANOVA was performed on the six-month-olds' test data. A main effect of pair emerged ($F(3,90) = 7.39$, p = .001, η^2_P= .20) indicating that infants looked progressively less to the screen as the test phase continued, a typical pattern in infant looking time studies. No main effect of condition ($F(1,30)$ = .03, p = .858, $\eta^2_P < .01$) or sequence type ($F(1,30)$ = .02, p = .884, $\eta^2_P < .01$) emerged, but a significant interaction between condition and sequence type ($F(1,30) = 5.30$, p = .028, η^2_P = .15) revealed that the six-month-olds familiarized to congruent stimuli exhibited an alternating preference during test ($M_{\text{alt}} - M_{\text{non}}$ = .66 seconds, *SD* = .59), while those familiarized to incongruent stimuli exhibited a non-alternating preference ($M_{\text{alt}} - M_{\text{non}}$ = -.77 seconds, *SD* = .67).

2.3.2 Nine-month-olds

A two-way 2 (Condition) x 2 (Region of interest) mixed-design ANOVA on the nine-month-olds' familiarization data revealed no main effect of condition ($F(1,30) = 1.49$, p = .232, η^2_P= .05). A main effect of region of interest emerged ($F(1,30) = 7.80$, p = .009, η^2_P = .21), indicating that infants in both familiarization conditions looked longer to the mouth region of the model's face than to the eye region ($M_{\text{eyes}} - M_{\text{mouth}}$ = -3.09 seconds, *SD* = .68). As with the 6-month-olds, there was a medium-sized interaction between condition and region that nearly reached significance ($F(1,30) = 3.19$, p = .084, η^2_P = .10). Although both groups of infants looked more to the mouth, infants familiarized to incongruent stimuli ($M_{\text{eyes}} - M_{\text{mouth}}$ = -5.07 seconds, *SD* = .69) did so more than did infants familiarized to congruent stimuli ($M_{\text{eyes}} - M_{\text{mouth}}$ = -1.12 seconds, *SD* = .68).

A three-way 2 (Condition) x 2 (Sequence type) x 4 (Pair) mixed-design ANOVA was performed on the nine-month-olds' test data. Again, the main effect of pair emerged ($F(3,90) = 12.21$, $p < .001$, $\eta^2_P = .29$) indicating that infants looked progressively less to the screen as the test phase continued. No main effect of condition ($F(1,30) = 1.37$, $p = .251$, $\eta^2_P = .04$) or sequence type ($F(1,30) = 1.17$, $p = .288$, $\eta^2_P = .04$) emerged, nor did the crucial interaction between condition and sequence type ($F(1,30) = .02$, $p = .887$, $\eta^2_P < .01$), revealing that nine-month-olds, regardless of familiarization condition, looked equally to alternating and non-alternating trials at test.

2.3.3 11-month-olds

A two-way 2 (Condition) x 2 (Region of interest) mixed-design ANOVA on the 11-month-olds' familiarization data revealed a medium effect of condition that nearly reached significance ($F(1,30) = 3.82$, $p = .060$, $\eta^2_P = .11$). Infants familiarized to congruent stimuli looked at the model's face more ($M = 12.73$ seconds*,* $SD = 3.37$) than infants familiarized to incongruent stimuli ($M = 10.32$ seconds, $SD = 3.61$). A large main effect of region of interest also emerged ($F(1,30) = 24.56$, $p < .001$, $\eta^2_P = .45$), indicating that infants in both familiarization conditions looked longer to the mouth region of the model's face than to the eye region ($M_{\text{eyes}} - M_{\text{mouth}} = -6.11$ seconds, $SD = .69$). Crucially, the interaction between condition and region of interest did not reach significance ($F(1,30) = 2.12$, $p = .156$, $\eta^2_P = .07$), indicating that the difference in amount of looking to the eyes versus to the mouth did not differ as a function of condition in the 11-month-olds.

A three-way 2 (Condition) x 2 (Sequence type) x 4 (Pair) mixed-design ANOVA was performed on the 11-month-olds' test data. Again, a main effect of pair emerged ($F(3,90) = 16.38$, $p < .001$, $\eta^2_P = .35$) indicating that infants looked progressively less to the screen as the test phase continued. No main effect of condition ($F(1,30) = 1.91$, $p = .178$, $\eta^2_P = .06$) or sequence type ($F(1,30) = .01$, $p = .933$, $\eta^2_P < .01$) emerged, nor did the crucial interaction between condition and sequence type ($F(1,30) = .31$, $p = .582$, $\eta^2_P = .01$), revealing that 11-month-olds, regardless of familiarization condition, looked equally to alternating and non-alternating trials at test.

2.4 Discussion

Children's first year of life is characterized by rapid changes in the perception of speech. More specifically, much evidence indicates that the period between six and 12 months of age constitutes a sensitive period for language learning during which infants' perception of speech sounds becomes specialized for their native language(s), a process referred to as perceptual attunement. However, with the few exceptions reviewed above, most of the research concerning perceptual attunement has been conducted in a unisensory domain, despite increasingly robust evidence that speech perception is audiovisual and that infants (like adults) process speech using information from multiple modalities. We reasoned that the well-established trajectory of auditory perceptual attunement might be better understood by probing the interaction between the perception of visual and auditory speech signals. Specifically, we asked two questions: first, we probed whether infants before, during, and after auditory perceptual attunement are sensitive to the audiovisual congruence of seen and heard speech in a language with which they are unfamiliar. We hypothesized that infants in the early stages of perceptual attunement would be sensitive to the congruence of the auditory and visual signals while viewing speech, as measured by differences in their looking to distinct regions of the face, while older infants would not exhibit such sensitivity. Second, we probed whether the addition of visual information to the auditory signal would change subsequent auditory discrimination of a non-native speech contrast. We hypothesized that, for six-month-old infants prior to perceptual attunement, incongruent—but not congruent—visual information would change subsequent auditory discrimination of the non-native contrast. For infants undergoing perceptual attunement (at nine months), we predicted that congruent —but not incongruent— visual information would extend the observable sensitive period for non-native phoneme discrimination and thus boost subsequent auditory discrimination of the speech sounds. Finally, we predicted that familiarization with audiovisual stimuli would not affect the subsequent auditory discrimination of the 11-month-old infants, regardless of whether the stimuli were presented congruently or incongruently.

Analysis of the familiarization data from the six-month-olds revealed a nearly significant, medium-sized interaction between familiarization condition and region of interest. This result provides some evidence that six-month-old infants detected the content congruence of heard and seen speech as demonstrated by their increased visual fixation to the mouth region of a speaker's face when observing incongruent audiovisual speech. The present finding is consistent with Tomalski and colleagues' (2013) finding that infants shift their attention to the mouth region of the face when perceiving incongruent speech in their own language, but extends their finding by demonstrating that such a pattern is evident when infants are watching unfamiliar speech as well. Notably, nine-month-old infants, in the midst of perceptual attunement and at an age by which auditory discrimination of non-native sounds has declined, also exhibited such a pattern of detection. Although nine-month-olds, as a group, looked longer to the mouth region of the model's face, those familiarized to incongruent audiovisual speech did so more than those familiarized to congruent stimuli. Eleven-month-old infants, having concluded perceptual attunement, exhibited no pattern of incongruence detection. Taken together, these incongruence detection results are comparable to those that have probed the temporal trajectory of auditory speech discrimination and of auditory-visual matching, but indicate that the sensitive period for congruence detection in audiovisual speech may last somewhat longer into ontogeny than the sensitive period for detecting unisensory speech distinctions.

Remarkably, infants in the current set of experiments observed speech sounds with which they were unfamiliar, yet the six- and nine-month-olds exhibited some evidence of sensitivity to the (in)congruence between the auditory and visual signals. Crucially, both the congruent and the incongruent speech stimuli were constructed such that the onsets and offsets of the visual and auditory signals were aligned, thus ruling out the possibility that infants were sensitive to incongruence simply via detection of a temporal mismatch in the audiovisual signal. Instead, it appears as though infants are sensitive to the congruence of finer details in the acoustic and visual signals, despite having had no prior experience with these specific speech sound contrasts, and that such a sensitivity declines in tandem with auditory perceptual attunement. While it is probable that the neural architecture of speech perception in the infant, like in the adult (Campbell, 2008), supports links between heard and seen speech, it is difficult to explain how the mapping could be so precise without specific experience as to enable detection of the differences between congruent versus incongruent auditory-visual dental (/d̪ɑ/) vs retroflex (/ɖɑ/) speech syllables.

One possibility is that infants' sensitivity to audiovisual congruence is mediated by information from infants' proprioception of their own pre-verbal oral-motor movements. Even prenatally, infants engage in frequent sucking and swallowing behaviour (Arabin, 2004; Kurjak et al., 2005), which provides corresponding acoustic information (see also Werker & Gervain, 2013). Moreover, prior to the age at which infants were tested in the current experiments, they begin to produce primitive vocalizations (Oller, Eilers, Neal, & Schwartz, 1999), and their own oral-motor movements affect their discrimination of unfamiliar speech sounds at six months of age (Bruderer et al., 2015). Indeed, one recent study has demonstrated that 4.5-month-old infants' articulatory configurations affect their matching of heard and seen speech, an effect that varies as a function of the specific oral-motor gesture that the infant makes (Yeung & Werker, 2013). Such results advance the proposal that infants' robust audiovisual speech perception may be grounded in early sensorimotor perception (Guellaï, Streri, & Yeung, 2014). Although infants in the current studies had not experienced the specific sound-sight pairings of Hindi in a language-learning environment, their endogenous experience with their own oral-motor movements (and corresponding acoustic productions), in addition to their experience perceiving audiovisual speech in their native language, may have provided them with sufficient information to establish a mapping of the relation between heard and seen speech. This in turn may have enabled them to detect the congruence in unfamiliar audiovisual speech.

Our second question probed whether the addition of visual information to the acoustic signal could change subsequent auditory discrimination of speech sounds before, during, and after auditory perceptual attunement. Analysis of the discrimination data provided evidence that, for the youngest group of infants, visual information indeed affected subsequent auditory discrimination. Infants at six months of age discriminate these non-native speech sounds auditorily, and the addition of congruent visual information to the auditory signal did not disrupt that discrimination. However, the addition of incongruent visual information changed auditory discrimination such that infants familiarized in that condition exhibited a preference for non-alternating acoustic stimuli, while infants in the congruent condition exhibited an alternating preference (as do infants with no audiovisual familiarization (Bruderer et al., 2015; Appendix A, this manuscript)). This result indicates that, prior to perceptual attunement, infants' auditory speech perception is altered by visual information, advancing the proposal that infants' perception of speech is audiovisual.

We were further interested in determining whether the addition of audiovisual information would boost discrimination for older infants who, undergoing or having completed perceptual narrowing, typically do not discriminate these non-native sounds auditorily. The analysis of the nine-month-old and 11-month-old test data revealed no such interaction between condition and sequence type, indicating that auditory perception of non-native speech sounds may only be affected by the addition of visual information prior to perceptual narrowing. Regardless of how they were familiarized, nine- and 11-month-old infants' auditory discrimination at test was not altered by familiarization to audiovisual speech.

The current pattern of results indicates that, prior to the closing of the sensitive period for non-native speech discrimination, auditory discrimination of speech sounds may be changed with the addition of visual information. Importantly, the content of the visual information appears to be crucial. While familiarization to congruent visual information resulted in the maintenance of auditory discrimination prior to perceptual attunement at six months, incongruent visual information changed the pattern of discrimination at this age.

The present findings augment a growing body of recent work aimed at better understanding sensitive periods in language learning from a multisensory perspective. Like ours, a few of these studies have similarly probed whether the addition of visual information to the speech signal would change auditory discrimination as the sensitive period for speech sound discrimination closes. For example, it was recently found that adding a visual display of a speaker producing either an /æ/ or an /ɛ/ to an auditory training procedure improved Dutch 8-month-old infants' sensitivity to this distinction, which they otherwise no longer discriminate at this age (Schure, Junge, & Boersma, 2016). Using a similar set of speech sounds as the ones used in these experiments, another study succeeded in changing infants' auditory discrimination after pairing the sounds with visual objects (Yeung & Werker, 2009; see also Yeung, Chen, & Werker, 2014). In that study, nine-month-old infants were familiarized to sight-sound pairings consisting of one visual novel object paired with one of the Hindi speech sounds (a voiced dental or a voiced retroflex consonant), and a second visual object consistently paired with the other Hindi speech sound. Although the sight-sound pairings were arbitrary, infants exhibited increased discrimination of the auditory speech sounds after familiarization to the object-sound pairings. However, that study used objects, not visual articulations, and no study to date has probed whether the congruence between seen and heard speech affects the manner in which perception is affected by the addition of visual information to the auditory signal.

Our results uniquely contribute to the understanding of how visual and auditory information interact in infant speech perception by demonstrating the differential impact of congruent versus incongruent visual articulatory gestures on auditory discrimination of speech. At the beginning of the sensitive period for speech discrimination, congruent and incongruent audiovisual stimuli affect subsequent auditory discrimination differently. This effect demonstrates that, unlike with arbitrary object-sound pairings, it is not simply the consistent temporal co-occurrence of a sight and a sound in audiovisual speech that can change auditory discrimination. Rather, as the relationship between the acoustic and visual signals of speech is non-arbitrary, the content congruence between the signals determines how visual information from speech will affect auditory perception.

2.5 Conclusion

In the study described in this chapter, we used infants' patterns of face scanning to demonstrate that their detection of congruence in non-native audiovisual speech follows a similar pattern of perceptual attunement to their discrimination of non-native auditory-only contrasts, declining in the period after six months of age. Moreover, we showed that infants' perception of non-native auditory speech is modified by the imposition of incongruent visual information, but only until an age at which discrimination of non-native auditory distinctions has declined. Together, this evidence provides support to the theory that infants are sensitive to the content congruence of the auditory and visual components of speech, and that such sensitivity does not rely upon experience with the sounds and sights of a specific language.

Chapter 3: Infants' use of content and temporal information in the processing of incongruent, asynchronous audiovisual speech

3.1 Introduction

In light of the accumulating evidence that speech perception is multisensory from early in life, and that infants are capable of relying on both low-level temporal information as well as higher-level phonetic content information to match heard and seen speech, the current study used artificially incongruent and asynchronous speech stimuli to probe the interaction between temporal dynamics and content congruence in infants' processing of speech. Specifically, this study tested whether, when presented with simultaneous incongruent audiovisual speech, young infants categorize speech events based primarily on the visual information that they see or the auditory information that they hear, and whether such a tendency can be changed by increasing the salience of one of those sources of information via a temporal offset. Importantly, this study utilized non-native speech to control for experience with, and hence the opportunity to learn the audiovisual pairings of, specific phones. Although the main findings of this study were inconclusive, post-hoc analyses conducted on the data revealed interesting asymmetries in the ways in which infants process incongruent, temporally asynchronous speech, with visual-first speech treated differently than auditory-first speech.

3.1.1 Background

The human speech signal is a complex combination of information from numerous modalities, and speech perceivers utilize multiple sensory systems to detect the various aspects of this informational stream. In addition to the rich acoustic information detected by the auditory system, when the talking face is perceptually available humans also simultaneously detect dynamic visual information in speech. Importantly, adult perceivers of speech do not detect these various streams of information as separate events, but rather as a unified percept. Increasingly, evidence indicates that even young infants detect audiovisual correspondences in speech, and do so even in languages (and non-human animal communication systems) with which they are unfamiliar. It has thus been proposed that the speech percept is supported from early in life by simultaneous information from multiple modalities, and that infants are capable of detecting matches in heard and seen speech even when they have not experienced the relevant sound-sight pairings.

How it is that infants detect such audiovisual (in)congruence in speech is the matter of some debate. Although experienced adults have expectations about how visual language should sound (and how auditory language should look), it is unknown whether infants have similar knowledge about the auditory characteristics of visual phones (and *vice versa*). Does such knowledge drive their detection of audiovisual (in)congruence? Much recent evidence has been interpreted to suggest that infants do not detect correspondences in phonetic content, *per se*, like adults do, but rather that they are sensitive only to low-level modality-general correspondences between heard and seen speech. Namely, it is possible that instead of relying on the match between heard and seen content, infants rely more heavily on the temporal synchrony between the onsets and offsets of auditory and visual speech events in order to match them (see Lewkowicz, 2010, and below, for a review). Such detection of temporal synchrony does not rely upon experience with the sound-sight pairings of syllables in a specific language, and is generalizable to all language (and non-language) events in the environment.

However, recent work (including the data presented in Chapter 2 of this manuscript) indicates that infants are sensitive to the content (in)congruence of unfamiliar audiovisual speech syllables even when the temporal simultaneity of the auditory and visual signals is maintained. Along with other recent literature that has demonstrated infants' abilities to match heard and seen speech even when temporal cues are absent, this result identifies the need for greater investigation into whether and how infants may use *both* temporal and content information when perceiving audiovisual speech.

The current study was designed to probe the interaction between the use of temporal and content properties during infant perception of audiovisual speech. As reported in this chapter, I first probed how infants categorize audiovisual speech in which the content of the two sensory signals is incongruent but the temporal synchrony of the signals is maintained. When they conflict, which of the two sensory signals do infants rely upon to process the speech that is presented? I operationalized this question by probing whether infants subsequently exhibited a matching preference for what they had heard or what they had seen during familiarization. I then tested my main question and probed the interaction between content incongruence and temporal synchrony by offsetting the auditory and visual signals of the incongruent speech by a small interval. I hypothesized that this temporal offset would increase the salience of one type of information (e.g., the auditory) over the other type of information (e.g., the visual), and would shift infants' use of the two types of information, resulting in a different pattern of speech perception than when the two signals were incongruent but simultaneously presented. Together with previous results (including those in this manuscript) indicating that infants can use phonetic content information and temporal information *separately* to match heard and seen speech, the results of this study illuminate how infants may use *both* content and temporal information simultaneously when perceiving audiovisual speech.

3.1.2 Multisensory speech perception

The perception of speech is multisensory. Although earlier models of speech perception focused predominantly on the acoustic perception of speech (see Campbell, 2008, for a review), there is no question that adults use information from multiple modalities to process speech. For example, adults are better at perceiving auditory speech in noise and in silence when that speech is accompanied by visible articulations (Sumby & Pollack, 1954; Summerfield, 1979; Ross et al., 2007; Remez, 2005), and, in a phenomenon called the McGurk effect, conflicting information from the auditory and visual information streams can result in an illusory percept (McGurk & Macdonald, 1976). Adults are also adept speechreaders, even if they have normal hearing, and use visual information from many areas of a speaker's body to enhance speech intelligibility (MacLeod & Summerfield, 1987; Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998; Grant & Seitz, 2000).

It has been known for some time that the speech percept is also multisensory for the infant. Infants match heard and seen speech signals in their own and in non-native languages, as well as in the vocalizations of non-human animals. Perhaps the best known evidence in support of the multisensory nature of speech perception in infancy comes from a series of studies in which infants hear auditory vowel segments of one type (e.g., /ɑ/) while simultaneously viewing silent side-by-side videos of a model producing vowels of two types (e.g., /ɑ/ and /i/). Infants as young as two months of age look longer at the video congruent with the vowel they hear, indicating that they match heard and seen speech signals (Kuhl & Meltzoff, 1982, 1984; Walton & Bower, 1993; Patterson & Werker, 1999, 2003; Yeung & Werker, 2013). Such sensitivity to the match between heard and seen speech signals is not limited to vowels. Infants also match auditory and visual presentations of consonants (Mugitani, Kobayashi, & Hiraki, 2007; Pons et al., 2009).

Moreover, infants match seen and heard speech even when the two information streams are presented sequentially, disrupting the temporal correspondence between the signals (Pons et al., 2009; Kubicek et al., 2014). Strikingly, in both of the above studies, infants also exhibited an ability to match auditory and visual speech signals even in languages with which they were unfamiliar. Infants also appear to have such a matching ability for the auditory and visual signals of non-human primate vocalizations (Lewkowicz & Ghazanfar, 2006). This ability declines midway through the first year of life, similar to the decline in infants' ability to discriminate between non-native speech contrasts.

More evidence for the multisensory nature of speech perception in infancy arises from studies probing the effects of visual facilitation on auditory discrimination. When infants are familiarized with audiovisual tokens of a non-native vowel contrast that they ordinarily do not discriminate, their auditory discrimination of the contrast is boosted (Ter Schure, Junge, & Boersma, 2016). Similarly, infants' discrimination of auditorily degraded consonants is boosted when those consonants are paired with silent visual articulations that match the auditory signal (Teinonen et al., 2008). Importantly, it appears as though visual effects on auditory discrimination vary as a function of audiovisual congruence. Congruent auditory and visual signals modify perception differently than do incongruent auditory and visual signals, even when those signals are from unfamiliar languages (Chapter 2 of this manuscript). It thus appears that the early speech percept is so robustly multisensory that inconsistencies in audiovisual content correspondence change the how speech is processed, and do so independently of experience with specific sound-sight pairings.

3.1.3 Perception of incongruent speech signals

That infants match the auditory and visual signals of speech has been relatively well established. And, as reviewed above, increasing evidence has indicated that infants do so even in languages with which they are unfamiliar. Moreover, in Chapter 2 of this manuscript, I presented evidence that infants not only detected content incongruence while observing non-native audiovisual speech, but also showed that the speech percept is modified when the two sources of content information are incongruent with one another. When six-month-old infants, who discriminate non-native auditory phonemes, were exposed during familiarization to congruent audiovisual speech from a non-native language, they (perhaps unsurprisingly) exhibited a pattern of discrimination at test that was comparable to their patterns of auditory-only discrimination without any audiovisual familiarization. However, when familiarized to incongruent audiovisual speech in which the content of the auditory and visual signals does not match, six-month-old infants' pattern of auditory discrimination shifted. This evidence indicates not only that infants at this age detect audiovisual incongruence in unfamiliar speech, but that such incongruence affects the way in which the audiovisual signal is processed, resulting in differences in looking time at test.

More evidence concerning infants' processing of incongruent audiovisual speech comes from studies probing the McGurk effect and similar phenomena. To review, the McGurk effect was originally demonstrated in adults, who, when presented with simultaneous incongruent audiovisual speech, perceived an illusory phone intermediate to the auditory and visual phones. When most adult English speakers observe an engineered video of a model producing a syllable consisting of the auditory bilabial stop /b/ and the visual velar stop /g/, they report perceiving an intermediate phone, the alveolar stop /d/ (McGurk & Macdonald, 1976). Some studies (e.g., Burnham & Dodd, 2004) have demonstrated that infants also exhibit the McGurk effect when observing this type of stimulus, categorizing the incongruent signals as a third, illusory percept. More recent work has demonstrated that infants change their face scanning patterns in response to McGurk-like speech, visually fixating more on a model's mouth when she produces incongruent, McGurk-like syllables (Tomalski et al., 2013). Such an effect indicates that young infants, like adults, use both auditory and visual information in the process of perceiving speech when the two signals are familiar (native) to the child.

Additional studies have examined *visual capture*, a phenomenon that is closely related to the McGurk effect. Under certain circumstances, when auditory and visual information are incongruent, the visual information overshadows (captures) the auditory information, and, rather than integrating the two sources of information into a McGurk-like illusory percept, adults report perceiving a speech sound that matches the visual phone. A classic example of visual capture consists of a stimulus consisting of an auditory /bɑ/ and a visual /vɑ/. Adults report perceiving /vɑ/, indicating that the visual information overshadows the auditory information in the perceptual process (Rosenblum & Saldaña, 1992, 1996). Infants, too, have been reported to exhibit visual capture using this same combination of auditory and visual phones (Rosenblum, Schmuckler, & Johnson, 1997).

Importantly, the visual and acoustic characteristics of the incongruent phones used in this type of study have an effect on how and whether the McGurk effect or visual capture will occur. In the same visual capture studies cited above, when adults were presented and infants were habituated with an incongruent stimulus consisting of the opposite audiovisual stimuli (an auditory /dɑ/ and a visual /vɑ/), visual capture did not occur. Instead, adults report perceiving simply the auditory stimulus /dɑ/, and infants similarly dishabituate to auditory sequences of /vɑ/

tokens, but not of /dɑ/ tokens, indicating that they, too, perceived /dɑ/. Likewise, in studies probing the McGurk effect, when adults are presented with a stimulus consisting of an auditory /gɑ/ and a visual /bɑ/, integration into an illusory percept does not occur. These asymmetries in audiovisual integration as a function of the specific acoustic and/or visual characteristics of the phones tested is further indication that infants' audiovisual speech perception, like that of adults, is affected by the specific content characteristics of auditory and visual phones.

Other evidence indicates that the visual influence on speech perception may be weaker in infancy than in adulthood (Desjardins & Werker, 2004), and that infants exhibit a wide range of individual differences in their treatment of McGurk-like stimuli (Kushnerenko et al., 2013). Crucially, the original McGurk and Macdonald (1976) study, as well as more recent studies, have demonstrated that the strength of the McGurk effect develops as a function of experience (Dupont, Aubin, & Ménard, 2005; Tremblay et al., 2007; Sekiyama & Burnham, 2008). Specifically, the effect of the incongruent visual signal on perception appears to increase as a child ages. This developmental trajectory is consistent with other investigations suggesting that visual influences on auditory speech perception increase throughout development (Baart, Bortfeld, & Vroomen, 2015; Hockley & Polka, 1994; Massaro, Thompson, Barron, & Laren, 1986; Massaro, 1984) and with experience with a specific language (Werker, Frost, & McGurk, 1992).

A review of the literature probing the McGurk effect in infants and children thus reveals an important effect of experience on an individual's treatment of an incongruent audiovisual speech event. What has not been well explored, however, is how young, inexperienced infants process the divergent sensory signals that they detect in incongruent audiovisual speech when that speech is from an unfamiliar language. How do infants deal with unfamiliar audiovisual speech in which the auditory and visual signals convey different information? Because the modality-specific information in natural speech is produced at the same source (the speaker) and the auditory and visual information conveyed is congruent in content, it is somewhat unclear (and is difficult or impossible to test) how infants use these two sources of congruent information in the perception of natural speech. However, by examining how infants use the auditory and visual signals when the two are divergent, it may be possible to better understand how infants rely on auditory and visual information when perceiving speech.

3.1.4 Temporal influences on audiovisual speech perception

An important consideration in attempting to determine how infants utilize auditory and/or visual information in the perception of speech is the temporal constraints on the presentation of these informational streams. In natural speech, auditory and visual information streams are not only congruent, but are also presented roughly simultaneously (with visual information preceding auditory information by a short interval; see further in this section for more detail). This natural temporal synchrony of audiovisual speech (and, indeed, of non-speech audiovisual events) may provide an important cue to infants, and, though perhaps to a lesser extent, to adults, when matching heard and seen speech.

Adults are particularly adept at detecting temporal synchrony in audiovisual speech and non-speech events. They detect asynchrony in the sound and sight of a bouncing ball or the sound and sight of a falling hammer when the auditory signal precedes the visual signal by as little as 65 ms and 75 ms, respectively (Lewkowicz, 1996a; Dixon & Spitz, 1980). Similarly, adults detect auditory-before-visual asynchronies greater than 80 ms in speech-like events in which a puppet produces a mouth aperture in tandem with a pure tone (McGrath & Summerfield, 1985). Another test of adults' detection of audiovisual synchrony comes from research probing the robustness of the McGurk effect under conditions of temporal asynchrony. When presented with incongruent audiovisual speech in which the auditory information precedes the visual information, adults experience a breakdown in the McGurk effect around 60 ms (Munhall et al., 1996; van Wassenhove, Grant, & Poeppel, 2007). It appears as though the detection of asynchrony in actual fluent human speech is somewhat weaker than in the above examples, with adults' detection of audiovisual asynchrony in fluent speech not occurring when the offset is less than about 130ms (Dixon & Spitz, 1980). It has been argued that detection of audiovisual asynchrony (and subsequent disruption of audiovisual integration) may be greater in syllabic speech than in fluent speech (Pandey, Kunov, & Abel, 1986).

Crucially, adults appear to be more 'tolerant of' (i.e., exhibit reduced sensitivity to) audiovisual asynchrony when the visual information precedes the auditory information than *vice versa*. Each of the examples in the paragraph above lists the threshold for audiovisual integration when auditory information *precedes* visual information. When the opposite is true, when visual information precedes auditory information, detection of asynchrony is weaker and adults require *greater* temporal offset to notice the mismatch in timing. In each of the studies reviewed in the above paragraph, the authors also tested detection of integration when the visual information preceded the auditory information. They found that adults required well over 100 ms of temporal offset to detect audiovisual asynchrony, an amount that varied based on task and stimulus type.

This asymmetry may be a result of the natural temporal precedence of the visual signal. That is, when auditory and visual signals are produced at the same source, as in the environment, the visual information arrives at the perceiver's eye milliseconds before acoustic information arrives at the ear, with the amount of delay of the acoustic signal increasing as a function of the distance between source and perceiver (Burr & Alais, 2006). In speech events, wherein the motor movements of the articulators are reflected in the visual signal prior to producing the acoustic signal, the precedence of visual information may be even more pronounced. However, in natural settings, the human sensory and perceptual systems appear to compensate for this delay. The transduction of the auditory signal, once it arrives at the ear, is a fast neuromechanical process (Corey & Hudspeth, 1979; King & Palmer, 1985). The photochemical and neurochemical process of converting light information to an optical signal is significantly slower (Lennie, 1981; Lamb & Pugh, 1992). Perceptually, the auditory signal appears to be processed more quickly than the visual signal as well, and this difference in processing speed appears to be related to the perceived distance between the source and the perceiver (Burr & Alais, 2006). These processes result in the *perception* of synchronous auditory and visual signals, even though the two are likely detected by the relevant sensory organs at different times. Thus, when visual information precedes auditory information in space, the two signals are perceived as synchronous. However, when the temporal mechanics of the auditory and visual signals are misaligned in the *opposite* direction (i.e., when the auditory signal artificially precedes the visual signal), the temporal asynchrony may be more salient to the perceiver. That is, the natural perceptual correction described above may result in *greater* perceived asynchrony when the auditory signal precedes the visual signal than *vice versa*. This increased perceptual asynchrony may explain adults' greater tolerance for visual-first stimuli over auditory-first stimuli as described in the experimental results above.

Like adults, infants detect asynchrony in the auditory and visual signals produced by multisensory objects and events. They detect asynchrony in moving and sounding objects, such as bouncing disks (Lewkowicz, 1986; Bahrick, 1988; Lewkowicz, 1992a, 1992b, 1996a; Scheier, Lewkowicz, & Shimojo, 2003). They also detect asynchrony in talking and singing faces (Dodd, 1979; Lewkowicz, 1996, 1998, 2000, 2003). Moreover, when habituated to synchronous audiovisual speech, infants dishabituate to asynchronous audiovisual speech, and *vice versa* (Lewkowicz, 2010). It is important to note, however, that the threshold for detection of audiovisual asynchrony appears to decrease across development. Infants are less sensitive than are adults to temporal asynchrony, and exhibit a window of 'tolerance' for asynchrony in fluent speech up to about 500 ms (Lewkowicz, 2010). When habituated to synchronous audiovisual speech, infants do not dishabituate to asynchronous speech when the offset is less than this 500 ms threshold.

Despite infants' lower sensitivity to temporal asynchrony in speech than that of adults, some authors (e.g., Lewkowicz, 2010) argue that it is this detection of temporal synchrony that allows infants to match heard and seen speech in the environment, and that, in turn, bootstraps later phonetic learning. The neural mechanisms necessary to detect audiovisual correspondence are present from early in ontogeny (Bushara, Grafman, & Hallett, 2001), and audiovisual synchrony is the norm in the environment. Early detection of such synchrony may allow infants to use it as a cue for source localization, thus honing attentional resources and improving learning. Moreover, a few studies have demonstrated that infants are capable of detecting audiovisual correspondence in speech even when content information (the information in the auditory and visual signals that allows for speech sound identification) is absent. When infants are familiarized to synchronous audiovisual speech in which the acoustic-phonetic information is replaced by an uninformative pure tone, they nonetheless dishabituate to asynchronous stimuli of the same type (Lewkowicz, 2010). Additionally, infants have been shown to match auditory and visual speech even when the acoustic signal is replaced by sine-wave speech, in which some (though not all) phonetic information is replaced by pure tones (Remez, Rubin, Pisoni, & Carrell, 1981; Baart, Vroomen, Shaw, & Bortfeld, 2014). These results have led to the possible interpretation that infants rely primarily on low-level temporal information (the detection of corresponding onsets and offsets in the auditory and visual signals) to detect congruence in speech. Importantly, it appears as though infants may rely on these cues more than do adults, whose audiovisual speech matching performance is reduced when the acoustic signal is replaced by sine-wave speech (Baart et al., 2014; but see Kamachi, Hill, Lander, & Vatikiotis-Bateson (2003), demonstrating that adults match unknown faces and voices, even when the latter are sine-wave synthesized).

Although it seems clear that infants *can* use temporal information to match auditory and visual speech and that such information may be sufficient to do so, recent evidence indicates that such information is not the only cue that infants are capable of using in detection of congruence between the auditory and visual signals. Pons and colleagues (2009) familiarized infants to auditory-only exemplars of English voiced bilabial stops (/b/) or voiced bilabial fricatives (/v/), and then tested infants using silent side-by-side videos of a speaker producing the two consonants. Infants looked longer at the videos articulating the sound that they heard than the opposite sound, indicating that they matched what they heard and saw. However, this procedure, which never presented the infants with simultaneous audiovisual speech, disrupted infants' possible use of temporal cues to match heard and seen speech. The results of this study seem to indicate that infants are capable of using higher level information, other than temporal synchrony, to match heard and seen speech. Still more evidence for such an ability comes from a similar study in which 4.5-month-old infants were familiarized to videos of silent native or non-native fluent speech and then tested with auditory-only exemplars from the matching or mismatching language. Infants looked longer during auditory test trials that matched what they saw during familiarization, despite the temporal simultaneity of the signals being disrupted by the experimental paradigm (Kubicek et al., 2014).

Most recently, in Chapter 2 of this manuscript, I familiarized infants to audiovisual syllables from an unfamiliar language. Half of the infants were familiarized to incongruent audiovisual speech and half were familiarized to congruent audiovisual speech. Crucially, the temporal synchrony between the auditory and visual signals was tightly maintained in both conditions. Despite not having temporal simultaneity as a cue to congruence (or, perhaps more specifically, despite having temporal simultaneity maintained *regardless* of congruence), infants' face-scanning patterns changed in response to incongruent speech. These results, when taken together with those of Pons and colleagues (2009) and Kubicek and colleagues (2014), indicate that infants must be relying on more than just temporal information when perceiving and matching audiovisual speech. However, as noted, prior studies have demonstrated that infants do use temporal information to match artificial heard and seen speech when higher level, more fine-grained content cues are experimentally removed (Lewkowicz, 2010; Baart et al., 2014). Given these bodies of evidence, an alternative possibility is that infants use *both* fine-grained content information as well as temporal information to make sense of the heard and seen signals of speech.

3.1.5 Current study

The current study was designed to test how the speech percept is affected when both content information and temporal information are mismatched in the auditory and visual signals of unfamiliar speech. In the first place, I probed how infants treat conflicting visual and auditory information when the two are presented synchronously, and sought to determine which of the two sources of information more affected the speech percept. I then manipulated the temporal synchrony of the auditory and visual signals, to determine whether infants' dependence on one signal over the other can be modified by temporal precedence (i.e., one of the two signals coming first). I operationalized these questions by familiarizing and testing English learning infants with phones from Hindi, an unfamiliar language, in order to control for the effects of specific environmental experience with the phones in question. As reviewed above, because visual influences on speech perception increase across development as a function of experience, it was important to use non-native speech in this paradigm. The two sounds used in the current study were the voiced dental (/d̪/) and retroflex (/ɖ/) stops, which differ from each other both acoustically and visually. These two sounds are typically indistinguishable by English speakers and learners older than 9-10 months, but are discriminable by younger English-learning infants and continue to be discriminable by Hindi speakers and learners throughout the lifespan (see Chapter 2 for a review). Thus, the English-learning infants in this study were tested at six months of age, when discrimination of this contrast is still present. I probed my questions by pairing the visual signal from one of the two phones with the auditory signal from the contrasting phone during familiarization. Crucially, six-month-old English learning infants not only auditorily discriminate the exact Hindi speech sounds used in this study (Bruderer et al., 2015; Appendix A, this manuscript), but are also sensitive to the audiovisual congruence of these specific stimuli (Chapter 2, this manuscript). It was therefore assumed that the infants tested in this study would detect the audiovisual incongruence in the engineered stimuli.

The first part of my question tested whether infants would resolve their detection of audiovisual incongruence by relying more on one of the two sensory sources, auditory or visual, to categorize the speech sounds. As a further probe of which modality, auditory or visual, might contribute the most to incongruence resolution, the order of the auditory and visual signals was manipulated in this study temporally. While infants in one condition of the study were tested with synchronous, incongruent audiovisual speech, infants in the two other conditions were tested with temporally asynchronous, but still incongruent, events. As reviewed, infants appear to detect asynchrony in speech (and, therefore, to not integrate the auditory and visual signals into one percept) when the asynchrony exceeds approximately 500 ms (Lewkowicz, 2010). Constructing asynchronous, incongruent stimuli that exceed this threshold would increase the likelihood that infants would perceive the auditory and visual signals separately and fail to integrate them. However, constructing asynchronous stimuli at adult thresholds (e.g., 100 ms) might not have the effect of increasing the salience of the auditory or visual signal for infants whose asynchrony detection levels are higher than adults. It was thus determined that manipulating the temporal precedence of the auditory and visual signals by an intermediate interval (333 ms) could increase the salience of modality-specific information without disrupting audiovisual integration.

Therefore, in one of the asynchronous conditions, visual information was presented before the auditory information by 333 ms. In the other condition, auditory information was presented first. In both cases, the interval by which the two information sources were offset was within the window for temporal integration for infants at this age. The offset was thus designed to manipulate the temporal dynamics of the speech signal, without disrupting the processing of the auditory and visual signals as a unified percept.

After familiarization with the three types of temporally modified, incongruent speech, I tested infants using auditory-only exemplars of the two Hindi consonant-vowel syllables used during audiovisual familiarization. For each infant, these auditory-only test sequences matched either what they saw or what they heard during familiarization. I predicted that infants would resolve the incongruence in the auditory and visual signals in the synchronous condition by relying more upon one source of information than the other, and by exhibiting a looking time preference for those sounds that matched that informational source. Moreover, I predicted that the temporal arrangement of the auditory and visual signals would change the way in which infants resolved audiovisual incongruence. Specifically, I predicted that infants familiarized to audiovisual speech in which the visual information preceded the auditory information would process the speech events by relying more on the visual information, and would exhibit a matching preference for the visually-matched auditory stimuli at test. Likewise, I predicted that infants familiarized to auditory-first stimuli would exhibit matching to the auditorily-matched stimuli at test. Given the review of the literature above, however, it is important to note that an asymmetry in looking time patterns could emerge between the auditory-first and visual-first stimuli. Because visual-first stimuli are more consistent with the natural temporal precedence of the visual signal in the environment, and because detection of visual-first asynchrony is lower than that of auditory-first asynchrony, it is possible that differences could be observed in the auditory-first condition that are not observed in the visual-first and synchronous conditions.

Overall, patterns in line with the ones outlined here would serve to indicate that, when infants are presented with simultaneous incongruent information, they consistently rely on one of the two modalities (auditory or visual) to resolve that incongruence. However, such patterns would also indicate that infants utilize temporal information in addition to their use of content information when categorizing speech, and that temporal precedence can change their reliance on the auditory or the visual signal.

3.2 Material and methods

3.2.1 Sample

Six-month-old infants were sampled from a database of families recruited from a maternity hospital in Western Canada. Parents of all infants tested reported that their children heard approximately 90-100% English; none heard a language that uses the dental-retroflex contrast phonemically, and none had been diagnosed with an audiological disorder. A power analysis prior to data collection (using a partial eta-squared of .07) revealed a minimum sample size of 19 infants per condition (totalling 57 infants). Thus, data was collected from 60 infants, 20 per condition (mean age = 179 days; age range = 167 to 195 days; 30 females). Infants were randomly assigned to one of three familiarization conditions (auditory first, visual first, or synchronous). Sixteen additional infants were tested but were excluded from analysis due to fussiness causing them not to finish the experiment (8), equipment failure (7), or parent interference (1). An additional baby was excluded because she never fixated on the stimuli.

3.2.2 Stimuli

Stimuli for this study were constructed from stimulus items used by Bruderer and colleagues (2015) and in Chapter 2 of this manuscript. As noted and detailed in that chapter, one female native speaker of Hindi was recorded to create the stimuli for these experiments. During recording, the speaker produced triads of monosyllabic utterances consisting of a target consonant (/d̪/ or /ɖ/) and a vocalic segment (/ɑː/) in infant-directed speech. The speaker was oriented at a 45° angle from the camera, to optimize the viewer's ability to see the orofacial and head motions associated with the two stimulus syllables. For example, in the production of the dental consonant /d̪/, but not for the retroflex consonant /ɖ/, the tongue tip is visible between the teeth. Moreover, the retraction and raising of the tongue tip for the retroflex, /ɖ/, should be produced with the jaw in a lower position and possibly slightly protruded. This may result in less jaw lowering for the following vowel, /ɑ:/, compared to that associated with the transition from the dental consonant, /d̪/, to the following vowel, /ɑ:/ (see Figure 2.1, Chapter 2).

From this raw material, experimental stimulus items were chosen from among the second items in each triad sequence, in order to control for list intonation effects. Final stimulus tokens were those that had a natural duration between 750 and 1000 ms, and which contained no abnormalities in pitch contour or phonation. Stimuli were then engineered in Final Cut Pro X (Version 10.2.3) to create familiarization and test stimulus items.

All familiarization items consisted of incongruent audiovisual stimuli, in which the auditory and visual tracks consisted of different consonant types (e.g., a visual dental consonant combined with an auditory retroflex consonant). Familiarization items then varied by condition. In the auditory-first condition, the auditory track of each token preceded the visual track by 333ms. In the visual-first condition, the visual track preceded the auditory track by 333ms. Finally, in the synchronous condition, the two incongruent signals began and ended at the same times. In order to ensure that the auditory- and visual-first conditions did not result in visual or auditory white space at the beginning or the end of the stimulus, the first or last frame of each item was frozen and extended by 333ms. Three tokens of each stimulus type were engineered.

Test sequences consisted of auditory-only dental or retroflex stimuli paired with a black and white checkerboard. Each test sequence consisted of eight tokens of the same type of stimulus (three unique tokens distributed randomly within the sequence). The interstimulus interval for test sequences was 2.2 seconds and the total length of each test sequence was 20 seconds.

3.2.3 Procedure

Infants were seated on their caregiver's lap in a dim room approximately 60 cm from a television screen (101 x 57 cm) equipped with a recording camera. Infant movement and eye gaze information was recorded using iMovie (Version 9.0.9). Stimuli were presented using Habit (Version 1.0). Caregivers were instructed to abstain from directing the infant's attention, either through talking or pointing. Caregivers also wore a pair of darkened sunglasses to maintain blindness to the experimental stimuli, and were instructed to raise their hands during experimentation if the infant became too upset to continue.

Experimentation consisted of three phases: pretest, familiarization, and test (Figure 3.1). A colourful bouncing ball that served as an attention attractor separated adjacent trials of each phase. During the first trial of the pretest, infants observed a colourful waterwheel accompanied by a ringing bell for 20 seconds. This trial served as a measurement for baseline looking time. The following two pretest trials consisted of the white and black checkerboard, paired with the test stimuli that infants would see at the end of the procedure. Infants observed one test trial of each type (dental or retroflex) during this phase, and the order in which they observed each type was counterbalanced within condition. This pretest ensured that infants would be familiar with the checkerboard-sound pairing in order to limit increased looking due to surprise at test.

The familiarization phase immediately followed pretest. During familiarization, infants were presented with tokens according to their familiarization condition (auditory-first, visual-first, or synchronous). From the three possible tokens of each stimulus type, Habit randomly presented one token at a time, with an interstimulus interval of 500 ms. Presentation of these familiarization tokens continued as long as the infant was looking at the screen. If an infant looked away from the screen for more than two seconds at a time, presentation ceased and the attention attractor appeared. This procedure ensured that infants did not perceive auditory-only stimuli during familiarization when they were not oriented toward the experimental apparatus. If an infant maintained attention to the screen, presentation continued for a maximum of 15 seconds before the attention attractor appeared. Infants accumulated a total of 120 seconds of looking before the familiarization phase concluded.

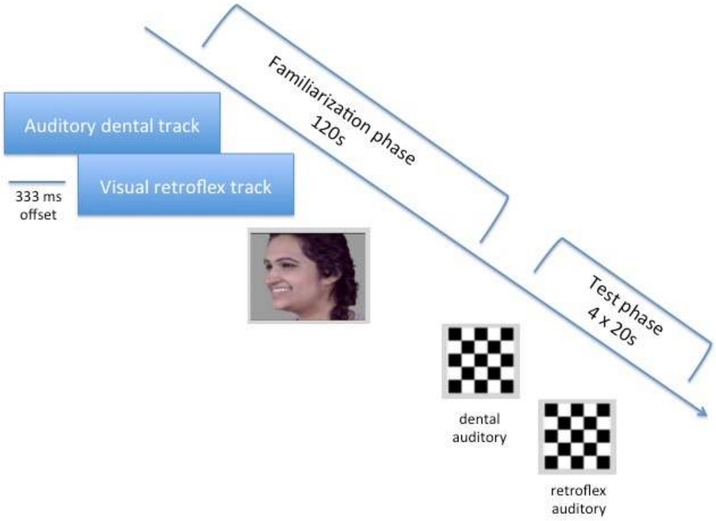

Figure 3.1: Graphical representation of the experimental paradigm, excluding pre-test (auditory-first condition with auditory dental segment and visual retroflex segment).

The four test trials consisted of the same stimuli used in the pretest: black and white checkerboards paired with auditory sequences of dental or retroflex syllables (20 seconds each). Depending on familiarization condition, these syllables matched either the stimulus type that the infant *heard* (auditory match) or the stimulus type that the infant *saw* (visual match). Counterbalancing ensured that half of the infants heard retroflex tokens first while the other half heard dental tokens first, after which point stimulus type alternated for the remaining three trials.

The experiment ended with another presentation of the waterwheel to measure whether looking time decreased substantially during experimentation.

Online coding was employed to ensure that stimuli were only presented when infants were looking at the screen, but no data from online coding were analyzed. Following experimentation, offline coding was performed in silence by a highly trained research assistant, who recorded when during the test trials infants were looking at the screen. From this raw looking time data, sums of looking time per test trial were calculated for each infant.

3.3 Results and analysis

3.3.1 Planned analysis

Test phase looking time data were processed as follows. First, data were examined to ensure that infants looked at each test trial for a minimum of 250 ms. As all 60 infants looked at each of four test trials for the minimum required time, no trials were excluded from analysis. Each infant was tested on two pairs of auditory-only test trials. One test sequence of each pair matched the content of auditory familiarization (auditory match trials) and one sequence of each pair matched the content of visual familiarization (visual match trials). Data from all three conditions are presented by pair (pair 1: test trials 1 and 2; pair 2: test trials 3 and 4) in Figure 3.2 as difference scores (auditory match minus visual match), such that positive scores indicate an auditory match preference.

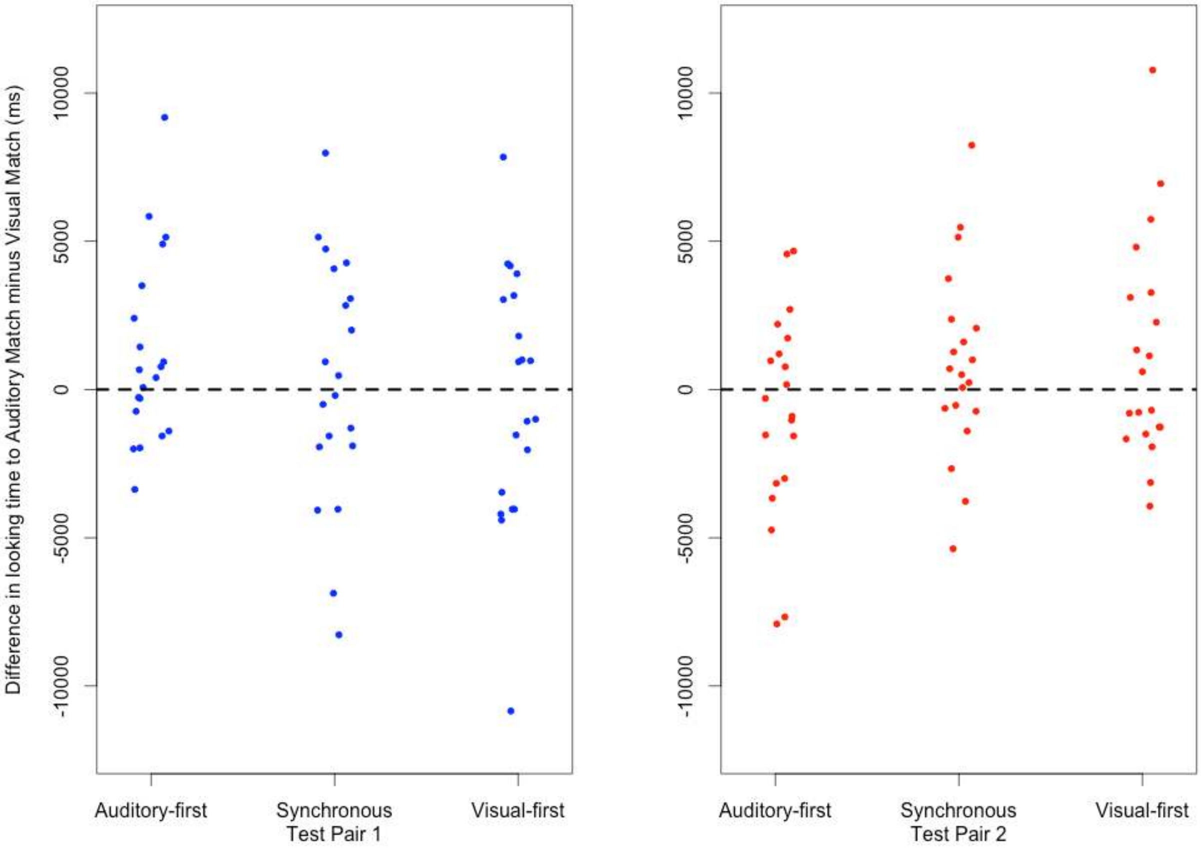

Figure 3.2: Difference scores (auditory match minus visual match) for both pairs of test trials across three conditions (auditory-first, synchronous, visual-first). Positive scores indicate a preference for the auditory match sequences.

As a first step, data from all test trials were analyzed together to determine whether there was any effect of an infant's sex on looking time during test. A 2 x 2 x 3 mixed-design ANOVA with sex (male, female) and familiarization condition (auditory-first, visual-first, synchronous) as between-subjects factors and match type (auditory, visual) as a within-subjects factor revealed no main effects of sex, condition, or match type (all *F*-values < 1.40; all *p*-values > .20). Male and female infants looked equally during the test phase ($t(236) = .69$, $p = .492$; $M_{males} = 12.62$ s,

SD_{males} = 3.85; M_{females} = 12.98 s, SD_{females} = 4.15). All subsequent analyses were conducted without including sex as a predictor.

In the main analysis, a 2 x 3 mixed-design ANOVA with match type as a within-subjects factor (auditory match, visual match) and familiarization condition (auditory-first, visual-first, synchronous) revealed no effects of match type ($F(1,57) = 1.40, p = .241, \eta^2_P = .02$) or of condition ($F(2,57) = 0.36, p = .703, \eta^2_P = .01$) on looking time. Importantly, the ANOVA also revealed no interaction between condition and match type ($F(2,57) = 0.11$, p = .893, $\eta^2_P < .01$), indicating that—regardless of familiarization condition—infants looked equally to the auditory match and visual match sequences at test.

3.3.2 Follow-up analyses

3.3.2.1 Consistent vs. inconsistent temporal arrangement

Examination of Figure 3.2 indicates that, despite the lack of interaction between familiarization condition and match type, infants familiarized to visual-first or synchronous stimuli appear to exhibit increased looking time to the auditory match syllables during the second pair of the test phase. Given the typical temporal arrangement of signals in the speech stream wherein the perception of visual information precedes that of auditory information, there is reason to expect that infants would treat visual-first and synchronous stimuli differently than auditory-first stimuli, which violate the natural temporal dynamics of speech. As such, in the following analyses, a new factor was used to organize the data in order to account for this difference in temporal consistency with natural speech. Infants in the visual-first and synchronous conditions are considered to have been exposed to “consistent” stimuli, while infants in the auditory-first condition are considered to have been exposed to “inconsistent” stimuli.

An exploratory 2 x 2 mixed-design ANOVA conducted on the data from the second pair of test trials was calculated with match type (auditory-match, visual-match) as a within-subjects factor and consistency (consistent, inconsistent) as a between-subjects factor. This analysis revealed a nearly significant interaction between match type and consistency ($F(1,58) = 3.79$, $p = .056$, $\eta^2_P = .06$) on looking time, indicating that infants familiarized to consistent stimuli (visual-first or synchronous) exhibited a greater proportion of their looking time to the auditory-match sequences at test ($M_{\text{auditory}} = 12.60$ s, $SD = 4.48$; $M_{\text{visual}} = 11.59$ s, $SD = 4.22$), while infants familiarized to inconsistent stimuli exhibited no such pattern ($M_{\text{auditory}} = 11.66$ s, $SD = 4.52$ s; $M_{\text{visual}} = 12.49$ s, $SD = 3.71$). Neither main effect emerged as significant (all F-values < .80; all p-values > .35), and an identical ANOVA conducted with the first pair of test trials revealed no main effects or interactions (all F-values < 1.30; all p-values > .260).

3.3.2.2 Phone type

An additional possibility—but one that this experiment was not directly designed to test—is that infants' use of auditory and visual information during familiarization was affected by the specific visual and/or acoustic characteristics of the phones presented. As noted in the introduction to this chapter, although adults' and infants' integration of incongruent auditory and visual signals in their native language is robust for certain phones, it does not occur for all combinations of auditory and visual speech tokens. For example, while the McGurk effect is robust for adults (and to a lesser degree for infants) when they observe a stimulus consisting of an auditory /bɑ/ and a visual /gɑ/, integration does not occur when the modalities of the two phones are reversed (an auditory /gɑ/ and a visual /bɑ/). Likewise, while visual capture occurs for adults and infants in response to an auditory /bɑ/ and a visual /vɑ/, it does not occur when the auditory phone is /dɑ/. Because integration of divergent auditory and visual signals has not been conducted with adult speakers of Hindi using sounds from their native language, it is difficult to predict whether such an asymmetry also exists for the dental and retroflex consonants used in the current study. However, an exploratory analysis can be conducted on the present data.

Infants were familiarized in six sub-conditions that varied in the identity of the auditory and visual phones and in the temporal order in which they were presented. Three of these sub-conditions consisted of a dental auditory signal and a retroflex visual signal, and varied in the order in which the auditory and visual information was presented temporally (auditory-first, visual-first, and synchronous). The other three sub-conditions consisted of a retroflex auditory signal and a dental visual signal, and also varied in the order in which the information in the two modalities was presented temporally. Because of the low number of subjects in each cell, any analysis of the test data by sub-condition is severely underpowered. Nonetheless, an exploratory mixed-design ANOVA conducted on the looking time data for both pairs of trials was calculated with match type (auditory-match, visual-match) as a within-subjects factor and sub-condition as a between-subjects factor. Unsurprisingly, the ANOVA revealed no main effect of match type or of sub-condition, and no significant interaction between the two (all *F*-values < 2.40; all *p*-values > .100).

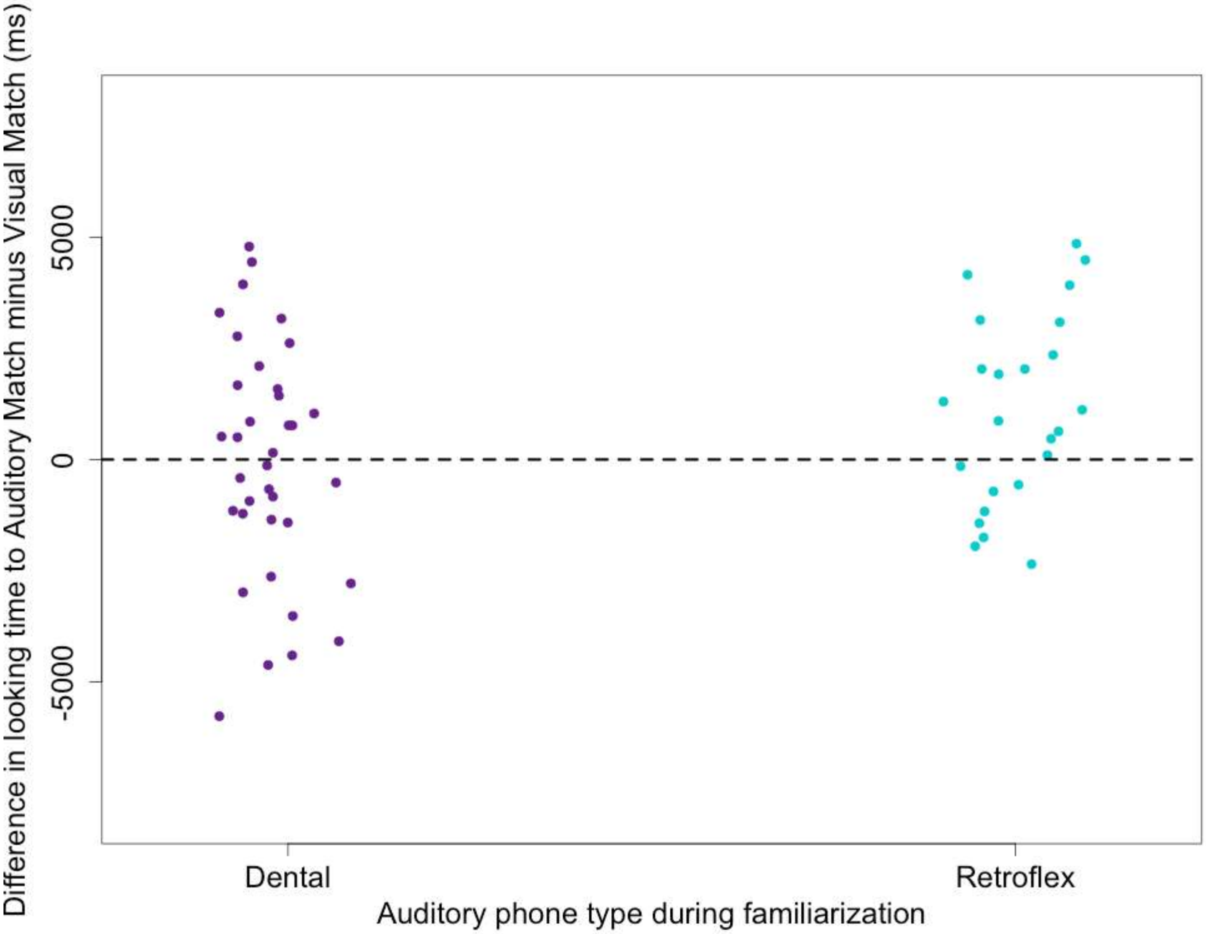

Figure 3.3: Difference scores (auditory match minus visual match) for both pairs of test trials, divided by auditory phone type during the familiarization phase.

In order to increase power and explore whether there was any effect of familiarization phone types on visual or auditory matching at test, data were then collapsed across sub-conditions and split based only on the identity of the auditory phone (dental or retroflex) to which infants were familiarized, regardless of whether visual or auditory information came first or whether they were presented synchronously. These reorganized data are presented in Figure 3.3 as difference scores (auditory match minus visual match). Visual examination of the figure indicates that infants familiarized to retroflex auditory stimuli, regardless of temporal synchrony, may have exhibited a matching preference for retroflex auditory stimuli at test. Another exploratory mixed design ANOVA was fitted to these looking time data, with match type (auditory match, visual match) as a within-subjects factor and auditory familiarization phone (dental, retroflex) as a between-subjects factor. While no main effects emerged as significant in this analysis, a marginal interaction between match type and auditory familiarization phone did emerge ($F(1,58) = 3.35$, $p = .072$, $\eta^2_P = .05$). Infants familiarized to stimuli in which the auditory component was a retroflex consonant exhibited a slight preference for retroflex test sequences. This small effect, though only exploratory, may indicate that infants exposed to auditory retroflex/visual dental syllables during familiarization may have relied on the auditory information to categorize the speech events, regardless of whether that auditory signal preceded, followed, or was presented synchronously with the visual dental signal. Such a result, though only preliminary, is consistent with previous findings testing audiovisual integration in English with adults and infants that have only shown visual capture for some, and not all, audiovisual combinations. As reviewed above, some audiovisual combinations, such as an auditory /dɑ/ paired with a visual /vɑ/, typically result in a /dɑ/ (auditory) percept, with no measurable visual influence on perception.

3.4 Discussion

Speech perception is multisensory, and is so from the early in life. Infants match heard and seen speech in their own and in other languages and non-human vocalizations (Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999, 2003; Pons et al., 2009; Kubicek et al., 2014; Lewkowicz & Ghazanfar, 2006); they (at least to some extent) exhibit the McGurk effect or visual capture when auditory and visual information are mismatched (Burnham & Dodd, 2004; Desjardins & Werker, 2004), and they detect audiovisual incongruence while watching speech

(Chapter 2 of this mansucript). Moreover, like adults, infants' perception or discrimination of auditory speech is improved by the imposition of the visual signal (Hollich, Newman, & Jusczyk, 2005; Teinonen et al., 2008; ter Schure, Junge, & Boersma, 2016), and their categorization of auditory speech is changed when audiovisual congruence is disrupted (Chapter 2 of this manuscript). These results, taken together, seem to indicate that infants are capable of detecting differences in phonetic content between heard and seen speech, and can do so even with sound-sight pairings from languages with which they are unfamiliar.

However, a review of the literature makes clear that infants are capable of relying on more than just content information to match heard and seen speech. Some authors argue that it is the temporal dynamics of speech, and specifically the detection of temporal synchrony in the auditory and visual signals produced in natural speech, that drives infants' ability to match auditory and visual speech events. Like adults, infants are sensitive to audiovisual asynchrony in speech (Lewkowicz, 2010), and—perhaps more so than are adults—infants are able to use temporal information alone, when content information is removed, to match heard and seen speech (Lewkowicz, 2010; Baart et al., 2014).

In the present study, I explored the interaction between content congruence and temporal synchrony in infants' perception of audiovisual speech. Specifically, I asked what information infants rely upon when detecting an incongruent audiovisual speech signal in which the auditory and visual information provide conflicting information, and whether the temporal precedence of one of these signals over the other changes that pattern. If infants integrate the divergent signals into one percept, does that percept more closely resemble what they see or what they hear? And if the auditory information or the visual information comes first, do infants use that preceding information more than if the two informational streams are presented simultaneously?

The planned analysis of the experimental data examined whether—as a function of familiarization condition—infants at test attended preferentially to the sound to which they were familiarized auditorily or the sound to which they were familiarized visually. This analysis revealed no significant effect of familiarization condition on looking time to the auditory match versus the visual match at test, nor did it reveal any general preference for auditory or visual match irrespective of familiarization condition. While the interpretation of a null result must be regarded with some caution, that infants looked equally to the auditory and visual match sequences during test could possibly indicate that the integration of the two signals may have resulted in an intermediate or illusory percept (similar to that observed in the McGurk effect). If such a percept were different than both the retroflex stop and the dental stop, it stands to reason that no matching preference would be shown at test for either of the two consonant types. On the other hand, an equally parsimonious and conflicting interpretation can be made in light of these data. Given infants' detection of audiovisual incongruence in a non-native language, it seems possible that infants were able to simultaneously process both sources of information when categorizing speech during familiarization, and that they perceived the auditory and visual signals separately. Such simultaneous processing of the conflicting visual and the auditory information might even indicate that infants' perception of the auditory and visual information in speech is not fully integrated at this age, particularly when the speech is derived from an unfamiliar language, and that the two percepts remain somewhat independent. Given that visual influences on speech perception appear to increase with age and linguistic experience, this explanation seems plausible.

If infants had used only one type of sensory information when perceiving speech during familiarization, it seems likely that a matching preference would have emerged during test.

Instead, infants may have relied on both sources of information during familiarization (either by integrating the two into a third, illusory percept, or by simultaneously processing both sources of information separately), disrupting any matching at test. Unfortunately, the current paradigm was not designed to differentiate between these two possibilities. Ongoing studies with adult native speakers of Hindi may help determine what illusory percept—if any—is present in the perception of these particular incongruent speech events. Specifically, because it is possible experimentally to ask adults to identify their speech percepts, it may be possible to determine how native speakers treat these engineered stimuli. Do they, like English speakers in McGurk-like studies, integrate the two sources of information into one intermediate percept? Do they process both sources of information separately, or rely only on one modality? Resolution of such questions, along with subsequent testing with young infants using an illusory percept, could explore the possibility that such infants, prior to the establishment of a stable phonological system, also integrate these non-native, incongruent audiovisual signals in a McGurk-like fashion.

Another possibility to explain the lack of overall difference between the temporally asynchronous conditions is that the temporal offset introduced experimentally was simply not long enough to influence infants' speech perception robustly. As reviewed, infants typically require more than 500 ms of temporal offset to detect audiovisual asynchrony in speech (Lewkowicz, 2010). An interval of 333 ms was chosen specifically in this study to be long enough to increase the salience of the preceding signal, but short enough to ensure that infants integration of the auditory and visual signals would not be disrupted. However, given infants relatively weak ability to detect audiovisual asynchrony, 333 ms may have been insufficient to increase perceptual salience of information from one modality.

Nonetheless, despite the null result obtained in the main analysis of these data, a follow-up exploratory analysis conducted on the second pair of test sequences revealed a somewhat different and interesting pattern. Infants familiarized to synchronous or visual-first stimuli did exhibit a matching preference for the auditory stimulus at test. As noted in the analysis, it is typical in infant looking time studies to observe effects on only one pair of test trials. This pattern is especially true when test trials consist of different visual and/or auditory stimuli than does familiarization. In this case, despite attempts to orient infants to the testing checkerboard prior to familiarization, the replacement of the model's face (during familiarization) with a checkerboard (during test) may have temporarily increased infants' looking time during the first test trial and concealed any matching effects.

The combination of the visual-first and synchronous test groups into one condition for exploratory analysis is a principled one. As reviewed in the introduction, although humans perceive the acoustic and light signals of speech (and other non-speech stimuli) simultaneously, the visual signal of speech is produced before the auditory signal, and travels through space more quickly than does the auditory signal. The human perceptual system is thus accustomed to visual information preceding auditory information, and a combination of sensory processing latencies and correction at the neural level typically lead to a resolution of this asynchrony at the perceptual level such that adults do not notice it (Burr & Alais, 2006). Additionally, some theories of multisensory perception (e.g., analysis-by-synthesis (van Wassenhove, Grant, & Poeppel, 2005)) suggest that this preceding visual information prepares the perceptual system to constrain later auditory processing. Moreover, the temporal mechanics of speech production are such that the visual signal, which reflects articulatory motor movements, must slightly precede the auditory signal that those movements produce. Although the perceptual system corrects for the slight visual precedence in natural audiovisual speech, reversing that precedence (as in the case of the experimental auditory-first stimuli) may result in more perceptually salient asynchrony. It thus seems reasonable that infants would process visual-first or synchronous stimuli, which are more consistent with natural stimuli, differently than they would auditory-first stimuli, which are not. In those two conditions, infants may have relied more heavily on one of those informational sources (in this case, the auditory information) when categorizing the speech events.

This pattern, though inconsistent with the main hypothesis laid out in the introduction (namely, that infants in the asynchronous conditions would exhibit a matching for whichever modality-specific information was presented first), might explain the preference for auditory match sequences of infants in the synchronous and visual-first conditions. Although infants familiarized to visual-first stimuli were expected to rely on visual information when processing speech, this preliminary result, that young, inexperienced infants may rely more heavily on auditory information than visual information whenever the stimuli are naturalistic, is consistent with some prior research indicating that visual influences on speech perception are not as robust in infancy and early childhood. The robustness of visual capture on auditory speech perception increases with age and with experience with a specific language, as does the strength of the McGurk effect (McGurk & Macdonald, 1976; Desjardins & Werker, 2004; Werker, Frost, & McGurk, 1992; Dupont, Aubin, & Ménard, 2005; Tremblay et al., 2007; Sekiyama & Burnham, 2008). Additional research using this design but varying the time interval by which signals are offset could better illuminate this asymmetry in the effects of auditory-first or visual-first stimuli.

Another possibility is that, given that the threshold for detection of temporal asynchrony is lower when auditory information precedes visual information than when visual information comes first, infants may not have integrated the auditory and visual signals in the auditory-first condition at all. Indeed, when the temporal asynchrony of two signals is high, it is typical for the integration of those signals to decline (e.g., Munhall et al., 1996). If infants in the auditory-first condition perceived the auditory and visual signals discretely, such an experience during familiarization may have disrupted any matching during test. Again, future work increasing the interval by which the auditory and visual signals are offset could test this possibility. If indeed infants in the auditory-first condition do not exhibit a matching preference at test because they are not integrating the two sources of information, longer intervals should interfere with a matching preference in the visual-first condition as well.

Given that differences in match preferences were only observed on the second pair of test trials, additional research might also examine whether a longer test phase would reveal greater divergences in infant looking time as the test phase progresses. Additionally, although infants were required to attend visually to the familiarization stimuli for two minutes prior to proceeding to test, such exposure may not have been sufficient given the complex, unnatural stimuli. Moreover, as the speech sounds used were unfamiliar to infants, and the checkerboard used during test had only been presented briefly to infants prior to familiarization, an overall novelty preference may have disrupted any preference for auditory-match or visual-match stimuli. Again, a longer test phase may reduce this confounding effect.

An additional exploratory analysis was conducted on the entire dataset to determine whether there was any effect of the specific phones used in this study. As noted in the introduction, studies examining the integration of incongruent audiovisual signals in infancy and adulthood have revealed an asymmetry in the tendency of certain phones to elicit visual capture or the McGurk effect. Because of a lack of power, the current design cannot directly test whether infants were more affected by the temporal order of the auditory and visual signals when the auditory signal was from a dental consonant or from a retroflex consonant. However, the second exploratory analysis indicates that infants familiarized to stimuli in which the auditory component was retroflex exhibited a matching preference for retroflex auditory stimuli at test. Such a finding may indicate that, as has been found in studies probing the perception of incongruent audiovisual speech in English, there is an asymmetry in audiovisual integration based on the specific acoustic and/or visual characteristics of the phones used. It is possible that the retroflex consonant may provide more reliable or salient acoustic information to the perceiver than the incongruent dental consonant provided visual information. When observing a multisensory event, perceivers typically rely on whichever sensory signal provides more reliable information, thus maximizing the accuracy of perception (Ernst & Banks, 2002). If the retroflex consonant provided acoustic information to the infants that was more salient than the visual information provided by the dental consonant, infants may have more heavily relied on the auditory information when processing auditory retroflex/visual dental stimuli. Additional planned testing with Hindi-speaking adults, who can verbally report their percepts, may help to clarify this asymmetry.

3.5 Conclusion

This study was designed to probe the interaction between infants' use of content and temporal informaiton when processing audiovisual speech. It was hypothesized that, when presented with incongruent audiovisual speech in which the auditory and visual signals provided conflicting information, infants would rely on *one* of those sources of information (the auditory or the visual) and would subsequently exhibit a matching preference for that information at test. Moreover, it was predicted that the addition of a temporal offset to the incongruent auditory and visual signals would bias infants’ use of the auditory or visual signal, causing them to rely more heavily on whichever signal was presented first temporally. Although no effects of familiarization condition were revealed in the overall study, an exploratory analysis on the second pair of test trials revealed that infants familiarized to visual-first and synchronous stimuli, which mimic the natural temporal arrangement of auditory and visual signals in the environment, were processed differently than auditory-first stimuli, which violate such an arrangement. Specifically, infants familiarized to more naturalistic stimuli exhibited a preference for auditorily matched stimuli at test, indicating that they may have better integrated and categorized the speech sounds during familiarization by using the auditory information therein. While additional research is necessary to probe the strength of this effect, this study provides some preliminary evidence that, although the early speech percept is multisensory, auditory information may be more informative than visual information when the two provide conflicting information.

Chapter 4: General discussion

4.1 Background and research questions

Perception of the world's objects and events is a multisensory process. Humans (and many non-human animals) detect auditory, visual, haptic, and olfactory/gustatory signals in the environment, and process those signals into unified perceptual entities. This process of integrating signals from multiple sensory modalities appears to be obligatory and automatic, and—as outlined in Chapter 1 of this dissertation—results in perception that is more accurate and faster than unisensory perception alone (e.g., Posner, Nissen, & Klein, 1976; Bahrick & Lickliter, 2000; Vroomen & de Gelder, 2000; Murray et al., 2004). The processing of the human speech signal, present in nearly every typical social interaction from the first moments after birth, is no exception to this pattern. Speech perception is not simply auditory, but multisensory, and particularly audiovisual (Rosenblum, 2005; Ghazanfar & Takahashi, 2014). The production of speech results in signals that can be seen, heard, and felt by a perceiver. These signals are highly correlated not only in time and space, but also in content, and adults and infants are both sensitive to these correspondences. Visual information boosts auditory perception for adults and infants alike (Sumby & Pollack, 1954; Summerfield, 1987; Remez, 2005; Navarra & Soto-Faraco, 2007; Teinonen et al., 2008; Ter Schure et al., 2016), and both groups detect even small asynchronies in the timing of auditory and visual speech (Munhall, Gribble, Sacco, & Ward, 1996; van Wassenhove, Grant, & Poeppel, 2007; Dixon & Spitz, 1980), although the asynchrony detected by infants is not as small as that of adults (Lewkowicz, 1996, 1998, 2000, 2003, 2010). Even hearing adults are capable of speechreading, in which they decipher the content of visual speech without corresponding auditory information (Summerfield, 1992). Infants, too, match heard and seen speech sounds from their native language (Kuhl & Meltzoff, 1982, 1984;

MacKain et al., 1983; Patterson & Werker, 1999, 2003) and, until such a sensitivity declines in the first year of life during perceptual attunement, from unfamiliar languages as well (Mugitani, Kobayashi, & Hiraki, 2007; Pons et al., 2009; Kubicek et al., 2014). Infants appear to be sensitive to the match between heard and seen speech by relying on low-level, modality-general information that the auditory and visual signals share, such as temporal synchrony in their onsets and offsets (Lewkowicz, 2010), and they may even rely more heavily on this type of information than adults do (Baart et al., 2014). But recent research also suggests that infants match heard and seen speech even when temporal information is unavailable or uninformative (Pons et al., 2009; Kubicek et al., 2014).

Nonetheless, speech perception has historically been conceptualized as a primarily auditory process, and not as a multisensory one (Campbell, 2008). As a result, a great deal of the research probing the development of early speech perception has been conducted in the auditory domain, and quite a few outstanding questions remain regarding the nature, development, and limits of audiovisual speech perception. In this dissertation, I have reported my attempts to further develop the understanding of multisensory speech perception by probing a series of questions with young infants at the outset of language acquisition:

1) How does our understanding of a relatively well-understood phenomenon of early auditory speech perception, perceptual attunement, change when we consider carefully the audiovisual nature of speech perception and of infants' early encounters with speech? Specifically:
    a. Do infants detect incongruence in unfamiliar audiovisual speech in the absence of temporal cues to congruence, and does such sensitivity decline in tandem with perceptual attunement?

   b. Can the well-established trajectory of perceptual attunement be shifted by exposing infants to richer audiovisual speech, rather than auditory-only speech?

2) How do infants use both content and temporal information when perceiving audiovisual speech? Specifically:
   a. When observing incongruent, synchronous audiovisual speech, on which modality (auditory or visual) do infants rely to process that speech?
   b. Does changing the temporal dynamics of audiovisual speech affect the way in which infants process the auditory and visual signals?

4.2 Detection of audiovisual (in)congruence and visual influences in auditory perceptual attunement

4.2.1 Detection of audiovisual incongruence

In Chapter 2, I first used a novel methodology to probe whether infants at six, nine, and 11 months of age, prior to, during, and after perceptual attunement, detect audiovisual incongruence in an unfamiliar language. To do so, I familiarized infants to audiovisual exemplars of consonant-vowel (CV) syllables from Hindi, a language with which none of the infants tested were familiar. These CV syllables were minimally contrastive, consisting of either the Hindi voiced dental stop /d̪/ or the voiced retroflex stop /ɖ/, followed by the Hindi long vowel /ɑ:/. In one condition, infants in all three age groups were familiarized to congruent exemplars of these syllables. In another condition, infants were familiarized to incongruent exemplars in which the auditory and visual signals were isolated from different phonetic categories (e.g, an auditory /ɖɑ:/ paired with a visual /d̪ɑ:/), but presented in temporal synchrony. I expected that, if infants at these ages are sensitive to audiovisual congruence even in this unfamiliar language, infants' face-scanning patterns would differ by familiarization condition. Specifically, I predicted that infants familiarized to congruent audiovisual speech would exhibit visual fixation patterns similar to those observed by Lewkowicz and Hansen-Tift (2012), who probed infants' face-scanning patterns when watching native and non-native audiovisual speech. In response to the incongruent stimuli, I predicted that infants would shift their face-scanning patterns, fixating more on the mouth region of the model's face than the infants familiarized to congruent speech. Such a pattern has been observed in response to incongruent speech in an infant's native language (Tomalski et al., 2013), but this present study is the first to probe whether infants exhibit such sensitivity to a language with which they are not familiar.

In the first place, the results of this phase of experimentation comprise a replication of the findings presented by Lewkowicz and Hansen-Tift (2012), as, when familiarized to congruent speech, the face-scanning patterns of infants in all three age groups of my study were similar to those presented in the non-native condition of their study. In addition, my results indicate that six- and nine-month-old infants, prior to and during perceptual attunement, respectively, detect content incongruence in the audiovisual speech of a non-native language. Infants at these two ages who were familiarized to incongruent speech deployed a greater proportion of their overall looking time to the mouth region of the model's face than did those infants familiarized to congruent speech. This result corroborates the findings presented by previous authors (Pons et al., 2009; Kubicek et al., 2014), and indicates that the period of sensitivity to non-native phonemic contrasts, mostly tested in the auditory domain, also extends to the detection of correspondences in audiovisual speech. Further, this result makes a novel contribution to the field by demonstrating that infants—until a certain age—detect content incongruence even when the signals of heard and seen speech are presented synchronously. As such, my finding indicates that infants may rely on more than low-level modality-general cues to match heard and seen speech, and rather that they are sensitive to the correspondence between the fine-grained content details in the two signals as well.

4.2.2 Modifying the trajectory of perceptual attunement

The second question posed in Chapter 2 was whether the well-established trajectory of perceptual attunement could be modified by the imposition of audiovisual speech. As reviewed in Chapters 1 and 2, young infants are sensitive to the acoustic distinctions between consonants and vowels in their native language as well as in many non-native languages, and exhibit such sensitivity until sometime after six months of age (depending on the specific contrast tested) (Trehub, 1976; Aslin, Pisoni, Hennessy, & Perey, 1981; Werker, Gilbert, Humphrey, & Tees, 1981; Werker & Tees, 1984; Werker & Lalonde, 1988; Polka & Werker, 1994; Anderson, Morgan, & White, 2003; see Maurer & Werker, 2014, for a review). However, with a few exceptions (e.g., Pons et al., 2009; Weikum et al., 2007; Sebastian-Galles et al., 2012; Lewkowicz & Ghazanfar, 2006), much of the research conducted on perceptual attunement in language has been conducted in the auditory domain. I proposed that the period of sensitivity to non-native consonant distinctions in the auditory domain might be extended by providing infants with additional information from another sensory modality (vision). As reviewed above, the addition of information from a second sensory modality has been shown to boost discrimination of speech and non-speech stimuli, and I probed whether such an effect might be evident in perceptual attunement as well. Previously reported results (Bruderer et al., 2015; see Appendix A) have established that English-learning six-month-olds, but not the nine- or 11-month-olds, would be sensitive to the acoustic distinctions in the Hindi dental and retroflex stop consonants and would exhibit discrimination of the two. I predicted that, by familiarizing infants to audiovisual exemplars of the consonants prior to testing them, the performance of the nine-month-old infants, in the midst of perceptual narrowing, could be boosted. Crucially, I hypothesized that infants' discrimination of the sounds would only be improved after familiarization to congruent, but not incongruent, audiovisual speech.

To test my question, after familiarizing all infants to either incongruent or congruent speech, as described above, I tested them using an auditory-only discrimination task. The results of the test did not support the hypothesis that discrimination of auditory speech sounds can be boosted with prior audiovisual familiarization, as both the nine- and 11-month-olds, regardless of familiarization condition, failed to exhibit evidence of discrimination at test. However, an interesting pattern of results emerged in the six-month-old sample. Those infants, who ordinarily do exhibit evidence of discrimination of these speech sounds (Bruderer et al., 2015; Appendix A), continued to do so after congruent audiovisual familiarization. However, after incongruent familiarization, infants shifted their pattern of discrimination. Rather than exhibiting greater looking time to the alternating test trials (the typical pattern of discrimination and one exhibited by the infants familiarized to congruent speech), the six-month-old infants familiarized to incongruent speech exhibited longer looking time to the non-alternating trials at test. Such a result, while not evidence that the incongruent familiarization *reduced* later auditory discrimination, does indicate that infants' processing of the audiovisual stimuli was affected by the incongruence in the heard and seen speech signals that they observed, in turn affecting the way in which they listened to subsequent auditory-only speech.

4.3 Infants' use of both temporal and content information in audiovisual speech perception

I tested my remaining questions in Chapter 3, here attempting to further investigate how infants use both fine-grained content information as well as low-level modality-general (temporal) information in their processing of audiovisual speech. Again, I tested these questions using non-native language in order to determine how infants process audiovisual speech prior to specific experience with the phones in question. First, given that six-month-old infants in Chapter 2 exhibited *detection* of audiovisual incongruence in non-native speech, I probed how infants *process* the divergent information from the two sensory signals. I hypothesized that, when observing this type of speech, infants would be sensitive to the content incongruence in the speech signal, and would rely more heavily on either the auditory information or the visual information to decipher the incongruent information presented. Second, I hypothesized that infants' use of auditory or visual content information would be modified by experimentally manipulating the temporal dynamics of the two sensory signals. Previous results have indicated that infants are sensitive to both modality-general (temporal) correspondences in seen and heard speech (Lewkowicz, 1996, 1998, 2000, 2003, 2010; Baart et al., 2014), as well as to fine-grained content correspondences in the auditory and visual signals (Pons et al., 2009; Kubicek et al., 2014; Chapter 2 of this manuscript). I hypothesized that by temporally offsetting the auditory and visual signals of speech such that one type of information preceded the other, infants' use of auditory and visual information in deciphering the speech signal would be modified. Specifically, I predicted that infants would rely more heavily on whichever sensory signal came first temporally, regardless of how they perform when the two signals are presented simultaneously. Such results, I reasoned, would provide evidence that infants can use both auditory and visual content information when processing speech, but that their reliance on one or the other of these signals varies as a function of the temporal dynamics in a speech event.

In one condition of the study presented in Chapter 3, I tested my first question by familiarizing infants to temporally synchronous, incongruent audiovisual speech consisting of the same Hindi dental and retroflex consonants utilized in Chapter 2. I then tested infants using an auditory matching procedure. I predicted that infants would exhibit evidence of matching to either the visual or the auditory signal from familiarization by deploying more of their looking time to the test sequences that were either auditorily or visually matched. Then, in two additional conditions, I again familiarized infants to incongruent audiovisual speech, but temporally manipulated the auditory and visual tracks of the stimuli such that the information from one modality preceded the other by a short interval (333 ms), one that falls outside infants' threshold for the integration of heard and seen speech (Lewkowicz, 2010). Infants in these two temporally asynchronous conditions were tested in the same auditory matching procedure as the infants in the synchronous condition.

The main results of the auditory matching phase in Chapter 3 were inconclusive. When analyzing the test phase in its entirety, infants did not exhibit a pattern of auditory or visual match, regardless of familiarization condition. Infants in all three conditions looked roughly equally to the auditorily and visually matched test sequences. However, post-hoc analyses conducted on the data revealed some interesting patterns. First, when examining only the second (last) pair of test trails, it appeared as though infants familiarized to synchronous or to visual-first stimuli may have exhibited a preference for the auditorily matched sequences at test. Such a pattern seems reasonable, as natural audiovisual speech is actually not audiovisually synchronous. Rather, the visual signal of speech slightly precedes the auditory signal, an asymmetry that is corrected by the sensory and perceptual systems of the perceiver (Burr & Alais, 2005). I therefore examined the possibility that the matching preferences of the infants in the synchronous and visual-first conditions patterned together, and differed from those of the infants in the auditory-first condition. Indeed, infants in the synchronous and visual-first conditions exhibited a preference for auditorily matched sequences in the second pair of test trials, while infants in the auditory-first condition exhibited no preference at test. Although this result is inconsistent with my hypothesis (that infants would rely more heavily on the information that was temporally precedent during familiarization), it provides some initial evidence that—when observing speech in which the temporal dynamics of the auditory and visual signal are more consistent with the natural environment—infants rely more heavily on auditory information to process the speech signal. This result of greater reliance on the auditory signal in these conditions is also consistent with previously reported evidence that humans' susceptibility to visual influences on speech perception increases across the course of development, and is not as strong in infancy as in adulthood (McGurk & Macdonald, 1976; Desjardins & Werker, 2004; Werker, Frost, & McGurk, 1992; Dupont, Aubin, & Ménard, 2005; Tremblay et al., 2007; Sekiyama & Burnham, 2008). It is similarly possible that, throughout the lifespan, the acoustic information in speech is more salient, and thus a more reliable perceptual signal, than is visual information.

A second post-hoc analysis was conducted on the data in Chapter 3 in order to determine whether infants use of auditory or visual information in processing the speech signal was affected by the phonetic characteristics (auditory or visual) of the specific phones used in testing. Previous investigations with English-speaking and –learning subjects have revealed an asymmetry in the tendency of certain visual phones to elicit visual capture or the McGurk effect.

For example, when observing an artificial syllable comprised of an auditory /bɑ/ and a visual /vɑ/, adults and infants appear to exhibit visual capture, categorizing the syllable as /vɑ/ (Rosenblum & Saldaña, 1992, 1996; Rosenblum, Schmuckler, & Johnson, 1997; Desjardins & Werker, 2004). Such an effect does not occur when the auditory syllable is /dɑ/ and the visual syllable is /vɑ/. When exposed to that stimulus combination, English-speaking adults exhibit no influence of the visual signal, perceiving the syllable as /dɑ/. In light of that evidence, it seems possible that the infants tested in Chapter 3 were more susceptible to auditory or visual influences as a function of whether the auditory syllable presented during familiarization was dental or retroflex. To explore this possibility, I collapsed the data across conditions, and reanalyzed infants' matching preferences at test based on whether they were familiarized to auditory retroflex or auditory dental sequences (with corresponding visual signals from the other phone type). I found that infants familiarized to sequences consisting of a retroflex (/ɖɑ:/) auditory component (and thus a dental (/d̪ɑ:/) visual component) exhibited a moderate auditory-match preference at test, while those infants familiarized to the opposite type of stimulus exhibited no matching preference. This result provides additional evidence that, even in a language with which they have no experience, infants are sensitive to the fine-grained content differences between the phones of the language. Such differences, particularly where they render the auditory or visual signal of a phone more salient, may drive infants' use of auditory or visual information in the perception of naturally occurring audiovisual speech.

4.4 Research questions revisited

In light of the evidence presented in Chapters 2 and 3, I revisit the main research questions posed in this dissertation.

How is our understanding of auditory perceptual attunement modified when we consider carefully the audiovisual nature of speech perception and of infants' encounters with speech? First, *do infants detect incongruence in unfamiliar audiovisual speech in the absence of temporal cues to congruence, and does such sensitivity decline in tandem with perceptual attunement?* The answer to both of these questions appears to be yes. Eyetracking results from the familiarization phase of the experiments presented in Chapter 2 indicate that six- and nine-month-old infants familiarized to incongruent audiovisual speech deployed more of their visual fixation time to the mouth region of the speaker's face, while those familiarized to congruent audiovisual speech exhibited face-scanning patterns that are typical of infants at the various ages tested (Lewkowicz & Hansen-Tift, 2012). Given that the only difference between the familiarization conditions was content congruence, this difference alone constitutes evidence that six- and nine-month-old infants in the incongruent conditions detected the content mismatch. However, these results are even more convincing in light of recently reported evidence that infants deploy more of their visual fixation time to the mouth region of a speaker's face when observing incongruent speech in their own language (Tomalski et al., 2013).

Crucially, in Ch 2, 11-month-old infants deployed more of their looking time to the mouth region of the speaker's face, regardless of familiarization condition. This pattern is consistent with results obtained in previous studies indicating that infants at this age deploy more looking time to the mouth region of the face while watching non-native speech (Lewkowicz & Hansen-Tift, 2012; Kubicek et al., 2013). However, the lack of difference in face scanning patterns between familiarization conditions at 11 months indicates that detection of incongruence in non-native audiovisual speech declines along with infants' reduction in sensitivity to non-native acoustic differences across the first year of life. This finding corroborates that of Pons and colleagues (2009) and Kubicek and colleagues (2014), demonstrating that infants' matching of non-native auditory and visual speech declines in tandem with perceptual attunement. However, the current result also extends their findings by demonstrating that, prior to this decline, infants detect content incongruence in audiovisual speech even when the auditory and visual information is presented simultaneously.

Second, *can the well-established trajectory of perceptual attunement be shifted by exposing infants to richer audiovisual speech, rather than auditory-only speech?* The results of the experiment presented in Chapter 2 do not provide evidence that the trajectory of perceptual attunement can be shifted by providing infants with audiovisual exemplars of speech, regardless of whether the auditory and visual signals of that speech is congruent or incongruent. Although pre-attunement six-month-old infants' perception of audiovisual speech was shifted as a function of whether they were familiarized to congruent or incongruent speech, nine- and 11-month-old infants exhibited no evidence of increased auditory discrimination at test. Such a result diverges somewhat from recently reported evidence that audiovisual speech can boost auditory discrimination. In one study, Ter Schure and colleagues (2016) reported preliminary evidence that, when familiarized to audiovisual exemplars of a non-native vowel contrast in a bimodal distribution, auditory discrimination of that contrast was moderately boosted. Although the formants of naturally produced vowels are not steady-state, the relatively long durations of vowel formants compared to the shorter durations of the frequency modulations caused by stop consonants (Repp, 1984; Bouchon, Floccia, Fux, Adda-Decker, & Nazzi, 2015) may provide additional information to the infant. It thus seems possible that the discrimination of non-native vowel contrasts may be more easily boosted than the discrimination of consonant contrasts, even though native-language vowel categories stabilize earlier in development than do consonant categories (Kuhl et al., 1992). Such a difference may explain the divergence in the results obtained in the current study and those of Ter Schure and colleagues (2016). However, the current results also diverge from those reported by Teinonen and colleagues (2008), who demonstrated that infants' otherwise poor discrimination of unimodally distributed, synthetic consonants could be improved by familiarizing them to pairings of the auditory consonants with corresponding visual displays. Although that study used synthetic stimuli, the consonant distinction tested was acoustically familiar to the English-learning infants (/bɑ/ vs. /dɑ/), as were visual articulations accompanying the two sounds. I was interested in probing whether infants' discrimination of a non-native consonant distinction could be boosted by audiovisual familiarization, and therefore sounds that were unfamiliar to the infants both acoustically and visually. Infants' familiarity with the stimuli presented by Teinonen and colleagues (2008), compared to their lack of familiarity with the stimuli used presently, may explain the difference between the results obtained.

In Chapter 3, I probed whether and how infants use both content and temporal information when perceiving audiovisual speech. More specifically, a) *when observing incongruent, synchronous audiovisual speech, on which modality (auditory or visual) do infants rely to process that speech?* And b) *does changing the temporal dynamics of audiovisual speech affect the way in which infants process the auditory and visual signals?* The results of the study reported in Chapter 3 are somewhat inconclusive with respect to these questions. Overall, infants familiarized to incongruent, temporally synchronous audiovisual stimuli did not exhibit a preference for auditorily or visually matched test sequences, and adding a small temporal offset to the auditory and visual signals did not change infants' overall looking time patterns at test. However, when considering the infants familiarized to synchronous, incongruent stimuli in combination with those familiarized to visual-first incongruent stimuli, a moderate preference for auditorily matched test sequences emerged on the second pair of test trials. Although only preliminary, this result indicates that infants may rely more heavily on auditory information in the perception of unfamiliar speech when temporal dynamics of that speech are consistent with those of the natural environment. Adults are less sensitive to asynchrony in audiovisual speech (and non-speech events) when the visual information temporally precedes the auditory information than when the reverse is true (see Chapter 3 for a review). It seems possible that infants, exhibiting the same asymmetry in detection of temporal asynchrony, perceived the visual-first and synchronous stimuli as natural speech, and relied more heavily on auditory information to decipher the speech signal. Such an interpretation is consistent with previously reported findings suggesting that the visual influence on speech perception evident in adulthood emerges across development. Infants and young children exhibit lower effects of visual influence than do adults (McGurk & Macdonald, 1976; Desjardins & Werker, 2004; Dupont, Aubin, & Ménard, 2005; Tremblay et al., 2007; Sekiyama & Burnham, 2008). Moreover, at least one study probing visual capture indicates that such an effect is dependent on language expertise (Werker, Frost & McGurk, 1992). As the infants tested presently were both young (six months of age) and inexperienced with the language used, a tendency to rely on auditory information is somewhat consistent with previously reported results.

Another supplemental analysis probing the effect of the specific speech sounds used in experimentation reveals that infants may have been more auditorily influenced by one type of stimulus (the retroflex consonant) than by the other (the dental consonant). This result is consistent with the widespread asymmetry in studies probing visual influences on speech perception, in which certain auditory phones are susceptible to visual influence and others are not (e.g., McGurk & Macdonald, 1976; Rosenblum, Rosenblum & Saldaña, 1992, 1996; Rosenblum, Schmuckler, & Johnson, 1997).

4.5 Limitations and future directions

The results obtained in the present series of research studies add to a growing body of knowledge about infants' audiovisual speech perception. They provide evidence that young infants detect content incongruence in audiovisual speech, even when they are unfamiliar with the specific sound-sight combinations tested and even in the absence of temporal cues. Additionally, they provide evidence that—when infants detect content incongruence—such incongruence modifies their speech percept. Further, they provide some preliminary evidence that infants may rely more heavily on auditory information when perceiving unfamiliar speech in which the auditory and visual signals conflict, and that their use of auditory or visual information may depend on the specific acoustic and/or visual properties of individual speech sounds. Nevertheless, there are a few ways in which these studies were limited in their generalizability and applicability.

First, although the decision to use non-native speech was a principled one in service of the broader research question, the unavailability of a native language control group[1] limits the interpretation of the current results. Although these studies provide evidence that English-learning infants detect content incongruence in audiovisual exemplars of these Hindi syllables, and although an assumption can be made that Hindi-learning infants would be equally sensitive, such a question has not been tested. As noted throughout this dissertation, Lewkowicz and

[1] Recruitment of Hindi-learning infants as a comparison group is ongoing, but the relatively small size of the population has rendered recruitment difficult. Over the course of 30 months of recruitment, only eight Hindi-learning infants have generated usable data in one condition of these experiments.

Hansen-Tift (2012) discovered that infants' face-scanning patterns were different when observing native speech than when observing non-native speech. Specifically, in their study, English-learning infants deployed more of their visual fixation time to the mouth region of the face when watching non-native (Spanish) speech. The face-scanning patterns observed in Chapter 2 (of English-learning infants observing Hindi speech) are consistent with this pattern. That English-learning infants deployed more of their visual fixation time to the mouth when observing the Hindi speech may have provided them with the opportunity to notice the incongruence in the auditory and visual speech signals. Hindi-learning infants, who—in keeping with the results presented by Lewkowicz and Hansen-Tift (2012)—may have exhibited less looking time to the mouth region in response to native speech, may have actually been less sensitive to content incongruence than were the English-learning infants. On the other hand, research with adults has indicated that deployment of visual fixation time to the mouth region of the face are not necessary to decipher speech (Vatikiotis-Bateson et al., 1998), and it is thus possible that Hindi-learning infants would detect content incongruence even while deploying more of their visual fixation time to the eyes. Probing this same question with Hindi-learning infants in the future would clarify whether the pattern of incongruence detection varies as a function of linguistic experience.

Similarly, it is yet unknown how native speakers and learners of Hindi would process the incongruent, asynchronous speech used in Chapter 3. As noted in that chapter, the answer to this question, particularly if asked of adult speakers who can verbally report what they perceive, would aid in better understanding the current results. As noted, among English speakers with whom most studies of visual capture and the McGurk effect have been conducted, there is an asymmetry in which syllables elicit visual influence on perception and which do not. However, it is not known which consonants elicit such effects in Hindi. Future research conducted with adult speakers of Hindi would provide insight into how this type of speech is processed, and specifically whether and how divergent visual and acoustic information from the dental and retroflex consonants can be integrated. In turn, those results might better explain the findings outlined in Chapter 3. For example, if Hindi-speaking adults report perceiving a retroflex consonant when presented with an incongruent auditory retroflex-visual dental syllable, such a finding might explain why infants familiarized to auditory retroflex syllables, regardless of temporal dynamics, exhibited an auditory match preference at test.

Methodologically, there are a few areas in which the current experimental designs could be modified in an attempt to further explore the current research questions. As detailed in Chapter 2, infants were exposed to audiovisual tokens of speech prior to an auditory-only test of discrimination. However, the familiarization phase was consistent in duration for all infants, regardless of the amount of looking that they deployed. As a result, some infants (i.e., those who looked at the experimental apparatus consistently) were exposed to more audiovisual speech than were others (i.e., those who spent a greater proportion of their time looking away from the screen). Although a supplemental analysis conducted to determine whether amount of looking during familiarization predicted discrimination at test did not reveal any effect, requiring infants to accumulate a certain amount of looking time during familiarization before proceeding to test could give them greater opportunity to use the audiovisual information to discriminate the non-native speech sounds. Indeed, in one study that succeeded in boosting nine-month-olds' discrimination of similar speech sounds by pairing them with novel, non-speech objects, 120 s of accumulated looking time during familiarization was required before infants proceeded to the test phase (Yeung & Werker, 2009).

Another methodological detail that may have affected the present findings is discussed briefly in Chapter 3. In the studies presented in that chapter, infants were familiarized to synchronous audiovisual speech and, in two separate conditions, to asynchronous audiovisual speech in which the two sensory signals were offset by 333 ms. The choice of this particular interval was guided by previous research. Prior studies have demonstrated that infants' detection of audiovisual asynchrony is weaker than that of adults, and infants require more temporal offset (greater than 500 ms) to detect asynchrony (see Lewkowicz, 2010, for a review). Because the questions posed in Chapter 3 concerned infants' use of the auditory and visual signals in tandem and probed the nature of the resulting percept, an interval of 333 ms was chosen in order to avoid infants' treating of the auditory and visual stimuli as separate events, while still being large enough to increase the salience of one signal over the other. However, given that infants do not detect temporal asynchrony explicitly when the auditory and visual signals are offset by less than 500 ms, it is possible that the temporal manipulation in the study presented in Chapter 3 was insufficient to change infants' perception. Further research conducted using multiple temporal offset windows could determine whether the size of the interval used presently caused an attenuation of the hypothesized effect.

4.6 Applicability to typically developing and special populations

It would be beneficial to determine to what extent infants' sensitivity to temporal and content correspondences in audiovisual speech actually aids in their acquisition of the phonological systems of their native language(s). Although the present research and the previous studies reviewed throughout this dissertation provide evidence that infants use both content and temporal information in matching seen and heard speech, and that the addition of visual information modifies auditory perception, it is not known how an individual infant's ability to detect congruence in audiovisual speech and/or sucseptibility to visual influences in speech perception correlate with later language outcomes. Previous studies have demonstrated that individual differences in infants' early phonological development (e.g., their individual pattern of auditory perceptual attunement) correlate with higher level linguistic skills later in ontogeny (e.g., vocabulary size) (Molfese, 2000; Tsao, Liu, & Kuhl, 2004; Kuhl, Conboy, Padden, Nelson, & Pruitt, 2005). Longitudinal studies examining the relationship between early phonological development and later language abilities should also take into account visual influences on speech perception that are evident in the first year of life.

Moreover, no known studies have examined how infants with visual impairments differ in their discrimination of auditory speech sounds, though it has been established that visually impaired adults exhibit greater better auditory spatial tuning in some (though not all) tasks (e.g., Röder et al., 1999; Finocchietti, Cappagli, & Gori, 2015) and that they exhibit superior auditory discrimination of certain vowel contrasts, compared to their sighted peers (Ménard, Dupont, Baum, & Aubin, 2009). If visual information is indeed an important tool in infants' development of perceptive language, as the results of the present studies suggest, differences might emerge in the discrimination of non-native speech sounds (and, perhaps, in native speech sounds as well) when examining infants with visual impairments.

Similarly, infants with auditory impairments might exhibit differences in their treatment of incongruent and/or asynchronous speech. Adults with auditory impairments exhibit enhanced visual speech perception, and often are better speechreaders than are normally hearing adults (Auer & Bernstein, 2007). However, at least one study has demonstrated that the ability to integrate auditory and visual signals from speech is reduced in hearing-deprived children with cochlear implants (Bergeson, Houston, & Mitamoto, 2010), and that their audiovisual speech perception is dominated by information from the visual modality (Schorr, Fox, van Wassenhove, & Knudsen, 2005). Importantly, the latter study also demonstrates that audiovisual integration outcomes are improved if cochlear implantation occurs before 2.5 years of age, suggesting a possible sensitive period for the integration of heard and seen speech (see also Werker & Hensch, 2015). These results indicate that it may be important to extend the basic findings gleaned from experimentation with typically developing infants in the laboratory to atypically developing populations, in order to determine the relative importance of intermodal sensitivity in the acquisition of language. Specifically, it would be informative to determine whether, first, there are differences between typically developing and sensory deprived populations in their performance on the type of task outlined in this thesis, and—if so—whether such differences correlate with higher level linguistic outcomes later in development.

It is also important to note that findings gleaned from testing typically developing infants in the laboratory may not be generalizable to the entire population. Even within the broad spectrum of typical development, there is significant cultural and socioeconomic variation in the amount of face-to-face and verbal interaction that infants have with their caregivers (LeVine et al., 1996; Ochs & Schieffelin, 1984; Richman, Miller, & LeVine, 1992). In general, North American caregivers and those from other Western, industrialized cultures engage in more frequent en face interactions with their infants (Fogel, Toda, & Kawai, 1988; Kärtner, Keller, & Yovsi, 2010), a pattern even more pronounced in high-SES families (Beckwith & Cohen, 1984). Given the relatively high amount of "face-time" likely provided to the high SES, Canadian infants tested in these studies, it is possible that—due to this additional exposure—they are better tuned to the complex correlations between the visual and auditory kinematics of speech. Such exposure would still not have allowed the infants tested to learn the specific sound-sight mappings of Hindi, an unfamiliar language, but may have boosted their sensitivity to the match between the auditory and visual signals of speech in general. Testing infants from other cultural and/or socioeconomic backgrounds, while also collecting data on the amount of en face interactions that infants have with their caregivers, could illuminate the importance of face-to-face interactions in boosting audiovisual speech processing during the period of rapid language acquisition in the first year of life.

4.7 Conclusion

Using the experiments described in this dissertation, I have used novel methods to advance evidence that young infants, prior to and during perceptual attunement, detect content incongruence between the auditory and visual speech signals of an unfamiliar language. Remarkably, they appear to do so without the use of low-level temporal cues, which were not available to them in my experimental design. Instead, I propose that infants at these ages are—like adults—sensitive to the fine-grained content details of the seen and heard signals of speech, and are able to use their expectations of correspondence between those signals to detect incongruence. Further, I demonstrate that incongruence in auditory and visual signals changes infants' perception of speech, even in a language with which they have no expertise. Finally, I present preliminary evidence and suggest future work to more clearly determine that—as infants have access to both temporal and content information in the auditory and visual signals of speech—they use both of these sources of information when processing audiovisual speech. Taken together, the results of this dissertation bolster the advancing theoretical orientation that speech perception is robustly multisensory from birth, and is so without a reliance on experience with the specific sound-to-sight pairings of an individual child's language(s).

References

Anderson, J. L., Morgan, J. L., & White, K. S. (2003). A statistical basis for speech sound discrimination. *Language and Speech, 46*(2-3), 155-182. doi:10.1177/00238309030460020601

Arabin, B. (2004). Two-dimensional real-time ultrasound in the assessment of fetal activity in single and multiple pregnancy. *The Ultrasound Review of Obstetrics & Gynecology, 4*(1), 37-46. doi:10.1080/14722240410001700258

Aslin, R. N., Pisoni, D. B., Hennessy, B. L., & Perey, A. J. (1981). Discrimination of voice onset time by human infants: New findings and implications for the effects of early experience. *Child Development, 52*(4), 1135-1145. doi:10.2307/1129499

Auer, E. T., & Bernstein, L. E. (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment. *Journal of Speech, Language, and Hearing Research, 50*(5), 1157-1165. doi:10.1044/1092-4388(2007/080)

Baart, M., Bortfeld, H., & Vroomen, J. (2015). Phonetic matching of auditory and visual speech develops during childhood: Evidence from sine-wave speech. *Journal of Experimental Child Psychology, 129*, 157-164. doi:10.1016/j.jecp.2014.08.002

Bahrick, L. E. (1988). Intermodal learning in infancy: Learning on the basis of two kinds of invariant relations in audible and visible events. *Child Development, 59*(1), 197. doi:10.2307/1130402

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology, 36*(2), 190-201. doi:10.1037//0012-1649.36.2.190

Beckwith, L. & Cohen, S. E. (1984). Home environment and cognitive competence in preterm children during the first 5 years. In A. W. Gottfried (Ed.), *Home Environment and Early Cognitive Development: Longitudinal Research* (pp. 235-271). Orlando: Academic Press, 1984.

Bergeson, T. R., Houston, D. M., & Miyamoto, R. T. (2010). Effects of congenital hearing loss and cochlear implantation on audiovisual speech perception in infants and children. *Resorative Neurology and Neuroscience*, *28*, 157-165. doi:10.3233/RNN-2010-0522

Bertelson, P., & Aschersleben, G. (1998). Automatic visual bias of perceived auditory location. *Psychonomic Bulletin & Review, 5*(3), 482-489. doi:10.3758/bf03208826

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Haskins Laboratories Status Report on Speech Research, 109/110*, 89-108.

Best, C., & Jones, C. (1998). Stimulus-alternation preference procedure to test infant speech discrimination. *Infant Behavior and Development, 21*, 295. doi:10.1016/s0163-6383(98)91508-9

Bouchon, C., Floccia, C., Fux, T., Adda-Decker, M., & Nazzi, T. (2014). Call me Alix, not Elix: Vowels are more important than consonants in own-name recognition at 5 months. *Developmental Science, 18*(4), 587-598. doi:10.1111/desc.12242

Bruderer, A. G., Danielson, D. K., Kandhadai, P., & Werker, J. F. (2015). Sensorimotor influences on speech perception in infancy. *Proceedings of the National Academy of Sciences Proc Natl Acad Sci USA, 112*(44), 13531-13536. doi:10.1073/pnas.1508631112

Buchan, J. N., & Munhall, K. G. (2011). The influence of selective attention to auditory and visual speech on the integration of audiovisual speech information. *Perception, 40*(10), 1164-1182. doi:10.1068/p6939

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. doi:10.1002/dev.20032

Burr, D. & Alais, D. (2006). Combining visual and auditory information. In S. Martinez-Conde, S. L. Macknik, L. M. Martinez, J.-M. Alonso, & P. U. Tse (Eds.), *Progress in Brain Research: Visual Perception, Part 2: Fundamentals of Awareness: Multi-Sensory Integration and High-Order Perception* (Vol. 155, Part B, pp. 243-258). Amsterdam: Elsevier.

Bushara, K. O., Grafman, J., & Hallett, M. (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. *Journal of Neuroscience, 21*(1), 300-304.

Campbell, R. (2008). The processing of audio-visual speech: Empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 1001-1010. doi:10.1098/rstb.2007.2155

Cassia, V. M., Turati, C., & Simion, F. (2004). Can a nonspecific bias toward top-heavy patterns explain newborns' face preference? *Psychological Science, 15*(6), 379-383. doi:10.1111/j.0956-7976.2004.00688.x

Cohen, J., MacWhinney, B., Flatt, M., & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using Macintosh computers. *Behavior Research Methods, Instruments, & Computers, 25*(2), 257-271. doi:10.3758/bf03204507

Corey, D. P., & Hudspeth, A. J. (1979). Response latency of vertebrate hair cells. *Biophysical Journal, 26*(3), 499-506. doi:10.1016/s0006-3495(79)85267-4

Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants? *Developmental Psychobiology, 45*(4), 187-203. doi:10.1002/dev.20033

Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception, 9*(6), 719-721. doi:10.1068/p090719

Dodd, B. (1979). Lip reading in infants: Attention to speech presented in- and out-of-synchrony. *Cognitive Psychology, 11*(4), 478-484. doi:10.1016/0010-0285(79)90021-5

Doupe, A. J., & Kuhl, P. K. (1999). Birdsong and human speech: Common themes and mechanisms. *Annual Review of Neuroscience, 22*(1), 567-631. doi:10.1146/annurev.neuro.22.1.567

Dupont, S., Aubin, J., & Ménard, L. (2005). A study of the McGurk effect in 4 and 5-year-old French Canadian children. *ZAS Papers in Linguistics, 40*, 1-17.

Ernst, M.O. & Banks, M.S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429-433.

Finocchietti, S., Cappagli, G., & Gori, M. (2015). Encoding audio motion: Spatial impairment in early blind individuals. *Frontiers in Psychology, 6*, 1357. doi:10.3389/fpsyg.2015.01357

Fogel, A., Toda, S., & Kawai, M. (1988). Mother-infant face-to-face interaction in Japan and the United States: A laboratory comparison using 3-month-old infants. *Developmental Psychology, 24*, 398-406.

Friederici, A. D., & Wartenburger, I. (2010). Language and brain. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*, 150-109. doi:10.1002/wcs.9

Gebhard, J. W., & Mowbray, G. H. (1959). On discriminating the rate of visual flicker and auditory flutter. *The American Journal of Psychology, 72*(4), 521-529. doi:10.2307/1419493

Ghazanfar, A. A., & Takahashi, D. Y. (2014). The evolution of speech: Vision, rhythm, cooperation. *Trends in Cognitive Sciences, 18*(10), 543-553. doi:10.1016/j.tics.2014.06.004

Goren, C. C., Sarty, M., & Wu, P. Y. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics, 56*(4), 544-549.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds “L” and “R”. *Neuropsychologia, 9*(3), 317-323. doi:10.1016/0028-3932(71)90027-3

Grant, K. W., & Seitz, P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108*(3), 1197. doi:10.1121/1.1288668

Grant, K. W., & Seitz, P. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *The Journal of the Acoustical Society of America, 108*(3), 1197. doi:10.1121/1.1288668

Guellaï, B., Streri, A., & Yeung, H. H. (2014). The development of sensorimotor influences in the audiovisual speech domain: Some critical questions. *Frontiers in Psychology, 5*, 812. doi:10.3389/fpsyg.2014.00812

Haith, M., Bergman, T., & Moore, M. (1977). Eye contact and face scanning in early infancy. *Science, 198*(4319), 853-855. doi:10.1126/science.918670

Hall, D. G. (1991). *Perceptual and Associative Learning*. Oxford: Clarendon Press.

Hockley, N. S., & Polka, L. (1994). A developmental study of audiovisual speech perception using the McGurk paradigm. *The Journal of the Acoustical Society of America, 96*(5), 3309. doi:10.1121/1.410782

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. doi:10.1111/j.1467-8624.2005.00866.x

Howard, I. P., & Templeton, W. B. (1966). *Human spatial orientation*. London: Wiley.

Hunnius, S., & Geuze, R. H. (2004). Developmental changes in visual scanning of dynamic faces and abstract stimuli in infants: A longitudinal study. *Infancy, 6*(2), 231-255. doi:10.1207/s15327078in0602_5

Johnson, M. H., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition, 40*(1-2), 1-19. doi:10.1016/0010-0277(91)90045-6

Kamachi, M., Hill, H., Lander, K., & Vatikiotis-Bateson, E. (2003). Putting the face to the voice. *Current Biology, 13*(19), 1709-1714. doi:10.1016/j.cub.2003.09.005

Kärtner, J., Keller, H., & Yovsi, R. D. (2010). Mother-infant interaction during the first 3 months: The emergence of culture-specific contingency patterns. *Child Development*, *81*(2), 540-554. doi:10.1111/j.1467-8624.2009.01414.x

Keil, J., Muller, N., Ihssen, N., & Weisz, N. (2011). On the variability of the McGurk Effect: Audiovisual integration depends on prestimulus brain states. *Cerebral Cortex, 22*(1), 221-231. doi:10.1093/cercor/bhr125

Kim, R. S., Seitz, A. R., & Shams, L. (2008). Benefits of stimulus congruency for multisensory facilitation of visual learning. *PLoS ONE, 3*(1). doi:10.1371/journal.pone.0001532

King, A. J., & Palmer, A. R. (1985). Integration of visual and auditory information in bimodal neurones in the guinea-pig superior colliculus. *Experimental Brain Research, 60*(3), 492-500. doi:10.1007/bf00236934

Kubicek, C., Boisferon, A. H., Dupierrix, E., L Venbruck, 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. doi:10.1177/0165025412473016

Kubicek, C., Boisferon, A. H., Dupierrix, E., Pascalis, O., Lœvenbruck, H., Gervain, J., & Schwarzer, G. (2014). Cross-modal matching of audio-visual German and French fluent speech in infancy. *PLoS ONE, 9*(2). doi:10.1371/journal.pone.0089275

Kuhl, P. K., Conboy, B. T., Padden, D., Nelson, T., & Pruitt, J. Early speech perception and later languge development: Implications for the 'critical period'. *Language Learning and Development*, *1*(3-4), 237-264.

Kuhl, P. K., & Meltzoff, A. (1982). The bimodal perception of speech in infancy. *Science, 218*(4577), 1138-1141. doi:10.1126/science.7146899

Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development, 7*(3), 361-381. doi:10.1016/s0163-6383(84)80050-8

Kuhl, P. K., Tsao, F., Liu, H., Zhang, Y., & Boer, B. (2006). Language/Culture/Mind/Brain. *Annals of the New York Academy of Sciences, 935*(1), 136-174. doi:10.1111/j.1749-6632.2001.tb03478.x

Kuhl, P. K. (2010). Brain mechanisms in early language acquisition. *Neuron, 67*(5), 713-727.

doi:10.1016/j.neuron.2010.08.038

Kurjak, A., Stanojevic, M., Azumendi, G., & Carrera, J. M. (2005). The potential of four-dimensional (4D) ultrasonography in the assessment of fetal awareness. *Journal of Perinatal Medicine, 33*(1), 46-53. doi:10.1515/jpm.2005.008

Kushnerenko, E., Tomalski, P., Ballieux, H., Ribeiro, H., Potton, A., Axelsson, E. L., . . . Moore, D. G. (2013). Brain responses to audiovisual speech mismatch in infants are associated with individual differences in looking behaviour. *European Journal of Neuroscience, 38*(9), 3363-3369. doi:10.1111/ejn.12317

Lamb, T. D., & Pugh, E. N. (1992). A quantitative account of the activation steps involved in phototransduction in amphibian photoreceptors. *The Journal of Physiology, 449*(1), 719-758. doi:10.1113/jphysiol.1992.sp019111

Latto, A. J., Sato, M., & Diehl, R. L. (2004). Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. In J. Slifka, S. Manuel, & M. Matthies (Eds.), *From sound to sense: 50 years of discoveries in speech communication*. Cambridge, MA: MIT Press.

Lawrence, D. H. (1949). Acquired distinctiveness of cues: Transfer between discriminations on the basis of familiarity with the stimulus. *Journal of Experimental Psychology, 39*(6), 770-784. doi:10.1037/h0058097

LeVine, R. A., Dixon, S., LeVine, S., Richman, A., Leiderman, P. H., Keefer, C. H., & Brazelton, T. B. (1996). *Child Care and Culture: Lessons from Africa*. Cambridge: Cambridge University Press.

Lennie, P. (1981). The physiological basis of variations in visual latency. *Vision Research, 21*(6), 815-824. doi:10.1016/0042-6989(81)90180-2

Lewkowicz, D. J. (1986). Developmental changes in infants' bisensory response to synchronous durations. *Infant Behavior and Development, 9*(3), 335-353. doi:10.1016/0163-6383(86)90008-1

Lewkowicz, D. J. (1992a). Infants’ responsiveness to the auditory and visual attributes of a sounding/moving stimulus. *Perception & Psychophysics, 52*(5), 519-528. doi:10.3758/bf03206713

Lewkowicz, D. J. (1992b). Infants' response to temporally based intersensory equivalence: The effect of synchronous sounds on visual preferences for moving stimuli. *Infant Behavior and Development, 15*(3), 297-324. doi:10.1016/0163-6383(92)80002-c

Lewkowicz, D. J. (1996a). The temporal basis of multimodal integration. *Infant Behavior and Development, 19*, 158. doi:10.1016/s0163-6383(96)90213-1

Lewkowicz, D. J. (1996b). Infants' response to the audible and visible properties of the human face: Role of lexical-syntactic content, temporal synchrony, gender, and manner of speech. *Developmental Psychology, 32*(2), 347-366. doi:10.1037/0012-1649.32.2.347

Lewkowicz, D. J. (1998). Infants' response to the audible and visible properties of the human face: Discrimination of differences between singing and adult-directed speech. *Developmental Psychobiology, 32*(4), 261-274.

Lewkowicz, D. J. (2000). Infants' perception of the audible, visible, and bimodal attributes of multimodal syllables. *Child Development, 71*(5), 1241-1257. doi:10.1111/1467-8624.00226

Lewkowicz, D. J. (2003). Learning and discrimination of audiovisual events in human infants:

The hierarchical relation between intersensory temporal synchrony and rhythmic pattern cues. *Developmental Psychology, 39*(5), 795-804. doi:10.1037/0012-1649.39.5.795

Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology*, *46*(1), 66-67. doi:10.1037/a0015579

Lewkowicz, D. J., & Ghazanfar, A. A. (2006). The decline of cross-species intersensory perception in human infants. *Proceedings of the National Academy of Sciences of the United States of America, 103*(17), 6771-6774. doi:10.1073/pnas.0602027103

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. doi:10.1073/pnas.1114783109

Lewkowicz, D. J., Leo, I., & Simion, F. (2010). Intersensory Perception at Birth: Newborns Match Nonhuman Primate Faces and Voices. *Infancy, 15*(1), 46-60. doi:10.1111/j.1532-7078.2009.00005.x

Lewkowicz, D. J., & Turkewitz, G. (1980). Cross-modal equivalence in early infancy: Auditory-visual intensity matching. *Developmental Psychology, 16*(6), 597-607. doi:10.1037/0012-1649.16.6.597

MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science, 219*(4590), 1347-1349. doi:10.1126/science.6828865

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21*(2), 131-141.

doi:10.3109/03005368709077786

Massaro, D. W. (1984). Children's perception of visual and auditory speech. *Child Development, 55*(5), 1777. doi:10.2307/1129925

Massaro, D. W., Thompson, L. A., Barron, B., & Laren, E. (1986). Developmental changes in visual and auditory contributions to speech perception. *Journal of Experimental Child Psychology, 41*(1), 93-113. doi:10.1016/0022-0965(86)90053-6

Massaro, D. W., Cohen, M. M., & Smeele, P. M. (1996). Perception of asynchronous and conflicting visual and auditory speech. *The Journal of the Acoustical Society of America, 100*(3), 1777. doi:10.1121/1.417342

Mattock, K., & Burnham, D. (2006). Chinese and English Infants' Tone Perception: Evidence for Perceptual Reorganization. *Infancy, 10*(3), 241-265. doi:10.1207/s15327078in1003_3

Maurer, D., & Werker, J. F. (2014). Perceptual narrowing during infancy: A comparison of language and faces. *Developmental Psychobiology, 56*(2), 154-178. doi:10.1002/dev.21177

Maye, J., & Gerken, L. (2000). Learning phonemes without minimal pairs. In S. C. Howell, S. A. Fish, & T. Keith-Lucas (Eds.), *Proceedings of the 24th Boston University Conference on Language Development* (pp. 522-533). Somerville, MA: Cascadilla Press.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101-111. doi:10.1016/s0010-0277(01)00157-3

Maye, J., Weiss, D. J., & Aslin, R. N. (2008). Statistical phonetic learning in infants: Facilitation and feature generalization. *Developmental Science, 11*(1), 122-134. doi:10.1111/j.1467-

7687.2007.00653.x

McGurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748. doi:10.1038/264746a0

McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *The Journal of the Acoustical Society of America, 77*(2), 678. doi:10.1121/1.392336

Ménard, L., Dupont, S., Baum, S. R., & Aubin, J. (2009). Production and perception of French vowels by congenitally blind adults and sighted adults. *Journal of the Acoustical Society of America*, *126*, 1406-1414.

Merin, N., Young, G. S., Ozonoff, S., & Rogers, S. J. (2006). Visual fixation patterns during reciprocal social interaction distinguish a subgroup of 6-month-old infants at-risk for autism from comparison infants. *Journal of Autism and Developmental Disorders, 37*(1), 108-121. doi:10.1007/s10803-006-0342-4

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics, 18*(5), 331-340. doi:10.3758/bf03211209

Molfese, D. (2000). Predicting dyslexia at 8 years of age using neonatal brain responses. *Brain and Language*, *72*, 238-245.

Mugitani, R., Kobayashi, T., & Hiraki, K. (2008). Audiovisual matching of lips and non-canonical sounds in 8-month-old infants. *Infant Behavior and Development, 31*(2), 307-310. doi:10.1016/j.infbeh.2007.12.002

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & Psychophysics, 58*(3), 351-362. doi:10.3758/bf03206811

Murray, M. M., Michel, C. M., Peralta, R. G., Ortigue, S., Brunet, D., Andino, S. G., & Schnider, A. (2004). Rapid discrimination of visual and multisensory memories revealed by electrical neuroimaging. *NeuroImage, 21*(1), 125-135. doi:10.1016/j.neuroimage.2003.09.035

Nagy, E. (2008). Innate intersubjectivity: Newborns' sensitivity to communication disturbance. *Developmental Psychology, 44*(6), 1779-1784. doi:10.1037/a0012665

Narayan, C. R., Werker, J. F., & Beddor, P. S. (2010). The interaction between acoustic salience and language experience in developmental speech perception: Evidence from nasal place discrimination. *Developmental Science, 13*(3), 407-420. doi:10.1111/j.1467-7687.2009.00898.x

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration. *Cognitive Brain Research, 25*(2), 499-507. doi:10.1016/j.cogbrainres.2005.07.009

Navarra, J., & Soto-Faraco, S. (2005). Hearing lips in a second language: Visual articulatory information enables the perception of second language sounds. *Psychological Research, 71*(1), 4-12. doi:10.1007/s00426-005-0031-5

Norcross, K. J., & Spiker, C. C. (1957). The effects of type of stimulus pretraining on discrimination performance in preschool children. *Child Development, 28*(1), 79-84. doi:10.1111/j.1467-8624.1957.tb04833.x

Ochs, E., & Schieffelin, B. (1984). *Culture Theory: Mind, Self, and Emotion* (pp. 276-320) (R. Shweder & R. LeVine, Eds.). Cambridge: Cambridge University Press.

Oller, D., Eilers, R. E., Neal, A., & Schwartz, H. K. (1999). Precursors to speech in infancy. *Journal of Communication Disorders, 32*(4), 223-245. doi:10.1016/s0021-9924(99)00013-1

Palmer, S. B., Fais, L., Golinkoff, R. M., & Werker, J. F. (2012). Perceptual narrowing of linguistic sign occurs in the 1st year of life. *Child Development*, *83*(2), 543-553. doi:10.1111/j.1467-8624.2011.01715.x

Pandey, P. C., Kunov, H., & Abel, S. M. (1986). Disruptive effects of auditory signal delay on speech perception with lipreading. *Journal of Audiology Research, 26*(1), 27-41.

Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development, 22*(2), 237-247. doi:10.1016/s0163-6383(99)00003-x

Patterson, M. L., & Werker, J. F. (2002). Infants' ability to match dynamic phonetic and gender information in the face and voice. *Journal of Experimental Child Psychology, 81*(1), 93-115. doi:10.1006/jecp.2001.2644

Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science, 6*(2), 191-196. doi:10.1111/1467-7687.00271

Polka, L., & Werker, J. F. (1994). Developmental changes in perception of nonnative vowel contrasts. *Journal of Experimental Psychology: Human Perception and Performance, 20*(2), 421-435. doi:10.1037/0096-1523.20.2.421

Polka, L., Colantonio, C., & Sundara, M. (2001). A cross-language comparison of /d/–/ð/

perception: Evidence for a new developmental pattern. *The Journal of the Acoustical Society of America, 109*(5), 2190-2201. doi:10.1121/1.1362689

Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastian-Galles, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences of the United States of America, 106*(26), 10598-10602. doi:10.1073/pnas.0904134106

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review, 83*(2), 157-171. doi:10.1037/0033-295x.83.2.157

Reese, H. W. (1972). Acquired distinctiveness and equivalence of cues in young children. *Journal of Experimental Child Psychology, 13*(1), 171-182. doi:10.1016/0022-0965(72)90017-3

Remez, R. E. (2005). Three puzzles of multimodal speech perception. In E. Vatikiotis-Bateson, G. Bailly, & P. Perrier (Eds.), *Audiovisual Speech* (pp. 12-19). Cambridge, MA: MIT Press.

Remez, R., Rubin, P., Pisoni, D., & Carrell, T. (1981). Speech perception without traditional speech cues. *Science, 212*(4497), 947-949. doi:10.1126/science.7233191

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 10, pp. 243-335). New York: Academic Press.

Richman, A. L., Miller, P. M., & Levine, R. A. (1992). Cultural and educational variations in maternal responsiveness. *Developmental Psychology, 28*(4), 614-621. doi:10.1037/0012-

1649.28.4.614

Rosenblum, L. D. (2005). Primacy of multimodal speech perception. In D. B. Pisoni & R. E. Remez (Eds.), *The Handbook of Speech Perception* (pp. 51-78). Malden, MA: Blackwell.

Rosenblum, L. D., & Saldaña, H. M. (1992). Discrimination tests of visually influenced syllables. *Perception & Psychophysics, 52*(4), 461-473. doi:10.3758/bf03206706

Rosenblum, L. D., & Saldaña, H. M. (1996). An audiovisual test of kinematic primitives for visual speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 22*(2), 318-331. doi:10.1037//0096-1523.22.2.318

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics, 59*(3), 347-357. doi:10.3758/bf03211902

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2006). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex, 17*(5), 1147-1153. doi:10.1093/cercor/bhl024

Röder, B., Teder-Sälejärvi, W., Sterr, A., Rösler, F., Hillyard, S. A., & Neville, H. J. (1999). Improved auditory spatial tuning in blind humans. *Nature, 400*, 162-166.

Scheier, C., Lewkowicz, D. J., & Shimojo, S. (2003). Sound induces perceptual reorganization of an ambiguous motion display in human infants. *Developmental Science, 6*(3), 233-241. doi:10.1111/1467-7687.00276

Schorr, E. A., Fox, N. A., van Wassenhove, V., & Knudsen, E. I. (2005). Auditory-visual fusion in speech perception in children with cochlear implants. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 18748-18750.

Sebastian-Galles, N., Albareda-Castellot, B., Weikum, W. M., & Werker, J. F. (2012). A Bilingual Advantage in Visual Language Discrimination in Infancy. *Psychological Science, 23*(9), 994-999. doi:10.1177/0956797612436817

Sekiyama, K., & Burnham, D. (2008). Impact of language on development of auditory-visual speech perception. *Developmental Science, 11*(2), 306-320. doi:10.1111/j.1467-7687.2008.00677.x

Setti, A., Burke, K. E., Kenny, R., & Newell, F. N. (2013). Susceptibility to a multisensory speech illusion in older persons is driven by perceptual processes. *Frontiers in Psychology, 4*, 575. doi:10.3389/fpsyg.2013.00575

Shimojo, S., & Shams, L. (2001). Sensory modalities are not separate modalities: Plasticity and interactions. *Current Opinion in Neurobiology, 11*(4), 505-509. doi:10.1016/s0959-4388(00)00241-5

Shipley, T. (1964). Auditory flutter-driving of visual flicker. *Science, 145*(3638), 1328-1330. doi:10.1126/science.145.3638.1328

Soto-Faraco, S., Navarra, J., Weikum, W., Vouloumanos, A., Sebastian-Galles, N., & Werker, J.F. (2007). Discriminating languages by speech-reading. *Perception and Psychophysics*, *69*(2), 218-231.

Spector, F. & Maurer, D. (2009). Synesthesia: A new approach to understanding the development of perception. *Developmental Psychology*, *45*(1), 175-189.

Stein, B. E. (1998). Neural mechanisms for synthesizing sensory information and producing adaptive behaviors. *Experimental Brain Research, 123*(1-2), 124-135. doi:10.1007/s002210050553

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America, 26*(2), 212. doi:10.1121/1.1907309

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society B: Biological Sciences, 335*(1273), 71-78. doi:10.1098/rstb.1992.0009

Sundara, M., Polka, L., & Genesee, F. (2006). Language-experience facilitates discrimination of /d-/ in monolingual and bilingual acquisition of English. *Cognition, 100*(2), 369-388. doi:10.1016/j.cognition.2005.04.007

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. doi:10.1016/j.cognition.2008.05.009

Ter Schure, S., Junge, C., & Boersma, P. (2016). Discriminating non-native vowels on the basis of multimodal, auditory or visual information: Effects on infants' looking patterns and discrimination. *Frontiers in Psychology, 7*, 525. doi:10.3389/fpsyg.2016.00525

Tomalski, P., Ribiero, H., Ballieux, H., Axelsson, E. L., Murphy, E., Moore, D. G., & Kushnerenko, E. (2013). Exploring early developmental changes in face scanning patterns during the perception of audiovisual mismatch of speech cues. *European Journal of Developmental Psychology, 10*(5), 611-624. doi:10.1080/17405629.2012.728076

Trehub, S. E. (1976). The discrimination of foreign speech contrasts by infants and adults. *Child Development, 47*(2), 466-472. doi:10.2307/1128803

Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., & Théoret, H. (2007). Speech and non-speech audio-visual illusions: A developmental study. *PLoS ONE, 2*(8). doi:10.1371/journal.pone.0000742

Tsao, F. M., Liu, H. M., & Kuhl, P. K. (2004). Speech perception in infancy predicts later language development in the second year of life: A longitudinal study. *Child Development*, *75*, 1067-1084.

Tsao, F. M., Liu, H. M., & Kuhl, P. K. (2006). Perception of native and non-native affricate-fricative contrasts: Cross-language tests on adults and infants. *The Journal of the Acoustical Society of America, 120*(4), 2285. doi:10.1121/1.2338290

Tsushima, T., Takizawa, O., Sasaki, M., Shiraki, S., Nishi, K., Kohno,, M., . . . Best, C. (1994). Discrimination of English /r-l/ and /w-y/ by Japanese infants at 6–12 months: Language-specific developmental changes in speech perception abilities. *1994 International Conference on Spoken Language Processing,* 1695-1698.

Valenza, E., Simion, F., Cassia, V. M., & Umiltà, C. (1996). Face preference at birth. *Journal of Experimental Psychology: Human Perception and Performance, 22*(4), 892-903. doi:10.1037/0096-1523.22.4.892

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 of the United States of America, 102*(4), 1181-1186. doi:10.1073/pnas.0408949102

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia, 45*(3), 598-607. doi:10.1016/j.neuropsychologia.2006.01.001

Vatikiotis-Bateson, E., & Munhall, K. G. (2015). Auditory-visual speech processing: Something doesn't add up. In M. A. Redford (Ed.), *The Handbook of Speech Production* (1st ed., pp. 178-199). New York: Wiley-Blackwell.

Vatikiotis-Bateson, E., Eigsti, I., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics, 60*(6), 926-940. doi:10.3758/bf03211929

Vouloumanos, A., Druhen, M. J., Hauser, M. D., & Huizink, A. T. (2009). Five-month-old infants' identification of the sources of vocalizations. *Proceedings of the National Academy of Sciences of the United States of America, 106*(44), 18867-18872. doi:10.1073/pnas.0906049106

Walker, J. T., & Scott, K. J. (1981). Auditory-visual conflicts in the perceived duration of lights, tones, and gaps. *Journal of Experimental Psychology: Human Perception and Performance, 7*(6), 1327-1339. doi:10.1037/0096-1523.7.6.1327

Walton, G. E., & Bower, T. (1993). Amodal representation of speech in infants. *Infant Behavior and Development, 16*(2), 233-243. doi:10.1016/0163-6383(93)80019-5

Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastian-Galles, N., & Werker, J. F. (2007). Visual Language Discrimination in Infancy. *Science, 316*(5828), 1159-1159. doi:10.1126/science.1137686

Weikum, W. M., Oberlander, T. F., Hensch, T. K., & Werker, J. F. (2012). Prenatal exposure to antidepressants and depressed maternal mood alter trajectory of infant speech perception. *Proceedings of the National Academy of Sciences of the United States of America, 109*, 17221-17227. doi:10.1073/pnas.1121263109

Welch, R. B., DuttonHurt, L. D., & Warren, D. H. (1986). Contributions of audition and vision to temporal rate perception. *Perception & Psychophysics, 39*(4), 294-300. doi:10.3758/bf03204939

Werker, J. F., Gilbert, J. H., Humphrey, K., & Tees, R. C. (1981). Developmental Aspects of Cross-Language Speech Perception. *Child Development, 52*(1), 349. doi:10.2307/1129249

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. doi:10.1016/s0163-6383(84)80022-3

Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology, 24*(5), 672-683. doi:10.1037/0012-1649.24.5.672

Werker, J. F., Frost, P. E., & Mcguirk, H. (1992). La langue et les lèvres: Cross-language influences on bimodal speech perception. *Canadian Journal of Psychology/Revue Canadienne De Psychologie, 46*(4), 551-568. doi:10.1037/h0084331

Werker, J. F., & Tees, R. C. (2005). Speech perception as a window for understanding plasticity and commitment in language systems of the brain. *Developmental Psychobiology, 46*(3), 233-251. doi:10.1002/dev.20060

Werker, J. F., & Hensch, T. K. (2015). Critical Periods in Speech Perception: New Directions. *Annual Review of Psychology, 66*(1), 173-196. doi:10.1146/annurev-psych-010814-015104

Yeung, H. H., & Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-Month-olds use distinct objects as cues to categorize speech information.

Cognition, 113(2), 234-243. doi:10.1016/j.cognition.2009.08.010

Yeung, H. H., Chen, K. H., & Werker, J. F. (2013). When does native language input affect phonetic perception? The precocious case of lexical tone. *Journal of Memory and Language, 68*(2), 123-139. doi:10.1016/j.jml.2012.09.004

Yeung, H. H., Chen, L. M., & Werker, J. F. (2013). Referential Labeling Can Facilitate Phonetic Learning in Infancy. *Child Development, 85*(3), 1036-1049. doi:10.1111/cdev.12185

Young, G. S., Merin, N., Rogers, S. J., & Ozonoff, S. (2009). Gaze behavior and affect at 6 months: Predicting clinical outcomes and language development in typically developing infants and infants at risk for autism. *Developmental Science, 12*(5), 798-814. doi:10.1111/j.1467-7687.2009.00833.x

Appendices

Appendix A: 6-month-olds' auditory discrimination of non-native Hindi phones

At six months of age, infants discriminate between similar consonant phonemes from non-native languages, a sensitivity that declines shortly thereafter (Werker & Tees, 1984, *inter alia*). A series of studies in my laboratory, including those presented in this thesis, rely on Hindi dental and retroflex consonants, which, in general, six-month-olds discriminate auditorily. However, in order to conduct the current studies, new audiovisual stimuli were engineered. Prior to using these stimuli in new studies, we first replicated prior findings to ensure that six-month-olds indeed discriminated our new stimuli auditorily.

The following analyses are modified from an article of which I was second author (Bruderer, Danielson, Kandhadai, & Werker, 2015):

Six-month-old English-learning infants were tested in a standard alternating/non-alternating procedure in which looking time to a checkerboard was the dependent variable (Best & Jones, 1998). Infants were presented with two types of trials: those in which tokens from the dental and retroflex phonetic categories alternated in presentation (alternating trials, 'Alt'), and those in which tokens from the same speech sound category were repeated for the duration of the trial (non-alternating trials, 'NAlt'). In this design, significantly longer looking time to the Alt over the NAlt trials is taken as evidence that infants discriminate between these two sound categories.

Twenty-four English-learning six-month-old infants participated in this experiment. Looking time data were analyzed across the 4 trials of each type (4 Alt, 4 NAlt). Following Yeung and Werker (2009), looking time data were analyzed in pairs of trials, where each pair contained one Alt and one NAlt trial: pair one included the first two trials of the study, pair two included the third and fourth trials, and so on; this allowed us to account for any changes in looking time across the series of trials. A 2 (Trial Type) X 4 (Pair) repeated-measures ANOVA was performed on the looking times, using the within-subjects factors of Trial Type (Alt or NAlt) and Pair (1st, 2nd, 3rd, or 4th). The main effect of Pair was significant, *F*(3,69) = 9.84, *p* < .001, η_p^2 = .30, indicating that infants' looking time significantly declined across the four pairs of trials, as is standard in familiarization or habituation looking time paradigms. Further, there was a significant main effect of Trial Type, *F*(1,23) = 4.32, *p* = .049, η_p^2 = .16, and no interaction with Pair, *F*(3,69) = 1.63, *p* = .19, η_p^2 = .066, suggesting that the difference in looking time between the two types of trials did not significantly differ across the four pairs. Follow-up investigation of looking time means for the significant Trial Type effect showed that infants looked longer during Alt trials (*M* = 9369.17 ms, *SD* = 4033.06) than NAlt trials (*M* = 8542.29 ms, *SD* = 4053.58).

The results from Experiment 1 replicate previous findings showing that six-month-old English learning infants are able to discriminate the non-native Hindi /d̪/-/ɖ/ contrast.