Development of Audiovisual Integration in Human Infants:

The Effects of Spatial and Temporal Congruency and Incongruency on Response Latencies

Thesis by

Patricia A. Neil

In Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy

California Institute of Technology
Pasadena, California

2006
(Defended May 3, 2006)

The human brains starts working the moment you are born
and never stops until you stand up to speak in public.
– *George Jessel*

Acknowledgements

I would first like to acknowledge my debt to my advisor and mentor, Shin Shimojo, for his support and guidance... and for literally providing his first-born son as a test subject.

I would also like to thank David J. Lewkowicz for his invaluable advise and experience in infant perception, as well as the helpful contributions to this research from Joydeep Bhattacharya, Christine Chee-Ruiter, Christian Scheier, Rui Wang, Misty Richards, Ben Matthews, Alice Lin, Susan Dao, Elinor Lin, Ewen Chao, and Chika Arakawa.

This work would not have been possible without the unwavering love, support, and sympathy of my friends, including but not limited to: Rebecca Adler, Susan Ayer, Karen Bernhardt, Sharlotte Bolyard, Titus Brown, Don and Wendy Caldwell, Athena Castro, Mary Dunlop, Anne and Patrick Earhart, Jeffrey Edlund, Jessica Edwards, Ken Fisher, Greg Fletcher, Nathan and Natale Gray, Theresa Grieco, Freddy Hansen, Fumiko Maeda Hoeft, Katie Homann, Martin Huber, Desiree Lavertu, Kate Magary, Myra Martino, Virgie Mateo, Eva Murdock, Dylan Nieman, Julie Poposki, Kerstin Preuschoff, Lara Pruitt, Lorian Churchill Schaeffer, Eiko Shimojo, Tracy Teal, Ann Marie Polsenberg Thomas, Angela Tooker, Scott Van Essen, Tomtor Varutbangkul, Dirk Walthers, June Wicks, Heather Wiencko, Kjerstin Easton Williams, Donna Wintzen, Nathan Wozny, Daw-an Wu, Nikki Young, and my family: Mike, Brenda, Jackie, Andy, Naiomi, Becky, Travis, Nancy, Don, Johnny, Betsey, and, of course, Mom and Dad. And if most of them cannot remember what I have been doing for the past six years besides "making babies cry", I won't name names.

Abstract

Every day we are inundated with a mass of sensory inputs providing a continual stream of relevant and irrelevant, redundant and conflicting, information about the external world. Mature brains are very capable in integrating this confusion of input into a unified percept, but this is a non-trivial task for infants, whose brains and sensory systems are still immature at birth and who rely on their current level of integration and interaction of these inputs in order to shape their future development. Failure in being able to properly process basic sensory interactions has been implicated in higher-level developmental problems like attentional or autistic spectrum disorders. Numerous studies have looked at how adults perceive and react to multisensory stimuli, including findings of improved response latencies and target detection for spatially and temporally congruent stimuli, but much less is known about the development of multisensory integration or how spatial or temporal disparities effect sensory interactions in young babies. We examined the role of spatial and temporal congruency and incongruency on the response latencies of infants under ten months of age orienting toward an audiovisual stimulus at ±25° and/or ±45°. In Study 1, we found the beginnings of adult-style non-linear integration for spatially and temporally congruent audiovisual targets in 8–10 month olds, but not in younger infants, as well as indications of a differential developmental profile for binaural versus monaural processing. In Studies 2 and 3, spatial and temporal disparities were found to significantly lengthen infants' response latencies to an audiovisual target. We also found clear indications of developmental changes for all three spatial and temporal conditions, as well as key dependencies in relative position, temporal order, and sensory dominance.

Contents

List of Tables

List of Figures

Chapter 1

Multisensory Integration and Development

1.1 Multisensory Interactions

Every day we are inundated with a mass of sensory inputs providing a continual stream of relevant and irrelevant, redundant and conflicting, information about the external world. The adult brain is able to pull together reliably the appropriate multisensory stimuli in order to gain a unified, veridical percept of the external world. One of the more complex problems in the study of multisensory interactions is understanding how the developing brain—the individual sensory systems and higher cortical areas being still immature at birth—is able to make sense of this overabundance of sensory stimulation and reach an adult level of perception and capability.

1.1.1 Perception, Behavior, and Physiology

Multisensory interactions are not limited to cases of feature binding for object representation, but cover a wide range of perceptual, behavioral, and neurophysiological responses. Some multisensory interactions are capable of modulating or outright changing the perceived property of one stimulus through the presence of a second stimulus in another modality, with such changes as perceptibility (detectability or salience) (McDonald, Teder-Salejarvi, & Hillyard, 2000), discriminability (Kennett, Taylor-Clarke, & Haggard, 2001), or an increase in subjective stimulus brightness

(Stein & Wallace, 1996) or duration (Vroomen & Gelder, 2000). Everyone has had the experience of trying to engage in a conversation in a crowded and noisy room. In this situation, the auditory system is bombarded with input from any number of distinct conversations of varying intensity and perceptibility. By looking at the face and lips of the speaker to whom you wish to listen, their words are more easily distinguished from the background noise. One well-known variation of this obviously useful interaction is the McGurk effect (McGurk & Macdonald, 1976). When individuals are presented with certain conflicting phonemes and lip movements (e.g. the sound /ba/ combined with the discrepant lip movements of /ga/), the final percept is different from both (perceived as /da/). Another example of perception being altered by the presence of multiple modalities is the stream-bounce illusion. Two identical disks, starting in opposite corners at the top of a display screen, move diagonally down the screen and intersect in the center before continuing on to different corners at the bottom. By itself, this is a visually ambiguous stimulus; the disks can be perceived as either streaming past or bouncing against one another. However, the introduction of an acoustic event with a sharp rise and fall time around the moment of coincidence is able to strongly bias the perception to one of collision, while presenting the same auditory stimulus temporally displaced before or after the moment of intersection reverses the perceived bias toward a streaming event (Sekuler, Sekuler, & Lau, 1997). An auditory sound has even been found to generate an illusory visual perception: A single visual flash is perceived as a double (or more) flash when accompanied by double (or more) sounds (Shams, Kamitani, & Shimojo, 2000). Some interactions are even able to modify—either for the better or for the worse and without conscious awareness—the overt response to a multisensory event, such as improved visual search and event identification (Perrott, Saberi, Brown, & Strybel, 1990). Adults have long been found to show significantly faster response latencies and improved accuracy in localizing audiovisual targets compared to auditory-only or visual-only targets (Hughes, Reuter-Lorenz, Nozawa, & Fendrich, 1994), or to have their localization of an auditory target displaced in the direction of a visual distractor (Hairston et al., 2003). Multisensory interactions have also been found at the level of individual neurons. Two characteristic features of these multisensory neurons are an enhanced activation response to a synchronous, spatially co-located multimodal stimulus and a depressed response to a temporally or spatially dis-located stimulus. The maximal response from these neurons is frequently when the components of a multimodal stimulus are at lower intensity levels than are ideal for unimodal stimuli (Meredith & Stein, 1983, 1986; Wallace, Wilkinson, & Stein, 1996).

1.1.2 Temporal and Spatial Factors

One of the more puzzling questions regarding the study of multisensory phenomena is trying to understand what are the critical factors in integration. Two factors known to be crucial are temporal and spatial coincidence (or displacement), the contribution of both being highly task- and situation-dependent.

Perceptual judgments can be varied based upon the relative timing or location of the multimodal stimulus. For example, the ventriloquist effect—the illusion whereby a synchronously presented auditory stimulus (within 300 ms) appears to be located at the position of a physically displaced visual target—can also be modified such that large enough temporal disparities (150 to 250 ms) cause the erroneous perception that spatially aligned auditory and visual targets are displaced (Jack & Thurlow, 1973; Radeau & Bertelson, 1977; Slutsky & Recanzone, 2001). Whereas the standard ventriloquist effect is the shortening of the perceived distance between spatially disparate but synchronously presented visual and auditory stimuli, a temporal correlate—temporal ventriloquism—has recently been found such that the perception of the magnitude of asynchrony between simple lights and noise bursts (temporal order judgment) is reduced when the stimuli are in the same location versus displaced locations (Bertelson & Aschersleben, 2003). In the McGurk effect, the modified perception breaks down if the synchrony between the lip movements and the voice exceeds 200–300 ms (Massaro, Cohen, & Smeele, 1996; Munhall, Gribble, Sacco, & Ward, 1996). The previously described stream-bounce illusion percept can be flipped from the "stream" percept to that of the "bouncing" by introducing the auditory event close to the point of intersection instead of well before or after. The strength of the bounce percept is strongest when the auditory event occurs at or 150 ms prior to the visual intersection, and is weaker when it occurs 150 ms post intersection. This is just one example of how it is not just absolute temporal disparity, but relative disparity that seems to matter.

Although the spatial and temporal constraints may at first appear hopelessly complex in trying to predict the outcome of audiovisual integration for a given set of temporal and spatial parameters, two outcomes are generally found regarding which stimulus is biased and which does the biasing: (i) asymmetry of weighting (the ability of one modality to bias the perception or response toward a second) and (ii) relative primacy/intensity of each modality (in spatial determination tasks, the auditory component is often mis-localized in the direction of the visual component, and in temporal determination tasks, the auditory component often biases the visual component).

In a temporal order judgment of two visual targets (temporal determination), the presentation of spatially irrelevant auditory tones 75–250 ms after the visual onsets significantly improved performance while the same tone presented prior to the visual onsets had no effect. Inserting two (but not one) auditory tones (16 ms apart) in between the visual target onsets diminished performance (Morein-Zamir, Soto-Faraco, & Kingstone, 2003), showing how a temporal determination task can be biased by an auditory stimulus. Adults asked to localize an auditory target in space (spatial determination) were significantly biased in their accuracy towards the direction of a spatially-displaced visual distractor, even for very large disparities (Hairston et al., 2003), showing the modulating effect of vision in a spatial determination task. Although vision is generally capable of greater spatial acuity than audition, relative intensities (weighting) also play a role. Auditory signals of a particular gaussian temporal envelope that have little to no effect on the perceived location of a moving visual target with low positional uncertainty (a small gaussian blob), are able to exert a much greater effect as the positional uncertainty of the visual target is increased (a larger gaussian blob). Increasing the spread of the auditory signal's temporal envelope produces a corresponding reduction is it's ability to bias the visual target (Heron, Whitaker, & McGraw, 2004).

The variety of effects found when multiple modalities are involved—the majority of which have been largely studied in adults—immediately raises the question as to how such interactions develop from infancy.

1.1.3 Infants

Research to date has shown that infants possess a variety of multisensory perceptual abilities (D. J. Lewkowicz, 2000b, 2002). For example, it has been shown that infants can perform a variety of types of audiovisual integration, including intensity matching (D. J. Lewkowicz & Turkewitz, 1980), detecting synchrony relations (Dodd, 1979), and even perceiving illusions based on auditory and visual interactions (Scheier, Lewkowicz, & Shimojo, 2003). Research also has shown that some multisensory abilities differ across early development; the ability to make duration-based multisensory matches emerges by six months of age (D. J. Lewkowicz, 1986) but the ability to perceive synchrony relations embedded in a rhythmic pattern does not emerge until ten months of age (D. Lewkowicz, 2003). Developmental differences such as these are not surprising given the rapid changes in basic sensory/perceptual abilities that occur during the first year of life.

However, an examination across the breadth of infant developmental studies has revealed several puzzling mixed results. For example, whereas two different groups (Lyons-Ruth, 1977; Lawson, 1980) both found that young infants are sensitive to the spatial and temporal contiguity of auditory and visual stimuli, Lawson also reported that infants could not associate auditory and visual inputs solely on the basis of temporal contiguity. Yet, the opposite has also been reported. Several studies have shown that infants as young as 3.5 months of age do respond to the temporal contiguity of auditory and visual information, even when the auditory and visual inputs are not spatially contiguous (Bahrick, 1988; D. J. Lewkowicz, 1992b, 1996; Spelke, 1979). Is this inconsistency real or could the different findings be attributable to the specific values of temporal and spatial separation that were used in the different studies? This is a clear possibility as recent studies in adults have shown that the exact values of temporal and spatial discrepancies influence multisensory perception (Lewald & Guski, 2003).

The vast majority of crossmodal developmental studies have focused on either the temporal aspects of multisensory perception in infants or their ability to integrate the auditory and visual attributes of speech, affect, or shape, and whether infants can use various forms of temporal information to perceive intermodally unified events (D. J. Lewkowicz, 2000b). Lewkowicz identified four basic temporal characteristics utilized in crossmodal events: temporal contiguity, duration, rate, and rhythm (D. J. Lewkowicz, 1994). Temporal contiguity (or synchrony) has been the focus of many studies and there is much supporting evidence that it is a property to which young infants are already responsive (Bahrick, 1987; D. J. Lewkowicz, 1986, 1992a, 1992b, 1996; Spelke, 1988; Spelke, Born, & Chu, 1983). Temporal synchrony, being such a simple crossmodal relation, may be fundamental to the perception of crossmodal relations in general (Edelman, 1992; D. J. Lewkowicz, 2000a).

Only a handful of studies have investigated the development of spatial multisensory integration and little is known about infants' audiovisual spatial perception. This is somewhat surprising considering the importance of spatial cues in the previously mentioned studies. Indeed, in unpublished studies, Lewkowicz reported evidence suggesting that spatial cues are important even for temporal integration. The temporal asynchrony (350 ms) that infants were capable of detecting when audiovisual stimuli were spatially concordant was no longer detected when the stimuli were separated by 40°. Some other previous work has found that young infants are sensitive to the spatial co-location and temporal contiguity of auditory and visual inputs (Lyons-Ruth, 1977; Lawson, 1980) while other studies indicate that infants as young as 3.5 months of age are insensitive to the spatial separation of auditory and visual inputs when temporally coincident (D. J. Lewkowicz, 1992a; Spelke, 1979; Bahrick, 1988). However, none of these studies has systematically investigated spatial intersensory integration skills across a wide enough age range during infancy to capture possible developmental changes.

This has motivated our current examination into the development of auditory and visual integration in human infants under conditions of spatial and temporal congruency (Study 1), spatial disparity (Study 2), and temporal disparity (Study 3).

1.2 Spatial Localization

The location of objects and events in our environment is often specified by concurrent auditory and visual inputs. Adults of many species, including humans, take advantage of such multisensory redundancy in spatial localization. Detection, discrimination, and localization are often performed more quickly and more reliably when bimodal as opposed to unimodal cues are available (Miller, 1982). For example, the spatial localization of simple audiovisual targets is significantly faster than the localization of the same auditory or visual targets alone (Hughes et al., 1994). Despite the fact that adults profit from multimodal source specification when performing spatial localization tasks, it is not known when multisensory facilitation of localization behavior first emerges in human development. Ascertaining when it does is important because the ability to integrate multisensory inputs is critical to the development of a unified perceptual world and ultimately to the acquisition of veridical knowledge (Piaget, 1952; J. Gibson, 1979; E. J. Gibson, 1982; D. J. Lewkowicz, 2000a).

1.2.1 Orienting Response and Saccade Generation

Integrating visual and auditory spatial information requires the merging of *retinotopically* organized visual information with *tonotopically* organized auditory information, into a common reference frame for the production of a response signal (Stein & Meredith, 1993). Countless multisensory interaction studies utilize saccade measurements (response latency, saccade amplitude, peak velocity) as a means of understanding the behavioral response to an audiovisual event as this is a very innate and precise orienting system in primates. From birth infants orient their eyes and head toward novel visual and auditory stimuli (Fantz, 1963; Wertheimer, 1961). Voluntary saccades in adults serve to center the fovea on an event or object in visual space, and their generation is controlled through the superior colliculus, a midbrain structure containing spatially aligned maps of visual, auditory, and somatosensory inputs.

In single-cell recording studies in cats and macaques, multisensory neurons have been found to have distinct, but large and overlapping, receptive fields to auditory and visual stimuli. When spatially and temporally congruent auditory and visual targets fall within this overlapping region, the neurons respond with an enhanced firing rate, especially when the relative intensities of the targets are less than that which produces the maximal response alone (Meredith & Stein, 1996; Wallace et al., 1996). When the spatial disparity between the two stimuli is increased such that one falls outside it's receptive field for the neuron, the firing rate can even be depressed below the baseline firing rate (Stein, 1998). The maximal response in a multisensory neuron has also been found to coincide with the temporal alignment of the peak discharge periods corresponding to each unimodal response, with that response dropping off monotonically with increased misalignment (Meredith, Nemitz, & Stein, 1987). These neurons may communicate directly with saccade-related neurons in the intermediate and deep layers of the colliculus or in other areas of the brainstem (Meredith & Stein, 1985). A spatially and temporally aligned audiovisual event at a particular location in space might be initiating an enhanced response in populations of multisensory neurons with matching receptive fields, leading to a strong orienting signal to the saccade generation pathways. As spatial and/or temporal disparity increases between the constituent parts of the audiovisual event, fewer numbers of neurons in this population would produce an enhanced firing rate—or perhaps begin providing an inhibiting response via a depressed firing rate—resulting in a gradual reduction of the saccade generation signal and a corresponding reduction in the overt behavioral response. This expectation of a gradual reduction in enhanced behavior toward audiovisual events (response latency, accuracy, detection, etc) as spatial and/or temporal disparity increases has been found in several studies (Harrington & Peck, 1998; Frens, Vanopstal, & Vanderwilligen, 1995). Several developmental studies in both cats and monkeys suggest that these multisensory neurons, abundant in the adult, are either lacking or incapable of this integrative function in newborns (Wallace & Stein, 1997, 2001). In addition, there is some indirect, behavioral evidence for postnatal changes in the superior colliculus in humans (Harman, Posner, Rothbart, & Thomasthrapp, 1994). Finally, multisensory neurons of the superior colliculus receive projections from many cortical and subcortical regions (Wallace, Meredith, & Stein, 1993), some of which appear to mediate the multisensory integration observed in the superior colliculus (Jiang, Wallace, Jiang, Vaughan, & Stein, 2001). It has been suggested that the projections from cortical association areas to the superior colliculus develop postnatally and may be the final stage necessary for multisensory integration to occur (Wallace & Stein, 2000). The overall result of these studies is the suggestion that a great deal of development—with a corresponding change in overt response behavior—may be taking place in infants in how they are processing multisensory inputs.

1.2.2 Sensory Maturation

Before the integration of the auditory and visual modalities can be studied, it is important to understand the limits and capabilities of the individual sensory modalities and motor system in infancy. While it has been found that a newborn is capable of orienting—with head and/or eye movements—toward a moderately loud, sustained sound presented in the horizontal plane (such as a rattle) or turning away from an excessively loud or startling one, the auditory system is still immature at birth and the various features of auditory perception (such as frequency and threshold sensitivity, temporal resolution, and sound localization), develop along different timelines, some not completely maturing until adolescence. It is not until four or five months of age that infants are even capable of localization in the vertical plane (Fantz, 1963; Morrongiello, 1988a; Ehret, 1988). Sound localization in the horizontal plane does tend to improve steadily over the first half year, with one benchmark behavioral exception: Infants between one and three months of age exhibit a significantly reduced probability of orienting toward sounds (D. W. Muir, Clifton, & Clarkson, 1989). This hiatus in the auditory localization response is not an indication that infants are not hearing, nor that it is a trade-off for improved visual function; infants show this deficit even when tested in the dark, possibly representing a period of development where the orienting response is changing from a basic, early system (subcortical) to another more complex and adult-like one (cortical) (D. Muir, Abraham, Forbes, & Harris, 1979; J. Field, Muir, Pilon, Sinclair, & Dodwell, 1980). After these first six months, sound localization continues to mature, but more slowly, until the child is well into adolescence. In terms of a baby's ability discriminate temporal durations of competing auditory stimuli, six month old infants require the relative durations to be approximately twice that of what an adult can discriminate, with their performance not yet at an adult level by six years of age, and their ability to discriminate closely spaced sounds not at an adult level until puberty (Morrongiello & Trehub, 1987).

Compared to the auditory system, the visual system tends to be more accurate in localization although visual localization skills in infants are still relatively poor and slow to develop. Visual acuity is low at birth and even infants at three months of age require much larger contrast and lower spatial frequency to distinguish between visual patterns than do adults (Gwiazda, Brill, Mohindra, & Held, 1978; Dobson & Teller, 1978). In addition to poor auditory and visual localization performance, young infants appear to utilize subcortical pathways for localization tasks, unlike the cortical-controlled response found in infants older than five months of age (Clifton, Morrongiello, Kulig, & Dowd, 1981b) and young infants have only crude control of their saccadic response. Infants under two months of age are five times as slow as adults to initiate a saccade to a peripheral target and when they do, it is a series of staggered saccades instead of a smooth ballistic trajectory (Aslin & Salapatek, 1975).

1.3 Infant Studies

The majority of previous infant multisensory or perceptual studies have focused on one particular form of multisensory interaction or age. This work represents a systematic approach toward advancing our understanding of the behavioral response—specifically the latency and orienting response—of infants in the crucial development range from birth to ten months when presented with simple auditory and visual stimuli at horizontal eccentricities that are spatially and/or temporally congruent or incongruent. We hoped to gain a better understanding of the factors pertinent to infants over a broad age range in their perception and response to audiovisual stimuli; in particular, the role of temporal and spatial congruency/incongruency and how their influence changes in the first ten months of life.

1.3.1 Babies Will Be Babies

As is not at all surprising, infants are perhaps the most strictly regulated subject pool for non-medical, scientific research. Parents are often hesitant to volunteer their baby for even the most benign of scientific studies, and even when willing, as any parent can tell you, there are the difficulties in getting the baby's cooperation. The logistics of working around naps, mealtimes, doctors' visits, and parents' work schedules make it a difficult prospect to get even new subjects through the door for those few brief minutes of usable presentation time when the baby is both alert and cooperative.

1.3.2 Innate Behavior and Practical Constraints

A second, and equally challenging, aspect of infant research is the lack of meaningful communication between the experimenter and the subject. Infants younger than ten months are not able to speak or accept verbal instructions, and it is necessary to rely on a few overt behaviors that infants naturally exhibit, namely their inclination to orient toward and maintain their gaze on a novel stimulus or event in their surroundings, or their asymmetrical preference in visually examining novel versus familiar, complex versus simple, or dynamic versus static stimuli. It is the former behavioral response we take advantage of in all three of the following studies. By placing the babies in a quiet environment we can present only the stimuli we are interested in getting them to attend. To accommodate the infants' very short attention span, experiments have been designed to be broken up into short blocks lasting no more than 3–6 minutes. Stopping when (not if) the infant becomes fussy or inattentive, or at the request of the parents, is not uncommon. Given the much shorter number of completed trials achieved compared to adult studies, results are pooled for infants within the same age range, and parents were encouraged to bring their child back for multiple visits to maximize the possibility of useful data as well as the benefit of gaining longitudinal data on individual infants.