Temporal Correlation and Its Role in Multisensory Feature Integration and Binding

By

Aaron Ross Nidiffer

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Hearing and Speech Sciences

August 31, 2018

Nashville, Tennessee

Approved:

Ramnarayan Ramachandran, Ph.D.

G. Christopher Stecker, Ph.D.

Adele Diederich, Ph.D.

Mark T. Wallace, Ph.D.

To those giants on whose shoulders I stood

so that I could see a bit further.

To those dwarves who want to grow to be giants

so others can gaze from upon their shoulders.

Acknowledgments

My work as a graduate student, having taken so long to complete, has allowed me to meet many people who have, perhaps unknowingly, contributed to this finished product. The list is long and while impossible to mention them all here, I am forever grateful for how I have been shaped by my teachers, peers, experiences, problems, and their solutions during my time as a student. The thank yous enumerated here cannot begin to pay off the debt that is owed.

My love for science and learning is set on a strong foundation built by mentors at King College. Drs. Kevin DeFord, John Graham, and Vanessa Fitsanakis inspired my fascination of the brain and how it makes us who we are. Dr. Craig MacDonald was the first to show me the value in learning and in the end made me realize that I needed to begin this journey.

When I began graduate school, I was unaware of how much of my training as a scientist would come from my peers. Fortunately, I have been able to work in a large lab. That, with my longevity in grad school, resulted in many lab mates—17 grad students, six post-docs, and five research assistants by my count—who were willing and available to supplement my classroom education with a more practical one. Four of these 28 mentors had a profound impact on me both professionally and personally. Dr. Al Powers exemplified the scientist that I strive to be: curious, excited, rigorous, and slightly neurotic. All the MATLAB help I have been able to give is because he helped me first. Many of the techniques and concepts that I learned were learned from the example of Dr. Juli Krueger. She was the first teacher I had at Vanderbilt and I still learn things from her. Dr. Diana Sarko shared with me the frustration and accomplishment of

building a lab from the ground up. Neither of us could have done it without the other. Dr. David Simon is my go-to problem solver and is one hell of a psychophysical observer. This dissertation would not be what it is without his input. All my lab mates have been great teachers and students. They will forever be great friends.

During my time as a graduate student, I have had outstanding mentorship. These scientists are committed to the discovery of knowledge and the process of training young scientists in our art. Drs. Dan Ashmead, Wes Grantham, Troy Hackett, Vivien Casagrande, and Sean Polyn served on early committees and helped guide my early development as a scientist. Dr. Chris Stecker exemplified that experiments can be rigorously designed and still be creative in their approach. Dr. Adele Diederich cultivated my appreciation for using math to describe the brain and our behavior. I will never forget that "math belongs to science," as she so elegantly put it. Dr. Ram Ramachandran nurtured my propensity to get caught up in the minutiae of science but directed me away from the meaningless to the important. I will miss all of the hours-long conversations that typically began "Quick question. Do you have a minute?" Dr. Mark Wallace conferred upon me an ability to think bigger. It wasn't an easy road navigating the meetings where Ram and I fought to see who could be more pedantic, but he tolerated my love for detail while making sure I didn't miss the big picture. Without his guidance my vision as a scientist would be too narrow to be practical.

Inside and outside the lab, I've had a dependable bunch to make the burden of graduate school easier to bear. My Nashville friends—Jimmy, Nick, Justin, Michael, Brandon—were always good for a drink, watching sportsball, making and eating too much food, and general

tom-foolery. I will miss their friendship as I leave Nashville. I am also thankful for the support of my family. We are a weird bunch, but it has made my life more enjoyable. My mom has always been my biggest champion. She always made choices with me and my best interest in mind and showed me how to make the right choices for myself. My dad taught me to question everything and why it's important to understand how things work. I've never seen anyone do so much with so little. Chris showed me that with enough effort, I could do whatever I wanted. They often tell me how proud they are of me, but anything they are proud of was their doing.

Most importantly, Kayla has been the singular constant through the variables of graduate school. She is there through good and the bad. She makes the difficult times bearable and the good times better. She's the Julian to my Ricky, the Patti to my Doug, and the Linda to my Bob. I can't love and appreciate her enough.

Table of Contents

Section Page

List of Tables

Table Page

List of Figures

Figure Page

Chapter 1. Introduction

> *"The senses which operate through external media, viz. smelling, hearing, seeing, are found in all animals which possess the faculty of locomotion. To all that possess them they are a means of preservation; their final cause being that such creatures may, guided by antecedent perception, both pursue their food, and shun things that are bad or destructive. But in animals which have also intelligence they serve for the attainment of a higher perfection. They bring in tidings of many distinctive qualities of things, from which the knowledge of truth, speculative and practical, is generated in the soul."*
>
> — Aristotle, De Sensu et Sensibilibus

The Multisensory World

I answer my ringing phone and hear my friend's voice. A group is meeting up to go out for lunch. Outside my apartment on the busy street, I join a mob of pedestrians rushing past. As I am walking down the sidewalk, a barrage of events occurs all around. Countless pedestrians engage in conversations all around me. A construction crew is working on a new high-rise building. I pass food carts where venders sell hot dogs and pizza. As I approach the intersection, an airplane slides across the sky above. Across the street in a park, a dog is barking and trees sway in the wind. The same wind ruffles my clothes. Cars are rushing past; one driver beeps their

horn. Finally, the cars stop and there is a beep as the crosswalk figure lights up, indicating that it is safe to cross.

These events in my environment generate a variety of energies—electromagnetic radiation, waves of air pressure, physical manipulation upon the skin, and chemical odorants—which propagate outward from their sources. As I navigate the urban jungle, a portion of these energies is captured by my sensory periphery. They are transduced into the neural language and transmitted to my brain for processing into sensations and perceptions of light, sound, touch, and smell. Somehow, from the cacophony of sensations that make up the sensory landscape, my brain has composed a coherent symphony and has identified the instrument corresponding to the signal to cross the street.

Our accurate perception and interaction with our environment depend critically on the appropriate integration of these sensory signals. The many ongoing events which cause the sensory landscape to become chaotic can make this a difficult endeavor. However, our brains accomplish this task seamlessly, using statistical regularities in sensory signals to combine and separate sensory features into distinct sensory objects (Bizley & Cohen, 2013; Shinn-Cunningham, 2008; Wagemans et al., 2012). Many of the events that occur in the environment

have features that span sensory modalities. The brain uses regularities that occur across modalities to group features from different modalities (Bizley et al., 2016). In doing so, the brain forms a coherent and unified picture of our multisensory environment.

The brain can synergistically combine redundant or complementary information across the different senses to enhance the representation of events in the environment. The behavioral and perceptual manifestations of multisensory integration are well established (Calvert, Spence, & Stein, 2004; Murray & Wallace, 2012). In general, multisensory integration often leads to behavior that is more accurate (Frassinetti, Bolognini, & Làdavas, 2002; Ohshiro, Angelaki, & DeAngelis, 2011; Stein, Huneycutt, & Alex Meredith, 1988) and also faster (Amlôt, Walker, Driver, & Spence, 2003; Colonius & Diederich, 2004; Diederich, Colonius, Bockhorst, & Tabeling, 2003; Frens, Van Opstal, Van der Willigen, Opstal, & Willigen, 1995; Hershenson, 1962; Hughes, Reuter-Lorenz, Nozawa, & Fendrich, 1994) than behavior based on unisensory signals. These multisensory interactions are ubiquitous and have been shown in domains such as detection (Bolognini, Frassinetti, Serino, & Làdavas, 2005; Frassinetti et al., 2002; Grant & Seitz, 2000; Lovelace, Stein, & Wallace, 2003), localization (Battaglia, Jacobs, & Aslin, 2003; Bolognini, Leo, Passamonti, Stein, & Làdavas, 2007; Hairston, Laurienti, Mishra, Burdette, &

Wallace, 2003; Nidiffer, Stevenson, Krueger Fister, Barnett, & Wallace, 2016; Wallace et al., 2004), speech discrimination (Erber, 1969; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954), target selection (Corneil, Van Wanrooij, Munoz, & Van Opstal, 2002; Kösem & van Wassenhove, 2012), and attention (Convento, Rahman, & Yau, 2018; Maddox, Atilgan, Bizley, & Lee, 2015; Mast, Frings, & Spence, 2017). They have been demonstrated in manual responses and saccadic eye movements (Amlôt et al., 2003; Colonius & Diederich, 2004; Corneil et al., 2002; Frens et al., 1995; Hughes et al., 1994).

Multisensory presentations can enhance perception in unisensory domains (Sumby & Pollack, 1954), even when the information in the other modality is irrelevant (Colonius & Diederich, 2011; Lovelace et al., 2003; Maddox et al., 2015). Multisensory combinations can give us the "best of both worlds," for example, the speed of auditory behavior and the spatial accuracy of visual behavior (Corneil et al., 2002). Conversely, multisensory interactions can make observers less sensitive to spatial and temporal conflicts in unisensory signals (Parise, Harrar, Ernst, & Spence, 2013; Vatakis & Spence, 2007), errantly bias unisensory perception (Alais & Burr, 2004; R. Sekuler, Sekuler, & Lau, 1997; Shams, Kamitani, & Shimojo, 2000; Wallace et

al., 2004), and even elicit a percept that is absent from the unisensory signals (Mcgurk & Macdonald, 1976).

Multisensory guiding principles

Some of the earliest descriptions of the neural instantiation of multisensory interactions were described in the optic tectum of the rattlesnake (Newman & Hartline, 1981) and in the superior colliculus (SC; the mammalian homologue of the optic tectum) of the cat (Meredith & Stein, 1983; Stein, 1978; Stein & Arigbede, 1972). The SC is a midbrain structure that receives converging inputs from a very large portion of the brain including visual, auditory, and somatosensory areas in cortex (Kawamura & Konno, 1979) and sub-cortex (Edwards, Ginsburgh, Henkel, & Stein, 1979). These inputs converge on single neurons in the SC and result in neural activity that is often profoundly changed by the presentation of multisensory signals (Meredith & Stein, 1986b). In terms of the strength of the multisensory response relative to the unisensory response, these interactions can be described as response depression (multisensory < unisensory) or response enhancements (multisensory > unisensory). Response enhancements can further be divided based on a comparison to the sum of unisensory responses into additive, sub-additive, and super-additive interactions (Stanford & Stein, 2007; Stevenson

et al., 2014). The magnitude of multisensory interactions were determined not to be a sole property of the neuron (c.f., Perrault, 2003, 2005) but were guided by a set stimulus factors: spatial and temporal proximity and stimulus effectiveness.

Temporal proximity

Stimuli that occur in close temporal proximity can produce multisensory interactions in single neurons whereas stimuli separated by a long interval are processed as distinct unisensory events (Meredith, Nemitz, & Stein, 1987). However, the temporal relationships that guide multisensory interactions are not quite this simple. Multisensory interactions are not dependent on matching the onsets of multisensory stimuli per se, but rather how their resultant activity patterns overlap in a neuron. This notion is further complicated by two key elements. First, the multisensory system is tasked with combining energies that propagate at very different velocities through the environment. At room temperature conditions, light travels at =299,792,456 m/s (Evenson et al., 1972) while sound is much slower at 343 m/s (Dean, 1979). In contrast, somatosensory energy does not travel through the atmosphere and activates receptors on the skin instantly. Second, once stimulus energy has reached the sensory receptors and is transduced into neural energy, the times it takes for neuronal activity to reach the SC are substantially different

across the sensory systems (Meredith et al., 1987). Auditory stimuli cause the discharge of action potentials in SC neurons in approximately 5-29 ms (Middlebrooks & Knudsen, 1984), spikes are present about 9-18 ms after a somatosensory stimulus is applied to the skin (Stein, Magalhaes-Castro, & Kruger, 1976), and visual stimuli takes quite a bit longer—39-145 ms—to induce spiking activity in SC (Meredith et al., 1987; Stein & Arigbede, 1972). These issues explain two somewhat surprising findings. First, the most intense interactions often occur with some asynchrony among sensory signals, and second, the duration of the "temporal window" over which interactions can occur is rather long. This window has been interpreted as a means by which we integrate multisensory signals across different distances. Because auditory energy propagates slowly in the environment relative to visual energy, multisensory signals at further distances result in longer delays in the arrival of the auditory portion of the signal.

Spatial proximity

Neurons in the SC have receptive fields that respond to stimuli which are present in well circumscribed areas of visual (McIlwain, 1975; Meredith & Stein, 1990), auditory (Gordon, 1973; Middlebrooks & Knudsen, 1984), and somatosensory (Meredith, Clemo, & Stein, 1991; Stein et al., 1976) space. These spatial receptive fields are organized topographically in the SC

with a common organizational structure across sensory modalities. Therefore, neurons in the SC that respond to multiple sensory signals have separate but overlapping spatial receptive fields for each of its sensory modalities. The result of such an organization is that spatially coincident multisensory stimuli tend to fall within their respective spatial receptive fields and thus produce response enhancements proportional to the spatial congruence (Meredith & Stein, 1986a, 1996). When multisensory stimuli are spatially disparate or when one stimulus falls outside its receptive field, SC responses are depressed or are not distinct from unisensory responses. Typically, spatially proximal stimuli fall within their respective receptive fields, and therefore produce multisensory enhancement. However, spatial proximity is not a strict requirement for multisensory enhancements. When animals are reared in an altered sensory environment where temporally-coupled audiovisual signals originate from different locations in space, SC neurons develop unisensory spatial receptive fields that are separated in space commensurate with the discrepancy in their environment (Wallace & Stein, 2007). When stimuli are spatially coincident, one stimulus falls within its receptive field while the other falls out outside its receptive field. These neurons show normal multisensory integrative ability so long as unisensory

signals are separated by a distance that corresponds to the statistics of the developmental environment and thus occur within their respective receptive fields.

Inverse effectiveness

Maximum multisensory enhancements occur when weakly effective stimuli are combined (Meredith & Stein, 1986b). One of the benefits of multisensory interactions is the ability to enhance the physiological salience of external events which can help an organism to maximize the extraction of information from their surroundings. An organism does not stand to benefit much from the integration of stimuli that are effective on their own. Conversely, when stimuli are decreasingly effective (e.g., from low signal or high background noise), multisensory interactions become increasingly advantageous to that organism. Interestingly, this principle has been demonstrated by changing the spatial location of stimuli within the heterogeneous receptive fields of SC neurons (Carriere, Royal, & Wallace, 2008; Krueger, Royal, Fister, & Wallace, 2009). Large interactions are observed when stimuli are placed at receptive field locations which produce meager responses while highly excitable receptive field locations produce much smaller multisensory enhancements, if any.

Table 1.1: The multisensory principles outside single unit electrophysiology.

	Behavior	Electrophysiology	Imaging
Temporal Coincidence	(Dixon & Spitz, 1980; Hershenson, 1962; Wallace et al., 2004)	(Schall, Quigley, Onat, & König, 2009; Senkowski, Talsma, Grigutsch, Herrmann, & Woldorff, 2007)	(Macaluso, George, Dolan, Spence, & Driver, 2004)
Spatial Parity	(Bolognini et al., 2005; Frassinetti et al., 2002; Wallace, 2004)	(Zhou, Zhang, Tan, & Han, 2004)	(Macaluso et al., 2004)
Inverse Effectiveness	(Sumby & Pollack, 1954)	(Crosse, Di Liberto, & Lalor, 2016; Stevenson et al., 2012)	(Stevenson & James, 2009)

The principles outside the single neuron

Although the multisensory principles were codified using electrophysiological recordings of single SC neurons in cat and later in the primate SC (Wallace, Wilkinson, & Stein, 1996), they were first touch on in human nearly a century ago (Todd, 1912). These principles, based on the early physiology work (Meredith et al., 1987; Meredith & Stein, 1983, 1986a), have since been demonstrated in a wide range of human methodologies including behavior, electrophysiology, and imaging. Table 1 summarizes a few key papers.

Other considerations

Although the principles have been demonstrated and remain robust across a broad variety of tasks and techniques, there are select cases in which multisensory interactions do not conform to the principles. For example, although Sumby and Pollack (1954) found that enhancement increased monotonically with decreasing signal-to-noise ratio (SNR) of a speech signal (inverse effectiveness), others have found maximal enhancement occurring at intermediate SNRs, a so-called "sweet spot" (Foxe et al., 2015; Ross, Del Bene, Molholm, Frey, & Foxe, 2015; Ross et al., 2007), or in a manner unexplained by inverse effectiveness altogether (Chandrasekaran, Lemus, Trubanova, Gondan, & Ghazanfar, 2011). Additionally, there is evidence that the spatial misalignment can still cause integration under certain contexts (Murray et al., 2005).

Other factors have been presented that shape the multisensory product. One of these posits that multisensory signals must be semantically congruent to produce multisensory enhancements (Laurienti, Kraft, Maldjian, Burdette, & Wallace, 2004). When participants were shown a red or blue visual stimulus, they were faster at pressing a corresponding red or blue button when there was a semantically congruent auditory stimulus (spoken "red" or "blue"). Multisensory stimuli that are semantically incongruent (spoken "green", akin to a multisensory

Stroop Effect) were shown to impede behavioral performance. Further, two principles aimed to explain multisensory behavior have been proposed (Otto, Dassy, & Mamassian, 2013). The principle of congruent effectiveness states that multisensory behavioral enhancement is largest when behavioral performance in corresponding unisensory conditions is matched. The variability principle says that multisensory enhancements increase as unisensory behavior becomes less reliable. Interestingly, both of these can be thought of a nuanced reframing of the inverse effectiveness principle. First, mismatched unisensory behavioral performance reduces multisensory enhancements due to the increased effectiveness in one modality. Likewise, unisensory behavior becomes more variable as stimuli become less effective and therefore produces larger response enhancements.

The Binding Problem

Originally, the multisensory principles were described as a means to explain what the brain should integrate under the assumption that stimuli that are spatially and temporally congruent likely originate from a common event (i.e., causal inference; Körding et al., 2007; Magnotti, Ma, & Beauchamp, 2013) and therefore should be "bound" together. These statistical cues are important because the brain does not know *a priori* what sensory features should go

together and which should be separated into distinct objects. So the brain uses cues to *infer* a common cause for sensory events. This binding problem deals with the question of how we achieve the experience of a coherent world of integrated objects, and avoid seeing a world of disembodied or wrongly combined shapes, colors, motions, sizes and distances (Treisman, 1998). In order to evaluate and interact with our environment, we bind features from the same source into a single object while segregating features from different sources into multiple objects (Treisman & Gelade, 1980). In order to form these objects, our brains must first encode its component features and then *how* the features are combined (Treisman, 1996).

Gestalt psychologist have long been aware of the binding problem. One core principle of Gestalt psychology is prägnanz, which is the observation that humans tend to order our perceptual experience in predictable ways. These observations lead to the formalization of principles describing the conditions which facilitate perceptual grouping (Wertheimer, 1923, 1938). Their principles, such as similarity, proximity, and common fate – which describe grouping based on resemblance, closeness, and movements—relate closely to binding cues (i.e., spatial and temporal correlations) described in research today.

Unisensory grouping and segregation

Our sensory environment contains many signals that need to be appropriately combined. Returning to our early example, where I have just crossed the busy street, I am approached by my friend. For us to have a conversation, my auditory system must group the audible features (e.g., frequency components, spatial location) related to his voice and all the other ongoing sensory events into auditory objects and segregate them appropriately. These objects then must maintain continuity across longer time scales into streams. Humans are particularly adept at tracking these changes across time, even when multiple sources overlap significantly (Woods & Mcdermott Correspondence, 2015). Finally, we must focus attention on the target voice (Shinn-Cunningham, 2008). This process has been called the cocktail party problem (Cherry, 1953), which is a specific instance of auditory scene analysis (Bregman, 1990) and represents a non-trivial problem for our auditory perceptual system.

Auditory scene analysis takes advantage of a number of cues based in part on the spectral features of sounds (Darwin, 1997). One well-documented cue for grouping and segregation of sound sources is frequency proximity (Darwin, 1992). If two vowels are presented simultaneously, larger differences in the fundamental frequency—which is potentially indicative

of two speakers—result in better identification of both vowels (Assmann & Summerfield, 1994; Culling & Darwin, 1993; Culling & Darwin, 1994; de Cheveigné, Kawahara, Tsuzaki, & Aikawa, 1997). Similarly, in a classic paradigm designed to manipulate the perception of the number of perceived streams, a repeating sequence of alternating tones (A and B) are separated in frequency by varying amounts. In this paradigm, if the frequency separation between the tones is small, the sequences will be perceived as a single trill-like stream (ABAB). However, if the tones are separated sufficiently in frequency, the sequence will be perceived as two separate streams (A-A- and -B-B; Bregman & Campbell, 1971; Miller & Heise, 1950; Noorden, 1977).

Although frequency separation is important in the formation of auditory streams, another factor—the temporal structure of sounds—is also crucial in this process (Zion Golumbic, Poeppel, & Schroeder, 2012). Listeners are able to detect onset and offset asynchrony in sound features and can use these cues for segregating sounds (Darwin & Carlyon, 1995). In the ABAB paradigm, tone streams that are synchronous are more likely to be grouped into a single stream, regardless of frequency separation (Shamma et al., 2013). Tone sequences that are presented at different speeds, thus decoupling the temporal correlation between tones sequences, leads to the perception of separate of streams, regardless of frequency similarity, a result that is evident in

behavioral and neurophysiological data (Elhilali, Ma, Micheyl, Oxenham, & Shamma, 2009; Micheyl, Hanson, Demany, Shamma, & Oxenham, 2013). Listeners can also take advantage of temporal coherence independent of onset in complex sounds that change stochastically over time to segregate a figure from a background (O'Sullivan, Shamma, & Lalor, 2015; Teki, Chait, Kumar, Shamma, & Griffiths, 2013; Teki, Chait, Kumar, von Kriegstein, & Griffiths, 2011).

Just as auditory grouping relies strongly on cues encoded topographically (i.e., frequency), visual grouping relies on the topographically organized spatial structure (Blake & Lee, 2000; Kramer & Yantis, 1997; Rikhye & Sur, 2015). And paralleling auditory grouping, correlated temporal structure is an important cue in visual feature binding as well (Blake & Lee, 2005; Fahle, 1993; Treisman, 1999) and is more important than precise synchrony in onset (Guttman, Gilroy, & Blake, 2007). Visual feature grouping has been demonstrated via correlated or synchronized motion cues (Kandil & Fahle, 2004; S. H. Lee & Blake, 1999), sinusoidal gratings (Alais, Van Der Smagt, Van Den Berg, & Van De Grind, 1998), and paralleling auditory streaming paradigm, luminance changes that do not rely on spatial grouping cues (A. B. Sekuler & Bennett, 2001). There is even evidence that temporal cues can override spatial cues in visual perception (J. M. Wallace & Scott-Samuel, 2007). The brain's use of temporal structure and

correlation in the analysis of auditory and visual scenes (Bizley & Cohen, 2013; Blake & Lee, 2000) suggests a general mechanism of feature binding built on temporal correlations that can be used to group features across sensory modalities.

Multisensory binding and temporal correlation

After my auditory system has grouped the spectrotemporal features of my friend's voice into a stream, I am left with the challenging task of understanding what my friend is saying. In quiet, we can easily understand audible speech but performance drops as the environment gets louder or when distractors are present (Hygge, Rönnberg, Larsby, & Arlinger, 1992; Miller, 1947). And so, on the loud city street and with everyone talking around us, I struggle to understand what a single friend is saying. So naturally, I look to his face for help.

Although it is natural for us to use facial cues to aid in our listening ability, our visual system is not very adept at understanding speech (Romano & Berlow, 1973; Ross et al., 2007). It is not immediately apparent *how* we derive a benefit from an input that carries so little information, but despite our poor lip reading abilities, when we are able to see the face of a person speaking to us, we are better able to understand what is said (Erber, 1969; Ross et al., 2007; Sumby & Pollack, 1954). It has been proposed that one way in which we derive this

benefit is from the enhancements bestowed by forming a *multisensory* object and directing attention toward it (Bizley et al., 2016). This proposition is rooted in the tenets of object-based visual attention (Desimone & Duncan, 1995) where objects are the recipients of our attention and that attention enhances all features of an object. Indeed, it has been posited that attention is a critical component of the binding process (Treisman, 1998).

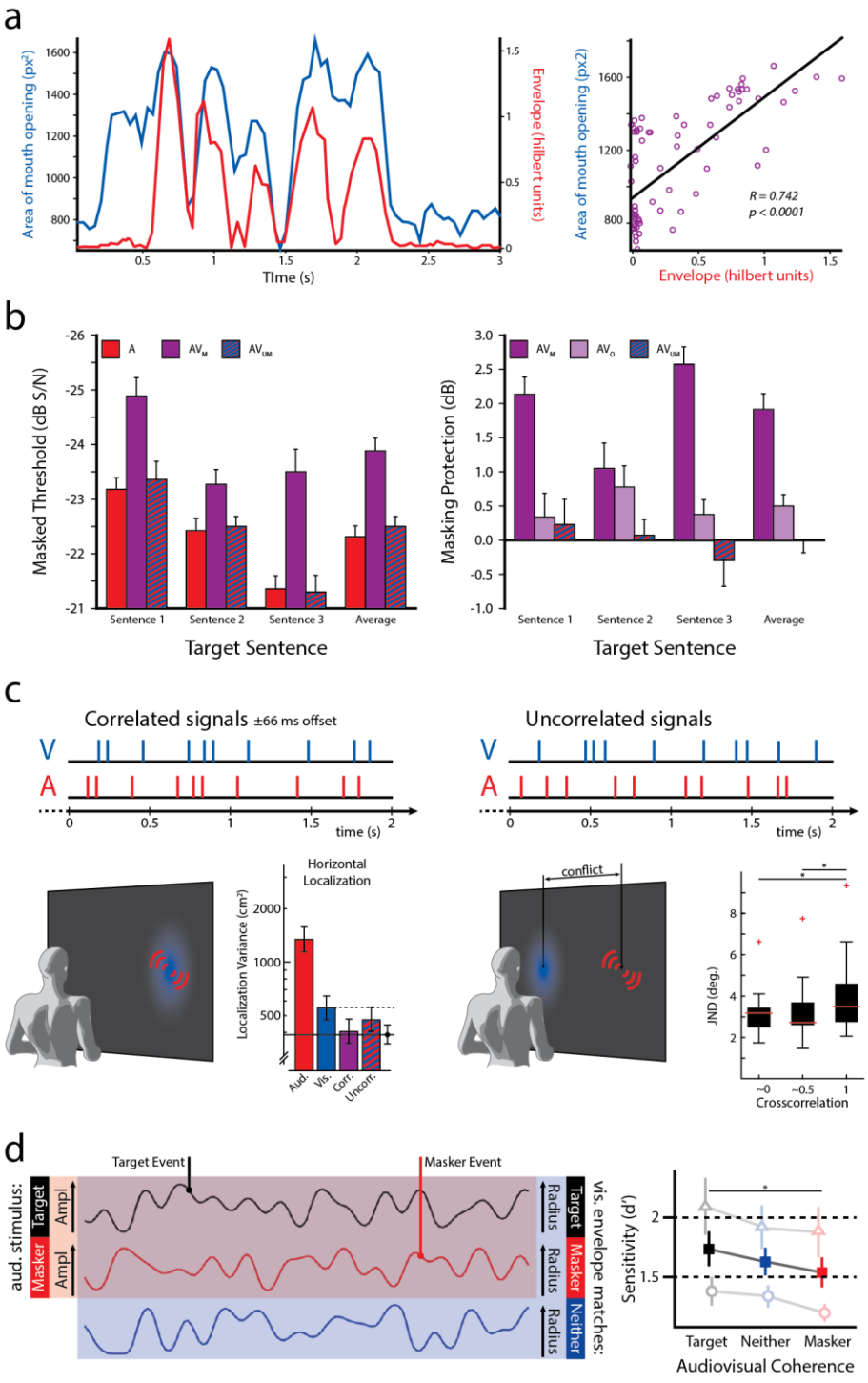
When a multisensory object emerges in our perception, that object becomes more salient than a unisensory one (Bizley et al., 2016). This becomes apparent in the context of stimulus competition and selective attention (e.g., the cocktail party). In a recent task that parallels the cocktail party problem, participants were engaged in streaming two auditory objects and were asked to report brief frequency or timbre events in one stream while ignoring events in the other. Performance was better when a separate visual stream was temporally correlated with the target stream than when it match the distractor (Maddox et al., 2015).

Just as we are likely to group unisensory features that change together over time, temporal correlation is said to be a strong determinant of multisensory binding (Bizley et al., 2016) and as mentioned above is a property of audiovisual speech (Figure 1.1a; Chandrasekaran, Trubanova, Stillittano, Caplier, & Ghazanfar, 2009). Visual speech has been shown to improve the detection

of auditory speech tokens when the two are temporally correlated (Bernstein, Auer, & Takayanagi, 2004; Grant & Seitz, 2000; Kim & Davis, 2004). This benefit is abolished when the visual stimulus is decorrelated by time-reversing the visual stimulus (Kim & Davis, 2004) or with presentation of unmatched visual speech (Figure 1.1b; Grant & Seitz, 2000). Interestingly, with a simple visual stimulus where the size is modulated by the speech envelope, thus preserving the temporal correlation, a small but reliable detection benefit is still observed (Bernstein et al., 2004) suggesting that temporal correlation, and not just the presence of a speaker's face—or any visual cue for that matter—is responsible for benefits conferred by visual speech on auditory speech perception.

Figure 1.1(next page): Multisensory temporal correlation and its consequences on behavior. *(a) Mouth opening area and auditory envelope intensity from an exemplar spoken sentence are plotted across time (left). Mouth opening and envelope are temporally correlated (right).* ***(b)*** *Detection thresholds in noise for three spoken sentences under auditory only (A) or paired with a matched (same sentence) visual stimulus (AV_M) or unmatched (different sentence) visual stimulus (AV_U; left). Masking protection (AV-A) conferred by the addition of matched, unmatched, or orthographic (AV_O; written sentence) visual stimulus.* ***(c)*** *Localization (left) and spatial discrimination task (right) using correlated or uncorrelated streams of Gaussian flashes and noise burst. When stimuli are correlated, stimuli are integrated in a statistically optimal fashion and are harder to separate in space.* ***(d)*** *An auditory event (frequency modulation, vowel shift) is presented in an auditory target or masker streams (tone complex or synthetic vowel) with randomly fluctuating amplitude envelopes (left). Participants were asked to detected target events and ignore masker events. Meanwhile, a visual stream containing a disk where the size could be modulated with an envelope that matched the auditory target, masker, or neither stream. When the*

visual stream matched the auditory target stream, sensitivity in detecting the target event was improved (right). Adapted from (Chandrasekaran et al., 2009; Grant & Seitz, 2000; Maddox et al., 2015; Parise, Spence, & Ernst, 2012; Parise et al., 2013).



In several tasks and stimulus types, temporal correlation of non-speech auditory and visual cues has been shown to facilitate multisensory binding and integration. In a localization task, streams of discrete auditory and visual events were combined in a statistically optimal fashion (i.e., reducing the perceptual variance; Ernst & Banks, 2002) only if the auditory and visual streams were correlated over time (Figure 1.1c; Parise et al., 2012). When the same stimuli were presented with varying degrees of spatial separation, auditory and visual streams that were correlated required larger separation than uncorrelated stimuli for participants to reliably judge the relative location of the unisensory parts (Figure 1.1c; Parise et al., 2013), suggesting that correlated stimuli are unified into a multisensory object that is resistant to deconstruction. During auditory stream segregation, selective attention toward a target stream in the presence of a distractor is improved with the presence of a temporally correlated, yet irrelevant, visual stream (Figure 1.1d; Maddox et al., 2015).

Although the spatial and temporal proximity of bimodal cues is an important feature that modulates multisensory processing (Bolognini et al., 2005; Frassinetti et al., 2002; Meredith et al., 1987; Meredith & Stein, 1986a, 1996), the formation of a unified object can occur despite large spatial and temporal discrepancies (Wallace et al., 2004). This is often demonstrated when

audiovisual stimuli are temporally correlated, and thus likely to be bound. Observer are more tolerant of spatial and temporal separation between stimuli that are temporally correlated (Chuen & Schutz, 2016; Parise et al., 2012, 2013; Vatakis & Spence, 2007). This phenomenon is the basis of an illusion termed the Ventriloquist Effect (Bertelson, 1999; Bertelson, Vroomen, Wiegeraad, & de Gelder, 1994). In this illusion, an auditory stimulus' perceived location is shifted toward a more salient visual stimulus (Alais & Burr, 2004; Wallace et al., 2004). When the temporal correlation between the auditory and visual components is removed, the ventriloquism effect, that is, the spatial tolerance, is extinguished. (Jack & Thurlow, 1973; Thurlow & Jack, 1973).

In another illusion, the McGurk Effect, which results from the binding and fusion of disparate bimodal cues (Nahorna, Berthommier, & Schwartz, 2012), the pairing of a semantically incongruent auditory (/ba/) and visual (/ga/) can result in a fused percept (/da/ or /tha/) (Mcgurk & Macdonald, 1976). The McGurk effect is also resilient to spatial incongruence (Bertelson et al., 1994). However, manipulation of the temporal structure present in the visual cue (such as presenting a time-warped cue or cues from different speaking speeds) degrades the temporal correlation between auditory and visual signals and results in a reduction of the fused

percept (Munhall, Gribble, Sacco, & Ward, 1996; Venezia, Thurman, Matchin, George, & Hickok, 2016). Additionally, presentation of incoherent auditory and visual streams prior to a McGurk target leads to fewer reports of a McGurk percept (Ganesh, Berthommier, & Schwartz, 2017; Nahorna et al., 2012).

Thus far, we have discussed multisensory objects in terms of stimuli in the environment, namely speech. However, a particular case of multisensory objecthood results from the binding of of exteroceptive and interoceptive cues (Blanke, Slater, & Serino, 2015), which has been described as the multisensory basis of embodied self-consciousness (Noel, Blanke, & Serino, 2018). Interestingly, we can demonstrate this binding by generating binding failures through spatial discrepancies between these cues. Importantly, temporal congruence is an important constraint in these examples. In one, repetitive and temporally coincident [but not temporally disparate (Tsakiris & Haggard, 2005)] stimulation of a rubber hand and a participant's hidden hand can alter bodily ownership as participants reported feeling the touch on the rubber hand (Botvinick & Cohen, 1998). In a more impressive demonstration of this concepts, participants can take ownership of an entire virtual body presented in front of them via virtual reality

(Lenggenhager, Tadi, Metzinger, & Blanke, 2007). Again, the illusion is dependent on temporal cues, demonstrating the importance of time in multisensory binding and perception.

Correlations in the Environment and the Brain

So far, we have discussed in depth how sensory features can have temporal structure and how those features can be correlated within and across sensory modalities. But correlations can arise in other domains in the environment. When we encounter a natural visual scene, it is composed of a wide range of features such as luminance, contrast, orientation, and spatial frequency. These features occur with a highly organized and auto-correlated spatial structure (Ruderman, 1994). That is, areas in natural scenes tend to be similar (i.e., correlated) to neighboring areas with similarity decreasing with distance. For example, the features (e.g., luminance and edges) of a visual object tend to be continuous throughout that object and distinct from other objects in a scene. These statistics in visual scenes are used by the visual system to efficiently encode spatial information from the environment (Barlow, 2001; Rikhye & Sur, 2015; Simoncelli & Olshausen, 2001).

Binaural hearing is perhaps the prototypical implementation of correlation detection in the nervous system (Jeffress, 1948). Sound signals reach the two ears with a temporal delay,

known as interaural time difference (ITD), which is proportional to the azimuthal location of the sound source (Feddersen, Sandel, Teas, & Jeffress, 1957). Neurons in the medial superior olive are sensitive to specific ITDs. These neurons behave like cross-correlators, responding cyclically to increasing ITD and maximally when the ITD produces a phase difference between the ears that correspond to neurons' preferred phase difference (Goldberg & Brown, 1969).

Beyond these examples, correlation detection has been implicated in several sensory and cognitive processes such as stereoscopic vision (Ohzawa, 1998), texture segregation (Bergen & Landy, 1991), and motion perception (Hassenstein & Reichardt, 1956). Additionally, correlation across sensory modalities cues has been posited as a mechanism for synchrony perception (Burr, Silva, Cicchini, Banks, & Morrone, 2009; Fujisaki & Nishida, 2005), cross-modal temporal encoding (Guttman, Gilroy, & Blake, 2005), and as a general mechanism for multisensory integration (Parise & Ernst, 2016).

The Drift-Diffusion Model in Perception

The current work also seeks to place the role of correlation in a decisional framework and thereby ascertain its role in the decision-making process. Decision making is a ubiquitous occurrence in our interaction with the environment. However, there are many factors, both

external (i.e., from the environment) and internal (i.e., from the observer) that contribute to this process (Gold & Shadlen, 2007). A univariate approach to studying behavior, such as analysis of reaction times or accuracy in isolation, can miss the interactions between these metrics and how they relate to and explain both internal and external factors (Voss, Nagler, & Lerche, 2013; Wagenmakers, 2009). However, by accounting for choice probabilities as well as shape of reaction time distributions for both correct and error responses [which univariate analyses often overlook (Luce, 1986)], decisional models can successfully disentangle these factors that contribute to decision-making (Ratcliff & Rouder, 1998).

During the formation of a simple decision, sensory evidence is accumulated in favor of a single choice between multiple alternatives (Churchland, Kiani, & Shadlen, 2008; de Lafuente, Jazayeri, & Shadlen, 2015; Gold & Shadlen, 2007; Roitman & Shadlen, 2002). When organisms have access to stronger sensory evidence we are able to accumulate evidence faster and with better accuracy, which is reflected in our decisions. In the above examples, sensory evidence is manipulated through the coherence of random dot motion stimulus, but can be as simple as stimulus intensity (Rach, Diederich, & Colonius, 2011) or a more complex measure such as

memory salience (Ratcliff, 1978). However, a number of other factors influence the decision-making process such as bias, evidence encoding time, and speed/accuracy trade-off.

These factors of the decision-making process have been formalized in several models of evidence accumulation (Ratcliff & Smith, 2004; Ratcliff, Smith, Brown, & McKoon, 2016). These include models include random walk (Laming, 1968; M. Stone, 1960), Wiener diffusion (Ratcliff, 1978), Ornstein-Uhlenbeck diffusion (Busemeyer & Townsend, 1992), leaky competing accumulator (Usher & McClelland, 2001), and linear ballistic accumulator (Brown & Heathcote, 2008) models, to name a few. Although these models differ in several nuances such as decisional criteria that are relative or absolute, number of accumulators and/or decision bounds, constant versus decaying drift rates, or the stochastic/deterministic nature of the evidence accumulation process, they share some core commonalities. In short, evidence begins at a starting point between two decisional thresholds or as a fraction of one threshold and accumulates toward a threshold at a certain rate. Once the evidence crosses the decision threshold, the model assumes a choice is made and that simulated choice and reaction time are recorded. Models typically include parameters that can account for differences in evidence starting point (bias), evidence threshold (speed/accuracy trade-off), non-decision time (pre-

decisional encoding of stimuli) and evidence accumulation rate (strength of sensory evidence). Often, there are one or more parameters involving variability of these processes so that simulating large numbers of decisions results in reaction time distributions that vary predictably with the manipulation of a given parameter (Ratcliff & McKoon, 2008). These parameters can then be fit to observed reaction time distributions (Voss et al., 2013) or mean reaction times (Wagenmakers, Van Der Maas, & Grasman, 2007).

One popular and well-documented evidence accumulator model, the drift-diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008; Voss et al., 2013) assumes accumulation of noisy evidence that is sampled sequentially. Thus, the decision variable is modeled as a stochastic process where the value takes on the cumulative sum of random changes over time (i.e., particle diffusion). A constant value, whose magnitude and sign are related respectively to strength and direction of sensory evidence, is added to this process at each time point and causes the decision variable to trend toward (i.e., drift) a positive or negative decision threshold. We've chosen to use this class of model due to its biologically relevant parameters that have been extensively validated.

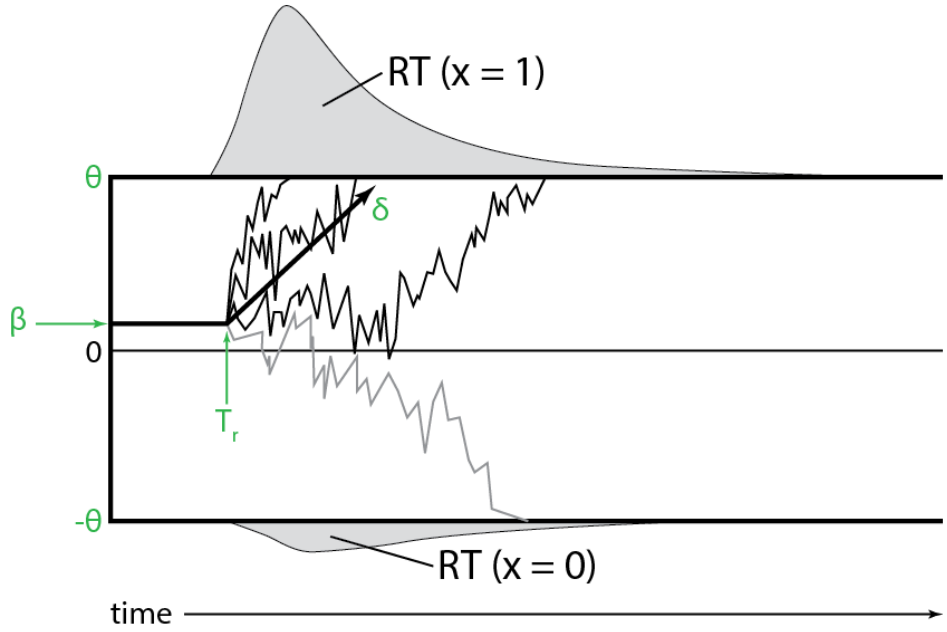


Figure 1.2: A depiction of the diffusion model and its parameters. *β: evidence starting point (i.e., participant bias), θ: decisional threshold (speed-accuracy trade-off), T_r: non-decision time (sensory encoding and motor time), δ: drift rate (magnitude of sensory evidence). Adapted from (Vandekerckhove, Tuerlinckx, Bulletin, Vandekerckhove, & Tuerlinckx, 2007).*

Figure 1.2 depicts the parameters of the drift-diffusion model and four simulated decision variables. The parameters have been shown to index a handful of relevant cognitive processes (Voss, Rothermund, & Voss, 2004). The decision variable begins with a value, β, which is constrained to be between the values of the decision thresholds, ±θ. A value for β that is different from 0 represents participant bias toward one decisional alternative and its sign represents the alternative that bias favors. The value of θ has been shown to index the speed-accuracy trade-off. Lower values of θ represent a more liberal strategy (higher speed, lower

accuracy) and larger values represents a more conservative strategy (lower speed, higher accuracy). The decision variable drifts toward ±θ with a rate, δ, proportional to the strength of stimulus evidence, with a sign reflecting the decisional alternative the evidence supports. A time adjustment, T_r, is added to each process to account for non-decision time and motor processes. In Figure 1, three decision variables are shown (thin black lines) that come from the drift rate that is shown (thick black line). Variability introduced through the diffusion process causes the decision variables to terminate at different times, allowing variability in the reaction time distributions. A single decision process from a negative drift rate is shown (thin grey line).

Thesis

On the busy sidewalk I have found myself walking down, I am likely to have a difficult time talking to my friend because of others talking around us. These other conversations produce signals that are fairly similar to the voice I am trying to listen to. So, the simple presence of audiovisual temporal correlation would be insufficient to induce appropriate binding and thus may lead to perceptual errors. The audible speech is, of course, strongly correlated with the articulatory movements of his mouth which produces that speech, but it also likely to be correlated (though with less strength) with mouth movements from the other speakers. Assessing

the strength of the correlations between stimuli could aid in the process of selecting the appropriate visual signal in an environment with many competing stimuli. Thus far, no systematic study of audiovisual correlation and its effect on multisensory integration has been undertaken. Studies on the importance of temporal correlations in multisensory binding have focused on the role of correlation in optimal cue combination (Parise et al., 2012), spatial fusion (Parise et al., 2013), auditory scene analysis (Maddox et al., 2015), and speech processing (Kim & Davis, 2004; Munhall et al., 1996; Venezia et al., 2016). These studies present a very coarse description of the role of correlation in multisensory processes, utilizing only two or three levels of correlation. The current work seeks to simplify the task and stimuli to explore and better characterize the role of temporal correlation in multisensory perception and binding.

In Chapter 1, we have laid out how temporal correlation is an important cue that is informative of whether sensory information belong to the same external event and therefore what sensory-related activity should be linked in the brain. This link (i.e., binding) in the brain relies, in part, on neural synchrony across regions, presumably related to the correlation of the different features of the external event. Here we hypothesize that the brain has a mechanism for binding that relies on a graded measure of the correlation. This effect across sensory domains should

manifest in multisensory integration. We hypothesize that weakly correlated stimuli will be less likely to bind together and thus produce smaller multisensory enhancements. As signals become more and more positively correlated, binding, and by extension multisensory perception, should be enhanced.

Chapter 2 presents a test of a first component of this hypothesis in which we measure the effect of temporal correlation strength on the process of multisensory integration. Following this, in Chapter 3 the role of the strength of temporal correlation in shaping the process of multisensory binding is tested. The role of correlation is placed in a bigger context of multisensory processes in Chapter 4. Here, we propose a developmental link between similarity and proximity whereby correlation, a measure of similarity, acts as a cue that signals across modalities belong together and thus shapes low-level multisensory processes represented by spatial and temporal proximity. A series of experiments to test predictions that follow from this frame work are proposed. The work is summarized and discussed in a broader context in Chapter 5 and future experiments are proposed to expand on the findings presented in this dissertation.

References

Alais, D., & Burr, D. (2004). Ventriloquist Effect Results from Near-Optimal Bimodal Integration. Current Biology, 14(3), 257–262. https://doi.org/10.1016/S0960-9822(04)00043-0

Alais, D., Van Der Smagt, M. J., Van Den Berg, A. V., & Van De Grind, W. A. (1998). Local and global factors affecting the coherent motion of gratings presented in multiple apertures. Vision Research, 38(11), 1581–1591. https://doi.org/10.1016/S0042-6989(97)00331-3

Amlôt, R., Walker, R., Driver, J., & Spence, C. (2003). Multimodal visual-somatosensory integration in saccade generation. Neuropsychologia, 41(1), 1–15. https://doi.org/10.1016/S0028-3932(02)00139-2

Assmann, P. F., & Summerfield, Q. (1994). The contribution of waveform interactions to the perception of concurrent vowels. The Journal of the Acoustical Society of America, 95(1), 471–484. https://doi.org/10.1121/1.408342

Barlow, H. (2001). The exploitation of regularities in the environment by the brain. Behavioral and Brain Sciences, 24, 602–607. https://doi.org/10.1017/S0140525X01000024

Battaglia, P. W., Jacobs, R. a, & Aslin, R. N. (2003). Bayesian integration of visual and auditory signals for spatial localization. Journal of the Optical Society of America A, 20(7), 1391. https://doi.org/10.1364/JOSAA.20.001391

Bergen, J. R., & Landy, M. S. (1991). Computational Models of Visual Texture Segregation. In Michael S Landy & J Anthony Movshon (Eds.), Computational Models of Visual Processing (pp. 253–271). Cambridge: MIT Press.

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. Speech Communication, 44(1–4), 5–18. https://doi.org/10.1016/J.SPECOM.2004.10.011

Bertelson, P. (1999). Chapter 14 Ventriloquism: A case of crossmodal perceptual grouping. Advances in Psychology, 129, 347–362. https://doi.org/10.1016/S0166-4115(99)80034-X

Bertelson, P., Vroomen, J., Wiegeraad, G., & de Gelder, B. (1994). Explorring the Relation Between McGurk Interference and Ventriloquism. In 3rd International Conference on Spoken Language Processing (pp. 559–562). Yokohama.

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. Nature Reviews Neuroscience. https://doi.org/10.1038/nrn3565

Bizley, J. K., Maddox, R. K., & Lee, A. K. C. (2016). Defining Auditory-Visual Objects: Behavioral Tests and Physiological Mechanisms. Trends in Neurosciences. https://doi.org/10.1016/j.tins.2015.12.007

Blake, R., & Lee, S.-H. (2000). Spatial and temporal structure jointly promote visual grouping. Invest Ophthalmol Vis Sci, 41(4), 2318B564.

Blake, R., & Lee, S.-H. (2005). The role of temporal structure in human vision. Behavioral and Cognitive Neuroscience Reviews, 4(1), 21–42. https://doi.org/10.1177/1534582305276839

Blanke, O., Slater, M., & Serino, A. (2015). Behavioral, Neural, and Computational Principles of Bodily Self-Consciousness. Neuron, 88(1), 145–166. https://doi.org/10.1016/J.NEURON.2015.09.029

Bolognini, N., Frassinetti, F., Serino, A., & Làdavas, E. (2005). "Acoustical vision" of below threshold stimuli: Interaction among spatially converging audiovisual inputs. Experimental Brain Research, 160(3), 273–282. https://doi.org/10.1007/s00221-004-2005-z

Bolognini, N., Leo, F., Passamonti, C., Stein, B. E., & Làdavas, E. (2007). Multisensory-mediated auditory localization. Perception, 36(10), 1477–1485. https://doi.org/10.1068/p5846

Botvinick, M., & Cohen, J. (1998, February 19). Rubber hands “feel” touch that eyes see [8]. Nature. Nature Publishing Group. https://doi.org/10.1038/35784

Bregman, A. S. (1990). Auditory scene analysis : the perceptual organization of sound. MIT Press.

Bregman, A. S., & Campbell, J. (1971). Primary Auditory Stream Segregation and Perception of Order in Rapid Sequences of Tones. Journal of Experimental Psychology, 89(2), 244–249. https://doi.org/10.1037/h0031163

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. Cognitive Psychology, 57(3), 153–178. https://doi.org/10.1016/J.COGPSYCH.2007.12.002

Burr, D., Silva, O., Cicchini, G. M., Banks, M. S., & Morrone, M. C. (2009). Temporal mechanisms of multimodal binding. Proceedings. Biological Sciences / The Royal Society, 276(February), 1761–1769. https://doi.org/10.1098/rspb.2008.1899

Busemeyer, J. R., & Townsend, J. T. (1992). Fundamental derivations from decision field theory. Mathematical Social Sciences, 23(3), 255–282. https://doi.org/10.1016/0165-4896(92)90043-5

Calvert, G., Spence, C., & Stein, B. E. (2004). The Handbook of Multisensory Processes. (G. Calvert, C. Spence, & B. E. Stein, Eds.), The handbook of multisensory processes. Cambridge, MA: MIT Press. https://doi.org/nicht verfügbar?

Carriere, B. N., Royal, D. W., & Wallace, M. T. (2008). Spatial heterogeneity of cortical receptive fields and its impact on multisensory interactions. Journal of Neurophysiology, 99(5), 2357–2368. https://doi.org/10.1152/jn.01386.2007

Chandrasekaran, C., Lemus, L., Trubanova, A., Gondan, M., & Ghazanfar, A. A. (2011). Monkeys and Humans Share a Common Computation for Face/Voice Integration. PLoS Computational Biology, 7(9), e1002165. https://doi.org/10.1371/journal.pcbi.1002165

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. PLoS Computational Biology, 5(7). https://doi.org/10.1371/journal.pcbi.1000436

Cherry, E. C. (1953). Some experiments on the recognition of speech, with one and with 2 ears. Journal of the Acoustical Society of America, 25(5), 975–979. https://doi.org/10.1121/1.1907229

Chuen, L., & Schutz, M. (2016). The unity assumption facilitates cross-modal binding of musical, non-speech stimuli: The role of spectral and amplitude envelope cues. Attention,

Perception, and Psychophysics, 78(5), 1512–1528. https://doi.org/10.3758/s13414-016-1088-5

Churchland, A. K., Kiani, R., & Shadlen, M. N. (2008). Decision-making with multiple alternatives. Nature Neuroscience, 11(6), 693–702. https://doi.org/10.1038/nn.2123

Colonius, H., & Diederich, A. (2004). Multisensory Interaction in Saccadic Reaction Time: A Time-Window-of-Integration Model. Journal of Cognitive Neuroscience, 16(6), 1000–1009. https://doi.org/10.1162/0898929041502733

Colonius, H., & Diederich, A. (2011). Computing an optimal time window of audiovisual integration in focused attention tasks: Illustrated by studies on effect of age and prior knowledge. Experimental Brain Research, 212(3), 327–337. https://doi.org/10.1007/s00221-011-2732-x

Convento, S., Rahman, M. S., & Yau, J. M. (2018). Selective Attention Gates the Interactive Crossmodal Coupling between Perceptual Systems. Current Biology, 28(5), 746–752.e5. https://doi.org/10.1016/j.cub.2018.01.021

Corneil, B. D., Van Wanrooij, M. M., Munoz, D. P., & Van Opstal, A. J. (2002). Auditory-visual interactions subserving goal-directed saccades in a complex scene. Journal of Neurophysiology, 88(1), 438–454. https://doi.org/10.1038/377059a0

Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration. Journal of Neuroscience, 36(38), 9888–9895. https://doi.org/10.1523/JNEUROSCI.1396-16.2016

Culling, J. F., & Darwin, C. J. (1993). Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0. The Journal of the Acoustical Society of America, 93(6), 3454–3467. https://doi.org/10.1121/1.405675

Culling, J. F., & Darwin, C. J. (1994). Perceptual and computational separation of simultaneous vowels: Cues arising from low-frequency beating. The Journal of the Acoustical Society of America, 95(3), 1559–1569. https://doi.org/10.1121/1.408543

Darwin, C. J. (1992). Listening to two Things at Once. In M.E.H. Schouten (Ed.), The Auditory Processing of Speech (pp. 133–147). Berlin: Mouton de Gruyter.

Darwin, C. J. (1997). Auditory grouping. Trends in Cognitive Sciences. https://doi.org/10.1016/S1364-6613(97)01097-8

Darwin, C. J., & Carlyon, R. P. (1995). Auditory Grouping. In Brian CJ Moore (Ed.), The Handbook of Perception and Cognition (pp. 387–424). London: Academic.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., & Aikawa, K. (1997). Concurrent vowel identification. I. Effects of relative amplitude and F0 difference. The Journal of the Acoustical Society of America, 101(5), 2839–2847. https://doi.org/10.1121/1.418517

de Lafuente, V., Jazayeri, M., & Shadlen, M. N. (2015). Representation of accumulating evidence for a decision in two parietal areas. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 35(10), 4306–4318. https://doi.org/10.1523/JNEUROSCI.2451-14.2015

Dean, E. A. (1979). Atmospheric effects on the speed of sound. El Paso.

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. Annual Review of Neuroscience, 18(1), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Diederich, A., Colonius, H., Bockhorst, D., & Tabeling, S. (2003). Visual-tactile spatial interaction in saccade generation. Experimental Brain Research. Experimentelle Hirnforschung. Experimentation Cerebrale, 148(3), 328–337. https://doi.org/10.1007/s00221-002-1302-7

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

Edwards, S. B., Ginsburgh, C. L., Henkel, C. K., & Stein, B. E. (1979). Source of subcortical projections to the superior colliculus in the cat. Journal of Comparative Neurology, 184(0021–9967 (Print)), 309–330.

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009). Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. Neuron, 61(2), 317–329. https://doi.org/10.1016/j.neuron.2008.12.005

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. Journal of Speech Language and Hearing Research, 12(2), 423. https://doi.org/10.1044/jshr.1202.423

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415(6870), 429–433. https://doi.org/10.1038/415429a

Evenson, K. M., Wells, J. S., Petersen, F. R., Danielson, B. L., Day, G. W., Barger, R. L., & Hall, J. L. (1972). Speed of Light from Direct Frequency and Wavelength Measurements of the Methane-Stabilized Laser. Physical Review Letters, 29(19), 1346–1349. https://doi.org/10.1103/PhysRevLett.29.1346

Fahle, M. (1993). Figure-ground discrimination from temporal information. Proceedings. Biological Sciences / The Royal Society, 254, 199–203. https://doi.org/10.1098/rspb.1993.0146

Feddersen, W. E., Sandel, T. T., Teas, D. C., & Jeffress, L. A. (1957). Localization of High-Frequency Tones. The Journal of the Acoustical Society of America, 29(9), 988–991. https://doi.org/10.1121/1.1909356

Foxe, J. J., Molholm, S., Del Bene, V. A., Frey, H.-P., Russo, N. N., Blanco, D., … Ross, L. A. (2015). Severe Multisensory Speech Integration Deficits in High-Functioning School-Aged Children with Autism Spectrum Disorder (ASD) and Their Resolution During Early Adolescence. Cerebral Cortex, 25(2), 298–312. https://doi.org/10.1093/cercor/bht213

Frassinetti, F., Bolognini, N., & Làdavas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. Experimental Brain Research, 147(3), 332–343. https://doi.org/10.1007/s00221-002-1262-y

Frens, M. A., Van Opstal, A. J., Van der Willigen, R. F., Opstal, a J. Van, & Willigen, R. F. Van Der. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. Perception & Psychophysics, 57(6), 802–816. https://doi.org/10.3758/BF03206796

Fujisaki, W., & Nishida, S. (2005). Temporal frequency characteristics of synchrony–asynchrony discrimination of audio-visual signals. Experimental Brain Research, 166(3–4), 455–464. https://doi.org/10.1007/s00221-005-2385-8

Ganesh, A. C., Berthommier, F., & Schwartz, J.-L. (2017). Audiovisual Binding for Speech Perception in Noise and in Aging. Language Learning. https://doi.org/10.1111/lang.12271

Gold, J., & Shadlen, M. (2007). The neural basis of decision making. Annu. Rev. Neurosci, 30(1), 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Goldberg, J. M., & Brown, P. B. (1969). Response of binaural neurons of dog superior olivary complex to dichotic tonal stimuli: some physiological mechanisms of sound localization. Journal of Neurophysiology, 32(4), 613–636. https://doi.org/10.1152/jn.1969.32.4.613

Gordon, B. (1973). Receptive Fields in Deep Layers of Cat Superior Colliculus. Journal of Neurophysiology, 36(2), 157–178.

Grant, K. W., & Seitz, P. F. 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 Pt 1), 1197–1208. https://doi.org/10.1121/1.422512

Guttman, S. E., Gilroy, L. A., & Blake, R. (2005). Hearing What the Eyes See. Psychological Science, 16(3), 228–235. https://doi.org/10.1111/j.0956-7976.2005.00808.x

Guttman, S. E., Gilroy, L. A., & Blake, R. (2007). Spatial grouping in human vision: temporal structure trumps temporal synchrony. Vision Research, 47(2), 219–230. https://doi.org/10.1016/j.visres.2006.09.012

Hairston, W. D., Laurienti, P. J., Mishra, G., Burdette, J. H., & Wallace, M. T. (2003). Multisensory enhancement of localization under conditions of induced myopia. Experimental Brain Research, 152(3), 404–408. https://doi.org/10.1007/s00221-003-1646-7

Hassenstein, B., & Reichardt, W. (1956). Systemtheoretische Analyse der Zeit-, Reihenfolgen- und Vorzeichenauswertung bei der Bewegungsperzeption des Rüsselkäfers Chlorophanus. Zeitschrift Für Naturforschung B, 11(9–10), 513–524. https://doi.org/10.1515/znb-1956-9-1004

Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. Journal of Experimental Psychology, 63(3), 289–293. https://doi.org/10.1037/h0055703

Hughes, H. C., Reuter-Lorenz, P. A., Nozawa, G., & Fendrich, R. (1994). Visual-Auditory Interactions in Sensorimotor Processing: Saccades Versus Manual Responses. Journal of Experimental Psychology: Human Perception and Performance, 20(1), 131–153. https://doi.org/10.1037/0096-1523.20.1.131

Hygge, S., Rönnberg, J., Larsby, B., & Arlinger, S. (1992). Normal-hearing and hearing-impaired subjects' ability to just follow conversation in competing speech, reversed speech, and noise backgrounds. Journal of Speech and Hearing Research, 35(1), 208–215.

Jack, C. E., & Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. Perceptual and Motor Skills, 37(3), 967–979. https://doi.org/10.2466/pms.1973.37.3.967

Jeffress, L. A. (1948). A place theory of sound localization. Journal of Comparative and Physiological Psychology, 41(1), 35–39. https://doi.org/10.1037/h0061495

Kandil, F. I., & Fahle, M. (2004). Figure-ground segregation can rely on differences in motion direction. Vision Research, 44(27), 3177–3182. https://doi.org/10.1016/j.visres.2004.07.027

Kawamura, K., & Konno, T. (1979). Various types of corticotectal neurons of cats as demonstrated by means of retrograde axonal transport of horseradish peroxidase. Experimental Brain Research, 35(1), 161–175. https://doi.org/10.1007/BF00236792

Kim, J., & Davis, C. (2004). Investigating the audio–visual speech detection advantage. Speech Communication, 44(1–4), 19–30. https://doi.org/10.1016/J.SPECOM.2004.09.008

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal Inference in Multisensory Perception. PLoS ONE, 2(9), e943. https://doi.org/10.1371/journal.pone.0000943

Kösem, A., & van Wassenhove, V. (2012). Temporal structure in audiovisual sensory selection. PLoS ONE, 7(7). https://doi.org/10.1371/journal.pone.0040936

Kramer, P., & Yantis, S. (1997). Perceptual grouping in space and time: Evidence from the Temus display. Perception & Psychophysics, 59(1), 87–99.

Krueger, J., Royal, D. W., Fister, M. C., & Wallace, M. T. (2009). Spatial receptive field organization of multisensory neurons and its impact on multisensory interactions. Hearing Research, 258, 47–54. https://doi.org/10.1016/j.heares.2009.08.003

Laming, D. R. J. (1968). Information theory of choice-reaction times. Information theory of choice-reaction times. Oxford: Academic Press. https://doi.org/10.1002/bs.3830140408

Laurienti, P. J., Kraft, R. A., Maldjian, J. A., Burdette, J. H., & Wallace, M. T. (2004). Semantic congruence is a critical factor in multisesnory behavioral performance. Experimental Brain Research, 158(4), 405–414. https://doi.org/DOI 10.1007/s00221-004-1913-2

Lee, S. H., & Blake, R. (1999). Visual form created solely from temporal structure. Science, 284(5417), 1165–1168. https://doi.org/10.1126/science.284.5417.1165

Lenggenhager, B., Tadi, T., Metzinger, T., & Blanke, O. (2007). Video ergo sum: manipulating bodily self-consciousness. Science (New York, N.Y.), 317(5841), 1096–1099. https://doi.org/10.1126/science.1143439

Lovelace, C. T., Stein, B. E., & Wallace, M. T. (2003). An irrelevant light enhances auditory detection in humans: A psychophysical analysis of multisensory integration in stimulus

detection. Cognitive Brain Research, 17(2), 447–453. https://doi.org/10.1016/S0926-6410(03)00160-5

Luce, R. D. (1986). Response times: their role in inferring elementary mental organization. New York: Oxford University Press.

Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. Neuroimage, 21(2), 725–732. https://doi.org/10.1016/j.neuroimage.2003.09.049

Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. ELife, 2015(4), 1–11. https://doi.org/10.7554/eLife.04995.001

Magnotti, J. F., Ma, W. J., & Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech. Frontiers in Psychology, 4, 798. https://doi.org/10.3389/fpsyg.2013.00798

Mast, F., Frings, C., & Spence, C. (2017). Crossmodal attentional control sets between vision and audition. Acta Psychologica, 178, 41–47. https://doi.org/10.1016/j.actpsy.2017.05.011

Mcgurk, H., & Macdonald, J. (1976). Hearing lips and seeing voices. Nature, 264(5588), 746–748. https://doi.org/10.1038/264746a0

McIlwain, J. T. (1975). Visual receptive fields and their images in superior colliculus of the cat. Journal of Neurophysiology, 38(2), 219–230. https://doi.org/10.1152/jn.1975.38.2.219

Meredith, M. A., Clemo, H. R., & Stein, B. E. (1991). Somatotopic component of the multisensory map in the deep laminae of the cat superior colliculus. Journal of Comparative Neurology, 312(3), 353–370. https://doi.org/10.1002/cne.903120304

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. The Journal of Neuroscience, 7(10), 3215–3229. https://doi.org/citeulike-article-id:409430

Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. Science, 221(4608), 389–391. https://doi.org/10.1126/science.6867718

Meredith, M. A., & Stein, B. E. (1986a). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. Brain Research, 365(2), 350–354. https://doi.org/10.1016/0006-8993(86)91648-3

Meredith, M. A., & Stein, B. E. (1986b). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. Journal of Neurophysiology, 56(3), 640–662. https://doi.org/citeulike-article-id:844215

Meredith, M. A., & Stein, B. E. (1990). The Visuotopic Component of the Multisensory Map in the Deep Laminae of the Cat Superior Colliculus. The Journal of Neuroscience, 70(11), 3727–3742.

Meredith, M. A., & Stein, B. E. (1996). Spatial determinants of multisensory integration in cat superior colliculus neurons. Journal of Neurophysiology, 75(5), 1843–1857. https://doi.org/citeulike-article-id:411558

Micheyl, C., Hanson, C., Demany, L., Shamma, S., & Oxenham, A. J. (2013). Auditory stream segregation for alternating and synchronous tones. Journal of Experimental Psychology. Human Perception and Performance, 39(6), 1568–1580. https://doi.org/10.1037/a0032241

Middlebrooks, J. C., & Knudsen, E. I. (1984). A neural code for auditory space in the cat's superior colliculus. The Journal of Neurosci Ence, 4(10), 2621–2634.

Miller, G. A. (1947). The Masking of Speech. Psychological Bulletin, 51(4), 327–358. https://doi.org/10.1037/h0061470

Miller, G. A., & Heise, G. A. (1950). The Trill Threshold. The Journal of the Acoustical Society of America, 22(5), 637–638. https://doi.org/10.1121/1.1906663

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. Perception & Psychophysics, 58(3), 351–362. https://doi.org/10.3758/BF03206811

Murray, M. M., Molholm, S., Michel, C. M., Heslenfeld, D. J., Ritter, W., Javitt, D. C., … Foxe, J. J. (2005). Grabbing your ear: Rapid auditory-somatosensory multisensory interactions in low-level sensory cortices are not constrained by stimulus alignment. Cerebral Cortex, 15(7), 963–974. https://doi.org/10.1093/cercor/bhh197

Murray, M., & Wallace, M. (Eds.). (2012). The Neural Bases of Multisensory Processes. Boca Raton: CRC Press. https://doi.org/10.1201/b11092

Nahorna, O., Berthommier, F., & Schwartz, J.-L. (2012). Binding and unbinding the auditory and visual streams in the McGurk effect. The Journal of the Acoustical Society of America, 132(2), 1061–1077. https://doi.org/10.1121/1.4728187

Newman, E. A., & Hartline, P. H. (1981). Integration of visual and infrared information in bimodal neurons in the rattlesnake optic tectum. Science (New York, N.Y.), 213(4509), 789–791. https://doi.org/10.1126/SCIENCE.7256281

Nidiffer, A. R., Stevenson, R. A., Krueger Fister, J., Barnett, Z. P., & Wallace, M. T. (2016). Interactions between space and effectiveness in human multisensory performance. Neuropsychologia, 88, 83–91. https://doi.org/10.1016/j.neuropsychologia.2016.01.031

Noel, J.P., Blanke, O., & Serino, A. (2018). From multisensory integration in peripersonal space to bodily self-consciousness: from statistical regularities to statistical inference. Annals of the New York Academy of Sciences. https://doi.org/10.1111/nyas.13867

Noorden, L. P. A. S. Van. (1977). Minimum differences of level and frequency for perceptual fission of tone sequences ABAB*. J. Acoust. Soc. Am., 61(4), 1041–1045. https://doi.org/10.1121/1.381388

O'Sullivan, J. A., Shamma, S. A., & Lalor, E. C. (2015). Evidence for Neural Computations of Temporal Coherence in an Auditory Scene and Their Enhancement during Active Listening. Journal of Neuroscience, 35(18), 7256–7263. https://doi.org/10.1523/JNEUROSCI.4973-14.2015

Ohshiro, T., Angelaki, D. E., & DeAngelis, G. C. (2011). A normalization model of multisensory integration. Nature Neuroscience, 14(6), 775–782. https://doi.org/10.1038/nn.2815

Ohzawa, I. (1998). Mechanisms of stereoscopic vision: The disparity energy model. Current Opinion in Neurobiology, 8(4), 509–515. https://doi.org/10.1016/S0959-4388(98)80039-1

Otto, T. U., Dassy, B., & Mamassian, P. (2013). Principles of Multisensory Behavior. Journal of Neuroscience, 33(17), 7463–7474. https://doi.org/10.1523/JNEUROSCI.4678-12.2013

Parise, C. V., & Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. Nature Communications, 7(12), 364. https://doi.org/10.1038/ncomms11543

Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. Current Biology, 22(1), 46–49. https://doi.org/10.1016/j.cub.2011.11.039

Parise, C. V, Harrar, V., Ernst, M. O., & Spence, C. (2013). Cross-correlation between Auditory and Visual Signals Promotes Multisensory Integration. Multisensory Research, 26, 1–10. https://doi.org/10.1163/22134808-00002417

Perrault, T. J. (2003). Neuron-Specific Response Characteristics Predict the Magnitude of Multisensory Integration. Journal of Neurophysiology, 90(6), 4022–4026. https://doi.org/10.1152/jn.00494.2003

Perrault, T. J. (2005). Superior Colliculus Neurons Use Distinct Operational Modes in the Integration of Multisensory Stimuli. Journal of Neurophysiology, 93(5), 2575–2586. https://doi.org/10.1152/jn.00926.2004

Rach, S., Diederich, A., & Colonius, H. (2011). On quantifying multisensory interaction effects in reaction time and detection rate. Psychological Research, 75(2), 77–94. https://doi.org/10.1007/s00426-010-0289-0

Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. Neural Computation, 20(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., & Rouder, J. N. (1998). Modeling Response Times for Two-Choice Decisions. Psychological Science, 9(5), 347–356. https://doi.org/10.1111/1467-9280.00067

Ratcliff, R., & Smith, P. L. (2004). A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. Psychological Review, 111(2), 333–367. https://doi.org/10.1037/0033-295X.111.2.333

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. Diffusion Decision Model: Current Issues and History, 20 Trends in Cognitive Sciences § (2016). https://doi.org/10.1016/j.tics.2016.01.007

Rikhye, R. V, & Sur, M. (2015). Spatial Correlations in Natural Scenes Modulate Response Reliability in Mouse Visual Cortex. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 35(43), 14661–14680. https://doi.org/10.1523/JNEUROSCI.1660-15.2015

Roitman, J. D., & Shadlen, M. N. (2002). Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 22(21), 9475–9489.

Romano, P. E., & Berlow, S. (1973). Lipreading Performance As Related to Visual Acuity. American Journal of Ophthalmology, 75(1), 136–141. https://doi.org/10.1016/0002-9394(73)90664-8

Ross, L. A., Del Bene, V. A., Molholm, S., Frey, H.-P., & Foxe, J. J. (2015). Sex differences in multisensory speech processing in both typically developing children and those on the autism spectrum. Frontiers in Neuroscience, 9, 185. https://doi.org/10.3389/fnins.2015.00185

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cerebral Cortex, 17(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

Ruderman, D. L. (1994). The statistics of natural images. Network: Computation in Neural Systems, 5(4), 517–548. https://doi.org/10.1088/0954-898X_5_4_006

Schall, S., Quigley, C., Onat, S., & König, P. (2009). Visual stimulus locking of EEG is modulated by temporal congruency of auditory stimuli. Experimental Brain Research, 198(2–3), 137–151. https://doi.org/10.1007/s00221-009-1867-5

Sekuler, A. B., & Bennett, P. J. (2001). Generalized common fate: grouping by common luminance changes. Psychological Science, 12(6), 437–444. https://doi.org/10.1111/1467-9280.00382

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. Nature, 385(6614), 308. https://doi.org/10.1038/385308a0

Senkowski, D., Talsma, D., Grigutsch, M., Herrmann, C. S., & Woldorff, M. G. (2007). Good times for multisensory integration: Effects of the precision of temporal synchrony as revealed by gamma-band oscillations. Neuropsychologia, 45(3), 561–571. https://doi.org/10.1016/j.neuropsychologia.2006.01.013

Shamma, S., Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., Pressnitzer, D., … Xu, Y. (2013). Temporal coherence and the streaming of complex sounds. Advances in Experimental Medicine and Biology, 787, 535–543. https://doi.org/10.1007/978-1-4614-1590-9-59

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. Nature, 408(December), 2000. https://doi.org/10.1038/35048669

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. Trends in Cognitive Sciences, 12(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural Image Statistics and Neural Representation. Annual Review of Neuroscience, 24, 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193

Stanford, T. R., & Stein, B. E. (2007). Superadditivity in multisensory integration: putting the computation in context. Neuroreport, 18(8), 787–792. https://doi.org/10.1097/WNR.0b013e3280c1e315

Stein, B. E. (1978). Development and organization of multimodal representation in cat superior colliculus. Federation Proceedings, 37(9), 2240–2245.

Stein, B. E., & Arigbede, M. O. (1972). Unimodal and multimodal response properties of neurons in the cat's superior colliculus. Experimental Neurology, 36(1), 179–196. https://doi.org/10.1016/0014-4886(72)90145-8

Stein, B. E., Huneycutt, S. W., & Alex Meredith, M. (1988). Neurons and behavior: the same rules of multisensory integration apply. Brain Research, 448(2), 355–358. https://doi.org/10.1016/0006-8993(88)91276-0

Stein, B. E., Magalhaes-Castro, B., & Kruger, L. (1976). Relationship Between Visual and Tactile Representations in Cat Superior Colliculus. JOURNAL OF NEUROPHYSIOLOGY Vnl, 39(2).

Stevenson, R. A., Bushmakin, M., Kim, S., Wallace, M. T., Puce, A., & James, T. W. (2012). Inverse effectiveness and multisensory interactions in visual event-related potentials with audiovisual speech. Brain Topography, 25(3), 308–326. https://doi.org/10.1007/s10548-012-0220-7

Stevenson, R. A., Ghose, D., Fister, J. K., Sarko, D. K., Altieri, N. A., Nidiffer, A. R., … Wallace, M. T. (2014). Identifying and Quantifying Multisensory Integration: A Tutorial Review. Brain Topography, 27(6), 707–730. https://doi.org/10.1007/s10548-014-0365-7

Stevenson, R. A., & James, T. W. (2009). Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. NeuroImage, 44(3), 1210–1223. https://doi.org/10.1016/j.neuroimage.2008.09.034

Stone, M. (1960). Models for choice-reaction time. Psychometrika, 25(3), 251–260. https://doi.org/10.1007/BF02289729

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. The Journal of the Acoustical Society of America, 26(2), 212–215. https://doi.org/10.1121/1.1907309

Teki, S., Chait, M., Kumar, S., Shamma, S., & Griffiths, T. D. (2013). Segregation of complex acoustic scenes based on temporal coherence. ELife, 2, e00699. https://doi.org/10.7554/eLife.00699

Teki, S., Chait, M., Kumar, S., von Kriegstein, K., & Griffiths, T. D. (2011). Brain bases for auditory stimulus-driven figure-ground segregation. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 31(1), 164–171. https://doi.org/10.1523/JNEUROSCI.3788-10.2011

Thurlow, W. R., & Jack, C. E. (1973). Certain determinants of the "ventriloquism effect". Perceptual and Motor Skills, 36, 1171–1184. https://doi.org/10.2466/pms.1973.36.3c.1171

Todd, J. W. (1912). Reaction to Multiple Stimuli. Archives of Psychology, 25, 1-65.

Treisman, A. (1996). The Binding Problem. Current Opinion in Neurobiology, 6, 171–178.

Treisman, A. (1998). Feature binding, attention, and object perception. Essent. Sources Sci. Study Conscious., 8, 226. https://doi.org/10.1016/j.ejpn.2004.03.003

Treisman, A. (1999). Solutions to the binding problem: Progress through controversy and convergence. Neuron, 24(1), 105–110. https://doi.org/10.1016/S0896-6273(00)80826-0

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. Cognitive Psychology, 12(1), 97–136.

Tsakiris, M., & Haggard, P. (2005). The Rubber Hand Illusion Revisited: Visuotactile Integration and Self-Attribution. Journal of Experimental Psychology: Human Perception and Performance, 31(1), 80–91. https://doi.org/10.1037/0096-1523.31.1.80

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. Psychological Review, 108(3), 550–592. https://doi.org/10.1037/0033-295X.108.3.550

Vandekerckhove, J. J., Tuerlinckx, F., Bulletin, P., Vandekerckhove, J. J., & Tuerlinckx, F. (2007). Fitting the Ratcliff diffusion model to experimental data. Psychonomic Bulletin & Review, 14(6), 1011–1026. https://doi.org/10.3758/BF03193087

Vatakis, A., & Spence, C. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. Perception & Psychophysics, 69(5), 744–756. https://doi.org/10.3758/BF03193776

Venezia, J. H., Thurman, S. M., Matchin, W., George, S. E., & Hickok, G. (2016). Timing in audiovisual speech perception: A mini review and new psychophysical data. Attention, Perception, & Psychophysics, 78(2), 583–601. https://doi.org/10.3758/s13414-015-1026-y

Voss, A., Nagler, M., & Lerche, V. (2013). Diffusion models in experimental psychology: A practical introduction. Experimental Psychology, 60(6), 385–402. https://doi.org/10.1027/1618-3169/a000218

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: an empirical validation. Memory & Cognition, 32(7), 1206–1220. https://doi.org/10.3758/BF03196893

Wagemans, J., Elder, J. H., Kubovy, M., Palmer, S. E., Peterson, M. A., Singh, M., & von der Heydt, R. (2012). A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. Psychological Bulletin, 138(6), 1172–1217. https://doi.org/10.1037/a0029333

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. European Journal of Cognitive Psychology, 21(5), 641–671. https://doi.org/10.1080/09541440802205067

Wagenmakers, E.-J., Van Der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. Psychonomic Bulletin & Review, 14(1), 3–22. https://doi.org/10.3758/BF03194023

Wallace, J. M., & Scott-Samuel, N. E. (2007). Spatial versus temporal grouping in a modified Ternus display. Vision Research, 47(17), 2353–2366. https://doi.org/10.1016/J.VISRES.2007.05.016

Wallace, M. T. (2004). The development of multisensory processes. Cognitive Processing, 5(2), 69–83. https://doi.org/10.1007/s10339-004-0017-z

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. Experimental Brain Research, 158(2), 252–258. https://doi.org/10.1007/s00221-004-1899-9

Wallace, M. T., & Stein, B. E. (2007). Early experience determines how the senses will interact. Journal of Neurophysiology, 97(1), 921–926. https://doi.org/10.1152/jn.00497.2006

Wallace, M. T., Wilkinson, L. K., & Stein, B. E. (1996). Representation and integration of multiple sensory inputs in primate superior colliculus. Journal of Neurophysiology, 76(2), 1246–1266.

Wertheimer, M. (1923). Untersuchungen zur Lehre yon der Gestalt II. Psychologische Forschung, 4, 301–350.

Wertheimer, M. (1938). Laws of organization in perceptual forms. In A source book of Gestalt psychology. (pp. 71–88). London: Kegan Paul, Trench, Trubner & Company. https://doi.org/10.1037/11496-005

Woods, K. J. P., & Mcdermott Correspondence, J. H. (2015). Attentive Tracking of Sound Sources. Current Biology. https://doi.org/10.1016/j.cub.2015.07.043

Zhou, B., Zhang, J. X., Tan, L. H., & Han, S. (2004). Spatial congruence in working memory: an ERP study. Neuroreport, 15(18), 2795–2799.

Zion Golumbic, E. M., Poeppel, D., & Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective. Brain and Language, 122(3), 151–161. https://doi.org/10.1016/j.bandl.2011.12.010

Chapter 2. Multisensory perception reflects individual differences in processing temporal correlations

Introduction

Our environment provides us with an enormous amount of information that is encoded by multiple sensory modalities. One of the fundamental tasks of the brain is to construct an accurate and unified representation of our environment from this rich array of sensory signals. To accomplish this, the brain must decide which signals arise from a common source. For example, during conversation among a group of individuals, listeners can group appropriate words from the same voice and further associate voices with the appropriate speakers, a process greatly facilitated by the availability of both audible and visible cues (Stein, 2012). Benefits that are associated with the presence of multisensory signals include increased detection (Frassinetti et al., 2002) and localization accuracy (Odegaard, Wozny, & Shams, 2015), improved speech intelligibility (Sumby & Pollack, 1954) and speeding of reaction times (Frens & Van Opstal, 1995; Hershenson, 1962).

A number of principles have been proposed that relate the spatial and temporal proximity of multisensory signals and the manner in which these enhance neural and behavioral responses (Bolognini et al., 2005; Frassinetti et al., 2002; Meredith et al., 1987; Meredith & Stein, 1986a). These factors have also been related to our brain's determination that multisensory signals come from the same source (Körding et al., 2007; Magnotti et al., 2013). In addition to these principles, it has been demonstrated that the temporal similarity (i.e., correlation) of these signals are also important in shaping our multisensory perception and causal inference (Chuen & Schutz, 2016; Jack & Thurlow, 1973; Parise et al., 2013; Vatakis & Spence, 2007). Indeed, temporal similarity is a hallmark feature of signals originating from the same source, such as the voice and mouth movements of a speaker (Chandrasekaran et al., 2009), and has been shown to be a robust cue for the binding of unisensory (Blake & Lee, 2005; Elhilali et al., 2009) and multisensory (Bizley et al., 2016; Maddox et al., 2015; Munhall et al., 1996; Parise et al., 2012) features. Observers can utilize these temporal correlations in multisensory signals to enhance behavioral performance (Grant & Seitz, 2000; Maddox et al., 2015; Parise & Ernst, 2016).

Although we know that temporal correlation between unisensory signals leads to a unified multisensory percept and enhancement of multisensory behaviors, it is not known

whether, and if so how, multisensory behavioral performance varies with the strength of the correlation. We hypothesize that audiovisual temporal correlation provides sensory evidence for multisensory decisions that is proportional to the sign and magnitude of the correlation. Further we hypothesize that these graded changes in sensory evidence will result in corresponding changes in multisensory behavior. To test these hypotheses, we presented participants with audiovisual signals with barely detectable (i.e., near threshold) amplitude modulation (AM). While manipulating the temporal correlation between the auditory and visual signals, we measured how observers' ability to detect these fluctuations changed with changes in stimulus correlation. We propose a mechanism—analogous to a phase shift—that approximates relative differences in unisensory temporal processing and that accounts for individual differences in behavioral results. Finally, we employed drift-diffusion modelling to test whether multisensory behavioral performance is better approximated by absolute stimulus correlation or by the adjusted correlations that account for this phase shift.

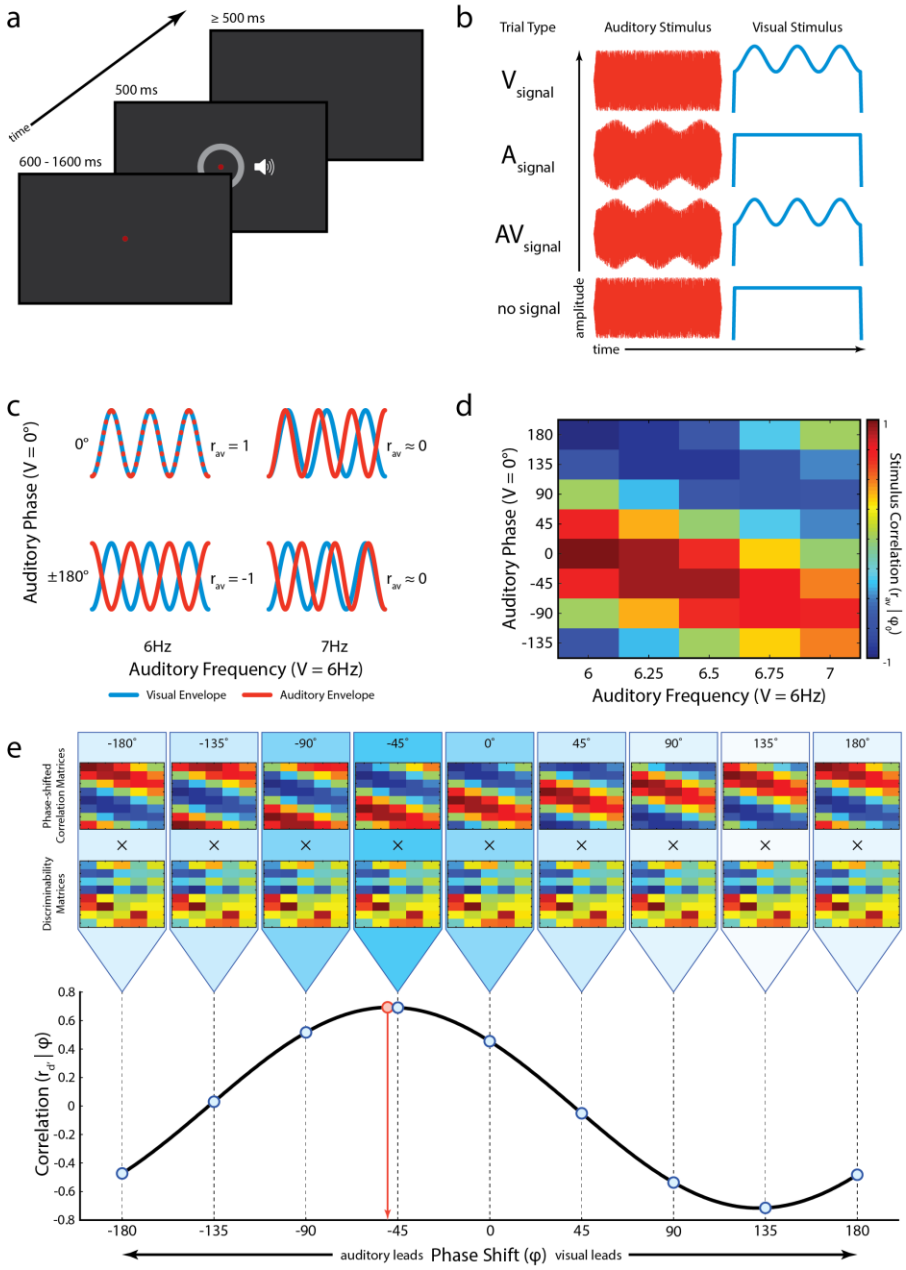
Results

Participants (n=12) detected near-threshold amplitude modulated (AM) audiovisual stimuli (Figure 2.1a-b). The temporal correlation of the AM signals was manipulated by

systematic changes in the phase and frequency relationship of the auditory and visual pairs (Figure 2.1c). Our central hypothesis was that multisensory behavioral performance would improve commensurate with increasing temporal similarity between the paired audiovisual stimuli (i.e., as correlation become more positive). To examine the potential dependence of behavior on stimulus correlation, a discriminability (*d'*) matrix and a reaction time (RT) matrix for each participant was constructed and related to the stimulus correlation (r_{av}) matrix (Δ frequency × Δ phase; Figure 2.1d).

Figure 2.1 (next page): Amplitude modulation detection task. ***(a)*** *Schematic representation of a single trial. Each trial began with the illumination of a fixation target. After a variable wait period, simultaneously presented auditory and visual stimuli appeared (see b). Participants indicated the presence or absence of amplitude modulation with a button press.* ***(b)*** *Auditory and visual stimuli were always present, but modulation was presented in auditory stimuli only (A_{signal} trial), visual stimuli only (V_{signal} trial), audiovisual stimuli (AV_{signal} trial), or neither stimulus (no signal, catch trial).* ***(c)*** *During audiovisual presentations, the frequency and phase of auditory modulation could be independently manipulated yielding a range of audiovisual correlations (r_{av}). Correlations were computed using the time series of the auditory and visual envelopes. Note that the visual envelope is always constant while the auditory envelope is varied. Four conditions out of forty are shown for illustration.* ***(d)*** *Stimulus Correlation Matrix ($r_{av} \mid \varphi_0$). All forty AV stimulus conditions are shown organized according to Δ frequency × Δ phase. Colors represent the correlation values of audiovisual stimuli across the different frequencies and phases presented where each color box represents one condition. In the task structure, there were 21 unique audiovisual stimulus correlations.* ***(e)*** *In order to account for phase shifts in individual participant data, the values in the stimulus correlation matrix ($r_{av} \mid \varphi_0$) were correlated to each participant's discriminability matrix ($r_{d'}$). In the top panel, a series of correlation matrices are shown in which a phase lag, φ_i, was applied to auditory (positive shifts) or visual (negative shifts)*

before correlations were computed, ($r_{av} \mid \varphi_i$). A total of 360 correlation matrices were correlated with the participant's discriminability matrix (r_d; middle panel, nine examples shown), approximating a cross correlation. In the bottom panel, each of the 360 correlations ($r_{d'}$) were plotted against phase lag [($r_{d'} \mid \varphi$); black line, examples shown by blue dots]. This function was fit to a sine wave and the phase of that fit was extracted (φ'; red dot and arrow) and was taken to represent a participant's individual phase shift. The stimulus correlation at that phase shift ($r_{av} \mid \varphi'$) was taken to represent a participant's "internal" correlation matrix.



While RTs did not show a robust systematic pattern (likely a result of the near-threshold nature of the stimuli, although see Table 2.1 for RT correlations in some participants), discriminability had a discernible pattern that reflected the nature of the stimulus correlations. In eight of 12 participants, discriminability was significantly correlated with stimulus correlation (Figure 2.2a). However, upon visual inspection, the discriminability matrices of two of the remaining four participants mirrored the stimulus correlation matrix but with an apparent shift along the Δ phase dimension (see Figure 2.2a-b, middle panels for one example). In fact, this phase shift appeared to be present in most participants to varying degrees and seemed to occur evenly across Δ frequency for each participant (i.e., any shift along the phase dimension was present for all auditory frequencies presented). We therefore hypothesized that this phase shift reflects an internal transformation that alters the relationship between stimulus correlation and behavior (and that is likely driven by individual differences in unisensory temporal processing).

Table 2.1: Reaction time (RT), hit rate (HR) and discriminability (d') correlations

	RT		HR		d'	
Ptc.	R	p	R	p	R	p
1	-0.24	0.14	0.77	4.5e-9	0.76	1.2e-8
2	-0.59	5.3e-5	0.91	4.0e-16	0.91	1.1e-15
3	-0.34	0.031	0.50	0.001	0.49	0.001
4	0.05	0.78	0.68	8.7e-7	0.68	1.2e-6
5	-0.25	0.14	0.14	0.36	0.12	0.45
6	-0.44	0.0042	0.70	4.3e-7	0.68	1.2e-06
7	-0.21	0.19	0.71	2.7e-7	0.70	4.7e-7
8	-0.54	3.3e-4	0.89	1.2e-14	0.88	5.8e-14
9	0.05	0.75	0.87	3.1e-13	0.86	8.0e-13
10	-0.39	0.014	0.57	1.3e-4	0.57	1.4e-4
11	-0.41	0.01	0.70	4.9e-7	0.62	2.2e-5
12	0.08	0.65	0.19	0.22	0.20	0.21

Nonsignificant correlations are in red.

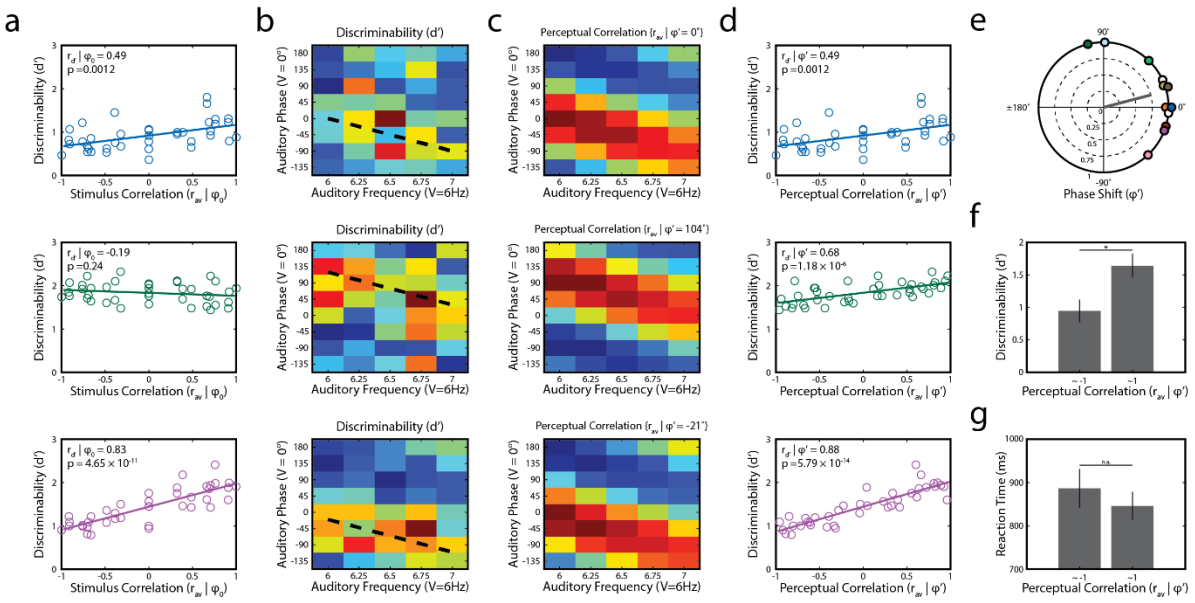


Figure 2.2: Individual participant data examples. ***(a)*** *Behavioral dependence on stimulus correlation ($r_{d'} \mid \varphi_0$) of three example participants. For parts a–d, each row represents a single participant. Participants are represented by the same color across the figures.* ***(b)*** *Discriminability matrices from three*

participants show how changes in phase (Δ phase; y-axis) and frequency (Δ frequency; x-axis) impact the ability to detect amplitude modulation (discriminability). Diagonal dashed lines represent the computed individual phase shifts (φ' = 0°, -104°, and 21°) corresponding to the approximate middle of the diagonal of positive correlations in (c). Color values have been scaled separately and range from the lowest to highest value (shown in panel a) for each participant. ***c****: Phase shifted ("perceived") correlation matrices ($r_{av} \mid \varphi'$) from each participant shown in (a). Note the strong positive (upward) shift in the second example participant and the moderate negative (downward) shift in the third example participant, relative to Figure 2.1d.* ***d****: Behavioral dependence ($r_{d'} \mid \varphi'$) on perceived stimulus correlation. Each participant shows a strong positive relationship between perceived stimulus correlation and detection behavior (i.e., discriminability). Note that the data in the middle panel was not significantly correlated to physical stimulus correlation (a) but reached significance when accounting for the phase shift. Further, note that the top participant shown did not differ between the two measures due to the lack of observed phase shift. Colors follow the convention described in Figure 2.2a.* ***e****: Distribution of observed phase shifts from all participants and mean resultant vector. Phase shifts were concentrated around the mean (14.7°, not uniform across phase). Phases were shifted toward positive values (visual leading) but were not significantly different from zero.* ***f-g****: Accuracy and reaction time effects between stimuli with the strongest negative and positive perceptual correlations. Strong positive correlation improves detection performance but has no impact on reaction times.*

Individuals display unique characteristics for auditory and visual temporal processing

We sought to measure and account for these individualized phase shifts. We modeled this by applying a phase shift to every condition in one of the unisensory modalities before recalculating a stimulus correlation matrix. We then measured the correlation between the discriminability matrix and a series of stimulus correlation matrices computed with phase shifts

ranging from -180° to +180° (Figure 2.1e; more detail in methods). We then fit this series of correlations to a sine wave. Due to the cyclical nature of the stimulus correlation matrix along the Δ phase dimension, we expected the correlations to be in the shape of a sine wave. As expected, each participant's phase-shifted correlations were well fit ($r^2 = 0.99999 \pm 2.9 \times 10^{-5}$). Another expectation is for these functions to have a period of 360° and to be centered about zero. Indeed, we found no evidence that their period was different from 360° (period = 360.06 ± 0.71; t_{11} = 0.2702, $p = 0.79$) or that their center was different from 0 (center = $1.3 \times 10^{-4} \pm 5.4 \times 10^{-4}$; t_{11} = 0.783, p = 0.45). Therefore, we calculated a participant's phase shift from these fits and then recomputed a unique correlation matrix for each participant using their individual phase shift.

As a test of the validity of phase shift, the pattern of data in the discriminability matrix should mirror the pattern of the phase-shifted stimulus matrix. This would manifest in several ways. First, if the perceived correlation matrix accounts for the data, large changes in the data should be accounted for by changes in the correlations. Therefore, the residual errors between the two measures should be very small relative to the data and centered on zero. Discriminability values (Figure 2.2b) were significantly above zero (d' = 1.30 ± 0.66; $z = 43.579$, $p = 8.75 \times 10^{-169}$). Subtracting the predicted d', which was computed from the perceptual correlation matrices (see

Methods; Figure 2.2c), from the observed *d'*, yielded residual errors which were substantially smaller and less variable compared to *d'* (mean error = 0.018 ± 0.33). Indeed, these residual errors did not differ significantly from zero (z = 1.210, p = 0.23). Second, we might question the validity of these phase shifts if the data do not mirror perceptual correlations equally for each Δ frequency (e.g., if the diagonal of high *d'* values in the discriminability matrix has a slope that doesn't match the slope of high *d'* values in the predicted discriminability matrix). To quantify this, we examined residual errors across different frequencies for any systematic changes. Residual error magnitude and variability showed no linear relationships across Δ frequency in any participant (magnitude: slopes = 0.047 ± 0.10, all p > 0.12; variability: slopes = 0.016 ± 0.07, all p > 0.09). Thus, phase shifts appear to be valid and systematic shifts in the phase dimension are independent of frequency. As such, the correlation matrices constructed using each participant's unique phase shift could be envisioned to represent the internal ("perceived") correlations of the external stimuli, accounting for differences in latency of sensory processing between the auditory and visual systems.

These perceptual correlations were used when determining the relationship between discriminability and stimulus correlation ($r_{d'}$; Figure 2.2d). The sine wave fits between phase shift

and correlation revealed the degree of participant audiovisual phase shift (φ'; Figure 2.2e). Phase shifts were not significantly different from 0 across participants but favored a visual leading shift (mean φ' = 14.7 ± 39.7°; 95% CI [42.2° -12.9°]). The distribution of shifts was concentrated about the mean as indexed by the mean resultant vector length (Figure 2.2e; MRVL = 0.76; z = 11.998, p = 1.5×10^{-8}, Rayleigh Test). To further probe the validity of these phase shifts, we tested whether the magnitude of phase shift was correlated to the strength of the relationship between behavior and stimulus correlation. Smaller correlations associated with larger phase shifts might suggest that the repeated phase shift approach returned spurious correlations. We found no evidence of such a relationship (rho = 0.25, p = 0.68).

Amplitude modulation discriminability varies with perceived stimulus correlation

Previously, it has been shown that strongly correlated multisensory stimuli provided behavioral and perceptual benefits relative to unisensory performance whereas poorly correlated stimuli fail to provide such benefits (Maddox et al., 2015; Parise et al., 2012, 2013). To examine whether a similar relationship is evident for the current task, we compared the discriminability of stimuli that had the highest and lowest correlation for each participant. We found that discriminability of audiovisual signals with the highest correlations was better than for

audiovisual signals with the lowest correlations (Figure 2.2f; $t_{11} = 4.312$, $p = 0.0062$, corrected). In contrast, reaction times failed to differ between correlated signals and uncorrelated signals (Figure 2.2g; $t_{11} = 3.384$, $p = 0.19$, corrected).

Our focus of the current study was to show that multisensory behavior varied proportionally with stimulus correlation. Although we demonstrated above that this relationship was robust in most participants (Figure 2.2a), there was evidence that this effect was weakened—and in some participants absent—due to significant individual variability. Thus, it still remained unclear whether phase shift plays an important role in this relationship. To test this, we measured the association between perceived stimulus correlation and discriminability ($r_{d'} \mid \varphi$'; Figure 2.3a). These correlations were significant in ten out of the twelve participants—two participants more than when not accounting for phase shift. This proportion, 10/12, was significantly greater than expected based on random effects ($p = 0.019$, binomial test). The significant correlations revealed effects that were very strong (Figure 2.3b). The correlation values for discriminability and hit rate are presented for each participant in Table 1.

Because we varied auditory parameters while holding visual parameters stationary, it remained possible that participant performance was driven by cues in the auditory modality

rather than by audiovisual correlation. In order to rule out that the effects reported here may be a result of unisensory auditory performance, four participants returned and completed a new experiment where visual modulation depth was set to zero while auditory depth was set at their individual threshold. We correlated auditory performance with AM frequency, AM phase, and perceived stimulus correlation. These data are summarized in Table 2. None of these correlations were significant in any of the four participants, even when computing perceived correlations based on potential phase shifts in auditory or audiovisual performance data. Moreover, phase shifts obtained from the auditory data were very different than those obtained from audiovisual data. As a final check, we subtracted the auditory data from the audiovisual data and measured the phase shift and resultant correlation. All four participants showed a significant correlation and the obtained phase shifts corresponded well to the phase shifts obtained from audiovisual data. These results suggesting that audiovisual correlations—rather than auditory modulations—are responsible for the behavioral effects presented here.

Table 2.2. Results of auditory only experiments

Ptc.		Stimulus Correlation Effect on AV	A[1]	A[2]	AV-A	Frequency Effect on A	Phase Effect on A
1	Shift	47	99	-	35	-	-
	R	0.76	0.29	0.09	0.55	-0.01	0.20
	p	1.2e-8	0.067	0.59	2.7e-4	0.93	0.43
6	Shift	-89	100	-	-89	-	-
	R	0.68	0.2	-0.18	0.6	-0.04	0.26
	p	1.2e-6	0.21	0.22	4.5e-5	0.83	0.25
8	Shift	21	-152	-	23	-	-
	R	0.88	0.27	-0.27	0.81	0.03	0.31
	p	5.8e-14	0.097	0.091	1.9e-10	0.85	0.14
10	Shift	-19	-18	-	-17	-	-
	R	0.57	0.05	0.04	0.35	-0.15	0.11
	p	1.4e-4	0.76	0.82	0.026	0.35	0.76

[1] Correlations were unconstrained and reflect best possible correlations. [2] Correlations were constrained by audiovisual phase shift. Nonsignificant correlations are in red.

When accounting for phase shift, the strength of these behavioral effects increased in all participants ($\Delta r_{d'} = 0.19 \pm 0.29$) and the increase was more pronounced in participants with larger magnitude phase shifts (Figure 2.3e, $a_{obs} = 0.706$). Due to the nature of the phase-shift fitting process, simulated random data (details can be found in methods) produces correlational improvement that peaks at ± 180° ($a_{null} = 0.205$, 95% CI [0.144 0.271]). Nonetheless, the observed effect was significantly larger than what would be expected by these random effects ($z = 21.80$, $p = 2.1\times10^{-93}$). Lastly, in contrast to the concentrated distribution of observed phase shifts

(Figure 2.2e), the distribution of simulated phase shifts was not significantly different from uniform (Figure 2.3f; MRVL = 0.04; z = 2.08, p = 0.125, Rayleigh Test). These findings provide strong support for the notion that phase shift reflects an important transformation between stimulus correlation as it occurs in the environment and how it manifests in perceptual performance.

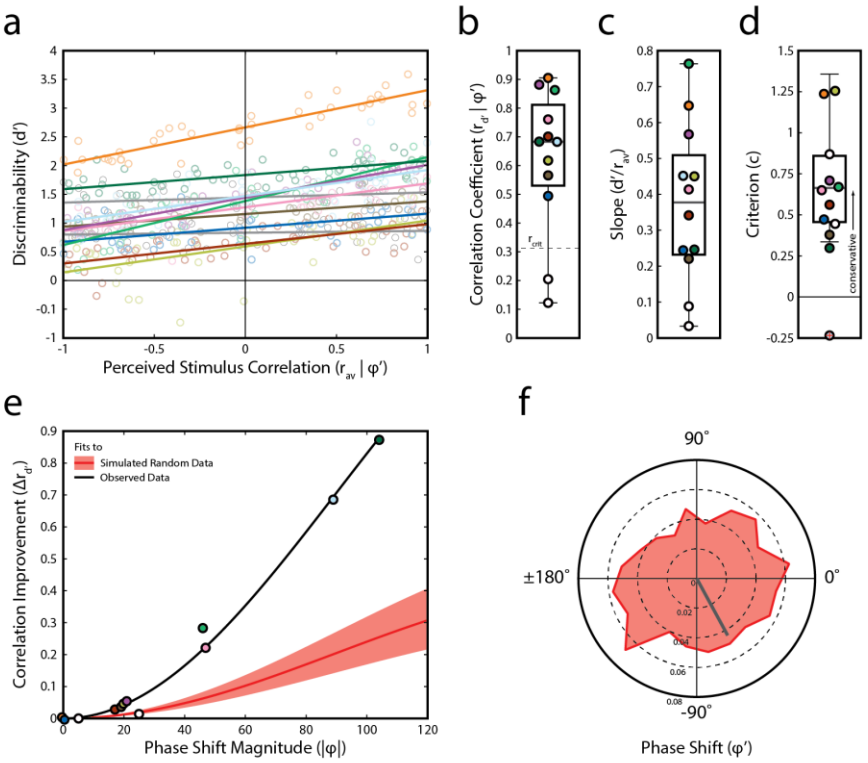


Figure 2.3 (previous page): Behavioral results. *(a) Behavioral dependence on perceived stimulus correlation across all participants (same as Figure 2.2d). Behavioral performance in 10 of 12 participants was driven by stimulus correlation. Non-significantly correlated data are represented in grey. Significantly correlated data is depicted in color. As in Figure 2.2, each participant retains the same color across the figures.* ***(b)*** *Correlation coefficients ($r_{d'} | \varphi'$) for each participant. The critical value of the correlation coefficient is denoted by a dashed line.* ***(c)*** *Slope of linear data fits shown in (a) for each*

participant. ***(d)*** *Criterion for each participant. Each participant but one held a conservative criterion indicating that participants weren't biased toward responding "yes."* ***(e)*** *Improvement in correlation ($\Delta r_d'$) is associated with phase shift and the effect is larger than expected by chance. Red line and shaded region represent the average fit and 95% confidence bands of random data from the Monte Carlo simulation fit to a sine wave. The black line represents the fit of the observed data to the sine wave. The amplitude of the data fit sine wave was significantly larger than expected by chance.* ***(f)*** *Distribution of phase shifts and the corresponding MRVL obtained from the Monte Carlo simulation. In contrast to observed data shown in Figure 2.2e, these phase shifts are not significantly concentrated about the circular mean. Note the scale difference in the radial axis between Figure 2.2e and here.*

Individuals showed widely varying dependencies on stimulus correlation as measured by the slope of a linear psychometric function fit to discriminability data (Figure 2.3c; sig. slopes = 0.43 ± 0.18). Lastly, despite the stimuli being presented at threshold levels, we were concerned about the possibility of participants adopting a strategy that exploits the low proportion of catch trials (i.e., they could be always reporting the presence of the stimulus modulation). We therefore quantified participant's willingness to respond with "modulation present." Figure 2.3d confirms that this strategy was not employed (c = 0.61 ± 0.41) with 11 of 12 participants adopting a conservative criterion. Further reinforcing this, 10 out of 12 participants (including the lone participant with a liberal criterion) were within one standard deviation of an unbiased criterion ($-1 < c < 1$).

Perceived stimulus correlation predicts audiovisual behavior via changes in evidence accumulation

Next, we sought to describe how audiovisual temporal correlation and phase shift influence behavioral performance in a decisional framework. Typically, changes in choice frequency and reaction time in a decision task are driven by changes in sensory evidence. We hypothesized that, in our task, sensory evidence was conferred by the temporal correlation of the stimuli. Further, we asked whether perceptual correlations rather than physical correlations better account for changes in behavioral performance on a participant-by-participant basis. To answer these questions, we employed two decision models.

The first model assumed that the drift rates, which index sensory evidence, are related to physical stimulus correlations ($r_{av} \mid \phi_0$) across conditions (Figure 2.4a). For the second model we assumed that the drift rates are related to the perceived stimulus correlations (Figure 2.4b), that is, correlations determined after a phase shift was applied ($r_{av} \mid \phi_i$). This design allowed the models not only to predict choice and reaction times with sensory evidence based on stimulus correlation, but also to measure participant phase shifts, providing converging evidence (in

conjunction with results provided above) of an internal phase shift of the representation of the physical stimuli.

Table 2.3. Model 1 parameters

Ptc.	θ	β	T_r	w	X^2	AIC
1	13	0.2151	0.5741	0.0182	97.719	-0.281
2	7	0.4954	0.7915	0.0762	87.192	-10.808
3	19	2.3893	0.6113	0.0111	175.113	77.113
4	9	3.5568	0.7605	0.0001	70.945	-27.055
5	13	-4.4554	0.8	0.0018	125.043	27.043
6	9	6.2835	0.6364	0.0001	95.408	-2.592
7	8	-0.6953	0.6018	0.0334	142.682	44.682
8	15	0.1974	0.788	0.0292	108.722	10.722
9	17	0.1539	0.7956	0.025	131.553	33.553
10	5	1.3622	0.7738	0.0339	88.424	-9.576
11	16	-11.9243	0.588	0.0333	163.444	65.444
12	9	3.2324	0.7994	0.0089	142.744	44.744

Participants (Ptc.) with nonsignificant correlations (Figure 2.3b) are in red. θ = boundary separation, β = evidence starting point, T_r = residual time, w = drift-rate scaling parameter.

Table 2.4. Model 2 parameters

Ptc.	Φ	θ	β	T_r	w	X^2	AIC
1	46	16	1.3753	0.4859	0.0214	214.55	14.55
2	0	7	-0.2467	0.7881	0.078	158.4	-41.6
3	0	19	2.3988	0.6074	0.011	281.68	81.681
4	-104	6	2.6569	0.7915	0.0348	136.03	-63.97
5	-2	13	-4.4437	0.8	0.0012	217.53	17.528
6	-92	8	4.6167	0.6443	0.0519	181.62	-18.382
7	13	10	-2.0867	0.5639	0.0298	252.04	52.041
8	19	15	0	0.7863	0.0317	168.74	-31.256
9	-46	17	-1.3041	0.8	0.0374	208.79	8.79
10	-16	14	2.3778	0.6013	0.0133	237.15	37.146
11	-8	15	-11.4521	0.6079	0.0375	225.2	25.201
12	2	9	2.2282	0.7951	0.0086	242.31	42.309

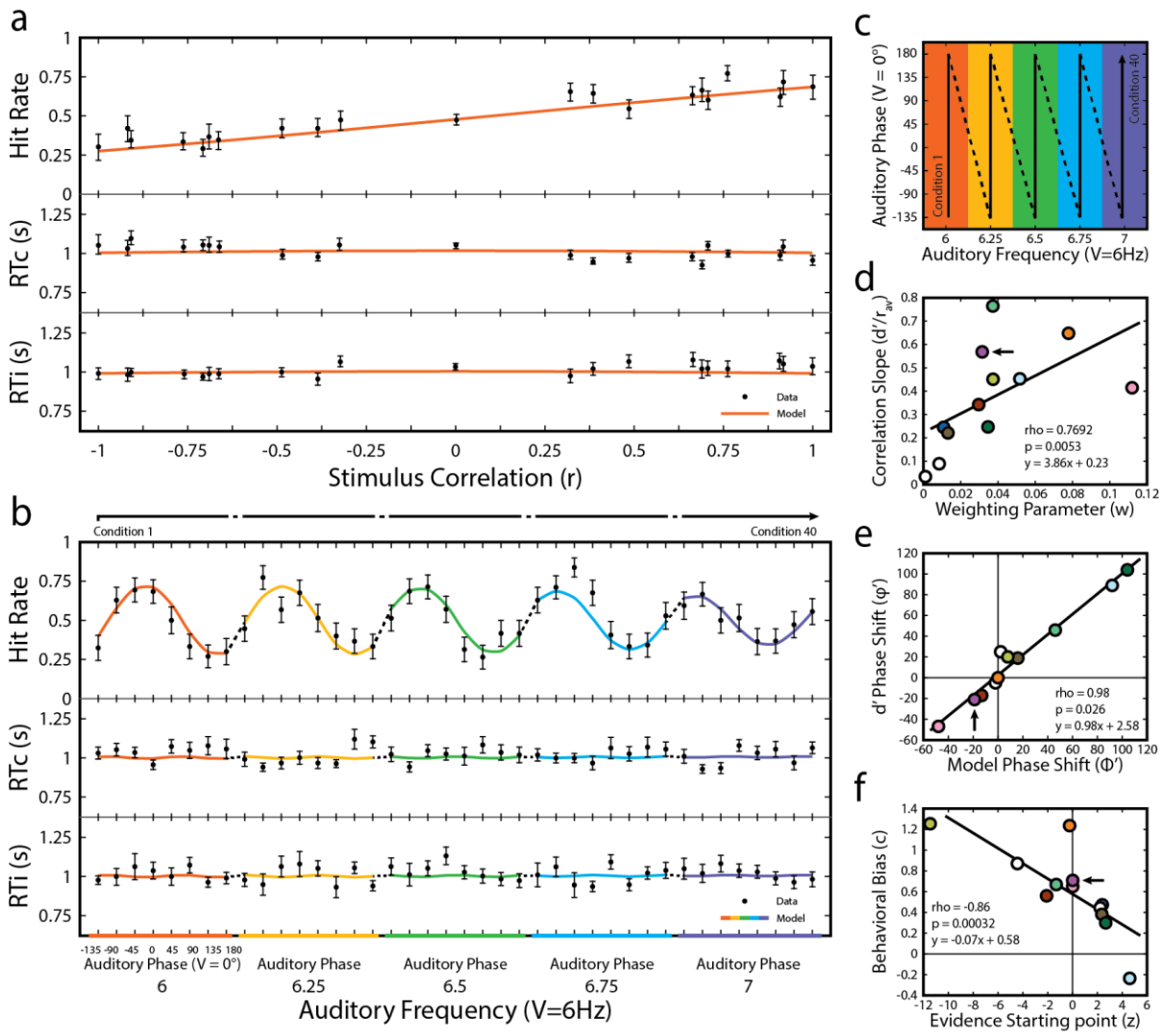
Participants (Ptc.) with nonsignificant correlations (Figure 2.3b) are in red. Φ = phase shift, θ = boundary separation, β = evidence starting point, T_r = residual time, w = drift-rate scaling parameter.

Table 3 and 4 show the estimated parameters for each model and their goodness of fit. Both models were well fit to the data and model 2 successfully incorporated the extra parameter for phase shift without compensation from other parameters meant to index bias, speed-accuracy tradeoff, and sensory encoding/motor preparation. As evidence that the models were not simply adjusting other parameters to adjust between models, we found that these parameters were strongly correlated between models when accounting for phase shift using partial correlations (θ: rho = 0.78, p = 0.0046; β: rho = 0.98, p = 7.67×10^{-8}; T_r: rho = 0.87, p = 0.00044). Using Akaike

Information Criterion (AIC) as a model selection metric, we found that most (8/12) participants' behavior was better described by the second model, in which the perceived correlation, included as a phase shift parameter, drives the decision process. Qualitatively, perceptual choice across conditions can be described as a dampening oscillator with dampening increasing with Δ frequency, a pattern which is also apparent in the model prediction of choice. Figure 2.4b shows the model fit (colored lines matching conditions shown in Figure 2.4c) to a single participant's data (filled circles).

Figure 2.4 (next page): Modeling results and comparison to behavioral results. *(a) Model 1 fit for a single participant. Proportion of correct responses (top panel), reaction times for correct responses (RTc; middle panel), and reaction times for incorrect responses (RTi; bottom panel) are shown (black dots, ± 1 S.E.M) for the 21 unique audiovisual correlations. The same data are shown from the model prediction (red lines).* ***(b)*** *Model 2 fit for the same participant. Proportion of correct responses (top panel), reaction times for correct responses (RTc; middle panel), and reaction times for incorrect responses (RTi; bottom panel) are shown (black dots, ± 1 S.E.M) for all 40 audiovisual conditions (top arrow). Model predictions and observed data are shown along a single continuous axis for simplicity with non-continuous data points connected by dashed lines (see panel c for key).* ***(c)*** *Representation of experimental conditions (frequency and phase) and how they are represented in panel (b). Conditions are organized in matrices (as in Figure 2.2b-c) with columns representing different frequencies and rows representing different phases. In (b), data have been reorganized column-wise such that Condition 1 is the first phase in the first frequency and Condition 40 is the last phase in the last frequencies. Colors of the model fit and bottom axis in (b) correspond to columns in the matrix with the same color. The top arrow in (b) correspond to the arrow in (c), unfolded.* ***(d)*** *Across all participants, phase shifts measured from discriminability matrices (Figure 2.2e) are strongly correlated with the phase shift parameters output by the diffusion model. Participant data shown in (b) correspond to the marker indicated by the arrow.* ***(e)***

Measures of bias, criteria (from Figure 2.3d) and evidence starting point parameters are correlated across participants. Participant data shown in (b) correspond to the marker indicated by the arrow. (f) Measures of dependence on stimulus correlation, psychometric slopes (from Figure 2.3c) and scaling parameters are correlated across participants. Participant data shown in (b) correspond to the marker indicated by the arrow.



Model 2 made accurate predictions of behavioral choice and reaction times based on the perceptual correlations and returned parameters that closely matched their signal detection theory counterpart. Each participant's model-fit phase shift parameter (ϕ') nearly perfectly matched their phase shift obtained from discriminability (φ', Figure 2.4d; rho = 0.98, p = 0.026, slope = 0.98). Additionally, evidence starting point, β, which is the parameter that measures the participant's bias toward one response over another (Laming, 1968; Voss et al., 2004), was also correlated with the signal detection theory measure of bias, c (Figure 2.4e; rho = 0.77, p = 0.0053). The bias reflects the participant's tendency to respond with "modulation present" or "modulation absent", which is unrelated to the sensitivity of the participant. Lastly, the drift-rate weighting coefficient was strongly correlated with the slope of their psychometric functions (Figure 2.4f; rho = -0.86, p = 0.00032), with both measures describing the dependence of behavior on changes in correlation. Moreover, the parameters were very consistent between models.

Discussion

Temporal factors such as (a)synchrony have long been known to influence multisensory processes in the brain (Bushara, Grafman, & Hallett, 2001; E. Macaluso, Frith, & Driver, 2002;

Macaluso et al., 2004; Meredith et al., 1987; Senkowski et al., 2007; Wallace et al., 1996) and in behavior (Colonius & Diederich, 2004; Dixon & Spitz, 1980; Frens et al., 1995; Fujisaki, Shimojo, Kashino, & Nishida, 2004; Hershenson, 1962; McGrath & Summerfield, 1985; J. V Stone et al., 2001). More recently, Parise and colleagues (2012) presented evidence that the fine temporal structure of an audiovisual stimulus *independent of asynchrony* can influence multisensory perception. They further showed that it is possible to explain a number of multisensory phenomena based on a general correlation detection mechanism (Parise & Ernst, 2016).The findings presented in the current study provide additional and unique support for the growing evidence implicating temporal correlation as an important cue in multisensory processing.

In the current work we extend this knowledge about multisensory temporal dependencies by showing that audiovisual detection behavior is a monotonic function of stimulus correlation. As the temporal similarity of two unisensory signals increased, detection of amplitude modulation embedded in the audiovisual signal improved in a linear manner (Figure 2.3a). Additionally, we qualify this finding in a way that provides mechanistic insight into how the brain combines dynamic stimuli across sensory modalities. Thus, the temporal correlation of the

audiovisual stimuli did not necessarily map directly onto multisensory behavioral performance; conditions in which physical stimulus correlation was highest did not always result in the best behavioral performance. Instead, it appears that a transform occurs in the brain of each individual and that results in a phase shift in behavioral performance relative to physical stimulus correlation ($r_{av} \mid \varphi_0$). Calculating temporal correlation after applying a phase lag to one of the stimuli ($r_{av} \mid \varphi'$), which simulates differential processing times for sensory signals in the brain, accounts for this difference. These phase-shifted correlations presumably represent the correlations as they are available to our decisional system.

Although our task did not reveal any measurable effects of temporal correlation on reaction times, we are not surprised. This lack of effect can be explained in terms of RT variability. Our stimuli employed near-threshold signals which are known to produce reaction times that are more variable than those produced by supra-threshold signals (McKendrick, Denniss, & Turpin, 2014). Additionally, the correlations in some stimulus conditions unfolded over time. In contrast, for some conditions the correlation does not change throughout the course of the signals. For example, when the auditory and visual modulations are both at 6Hz, across the entire stimulus, the relationship is maintained regardless of phase. However, when the

frequencies of visual and auditory AM are different (e.g., 6Hz and 7Hz, respectively), the starting and ending phase relationships change. In one phase condition (see Figure 2.1c), stimuli start out of phase (strong negative correlation) and end in phase (strong positive correlation). In another they start in phase and end out of phase. However, both of these conditions have an averaged correlation of 0 across the entire stimulus duration. This difference could introduce more reaction-time variability in some conditions than others, which may mask potential RT effects in some participants. To better measure any potential effect on reaction times, future experiments should be designed using supra-threshold signals that generate more reliable reaction times and take into account how correlations unfold over time.

The current study strongly grounds the relationship between stimulus correlation and multisensory processing in a decisional framework. Our model successfully incorporated the relationship between two signals (i.e., temporal correlation) into a dynamic-stochastic approach to account for choice frequency and response time. With only very few parameters (4 for model 1 and 5 for model 2) stimulus correlation was able to account for the observed patterns. Moreover, it was able to account for individual differences within and across participants. Our primary finding is related to the nature of how stimulus correlation influences the accumulation of

sensory evidence for a decision. Specifically, we found that perceived (phase-shifted) stimulus correlation serves as a good predictor of behavior when used to constrain drift rate. For perceptual tasks, drift rate is often interpreted as an index for the quality (e.g., strength) of sensory evidence that is available to the decisional system (Gold & Shadlen, 2007; Ratcliff & Smith, 2004). Typically, the strength of sensory evidence is provided by the physical attributes of the stimulus, for instance, the degree of motion coherence (Ratcliff & McKoon, 2008), intensity (Rach et al., 2011), line length (Diederich & Busemeyer, 2006), or numerosity (Leite, 2012). For simple multisensory behaviors (e.g., detection of simple stimuli), the drift rate relates to the combined evidence obtained from integrating the physical stimulus properties across modalities (Otto & Mamassian, 2012; Rach et al., 2011), especially when these properties are weak or ambiguous (e.g., low intensity, poor motion coherence, etc.) in the unisensory component stimuli (Rach et al., 2011).

In the current task, the key physical parameter that would presumably modulate the magnitude of evidence for detection is the depth of the amplitude modulation, with strength of evidence increasing with depth. However, modulation depth, and thus sensory evidence from the unisensory signals, is held constant across conditions. Although we cannot rule out that evidence

is supplied by integration of the unisensory stimulus properties, sensory evidence cannot come from these alone but instead is generated via a computation involving both stimuli. Different types of multisensory decisions require different architectures that depend on the structure of the task or stimulus (Bizley, Jones, & Town, 2016). The results presented here—that the strength of sensory evidence is based on a computation of the unisensory signals rather than the strength of the unisensory signals themselves—suggests that unisensory signals converge and evidence is computed prior to being evaluated by the decisional system. Other multisensory decisions such as simultaneity judgement (Simon, Nidiffer, & Wallace, 2018) and temporal order judgement (Diederich & Colonius, 2015; Mégevand, Molholm, Nayak, & Foxe, 2013), which require a similar comparison of the unisensory signals, have also been described in terms of their cross-modal computations.

It has recently been discussed that the presence or absence of audiovisual temporal correlation is a strong determinant of multisensory binding (Bizley et al., 2016) which manifests in a variety of behavioral enhancements (Grant & Seitz, 2000; Maddox et al., 2015; Parise et al., 2012). Results presented here extend this concept, despite the substantially different nature of the stimuli and task employed. According to our results, multisensory benefits—and likely by

extension the propensity to bind two signals—are monotonically related to the strength and sign of the temporal correlation (similarity) between unisensory signals. This notion implies that the process of binding signals is probabilistic. Stochastic binding related to temporal correlation could be an important mechanism in cognitive flexibility. It must be noted that weak, yet often significant, correlations exist in randomly paired stimuli (Chandrasekaran et al., 2009). In a sensory-rich environment, compulsory binding based on temporal similarity could lead to the perceptual unification of unrelated stimuli, creating great ambiguity in deciphering the sensory world. Instead, since the perceptual system has access to the strength of the correlation, the strongest and likely most appropriate signals can be bound. Further, it's likely that binding and integration are built on several other features such as spatial and temporal proximity. In the natural environment, these features are very often aligned; a single event will produce energies across different modalities that overlap in space and time and that are temporally correlated. Where these features are somewhat discrepant, the brain will appropriately weight (i.e., according to their reliability) proximity and similarity in the construction of a multisensory percept (Alais & Burr, 2004; Ernst & Banks, 2002).

The perceptual benefits of increased stimulus correlation are likely the result of mechanisms involving synchronized or coherent neural activity across brain regions (Nozaradan, Peretz, & Mouraux, 2012). Neural coherence has been hypothesized to play a role in shaping our conscious experience (Tononi & Koch, 2008) by underpinning mechanisms of sensory awareness (Engel & Singer, 2001), attentional selection (Schroeder & Lakatos, 2008), cognitive flexibility (Fries, 2005), and perceptual binding (Elhilali et al., 2009; Hipp, Engel, & Siegel, 2011; Senkowski, Schneider, Foxe, & Engel, 2008; Singer & Gray, 1995). Further, temporally correlated audiovisual streams have been shown to improve the representation of the auditory stimulus envelope and features in auditory cortex (Atilgan et al., 2018). This enhanced representation is likely the end result of why seeing a speaker's face improves speech intelligibility (Erber, 1969; Grant & Seitz, 2000; Sumby & Pollack, 1954). Rhythmic auditory and visual stimuli like the ones used in the current study are known to entrain neural oscillations (Henry & Obleser, 2012; Nozaradan et al., 2012; Thut, Schyns, & Gross, 2011) which index patterns of neuronal excitability over time (Bishop, 1933). Since uni- and multisensory stimuli can simultaneously entrain oscillations in multiple frequency bands (Henry, Herrmann, & Obleser,

2014; Nozaradan et al., 2012), it is likely that our stimuli do the same and thus induce coherent brain activity commensurate with the correlation in the stimuli.

In the current study, participants' behavioral performance was not necessarily best for the stimuli with highest physical correlation but were instead phase-shifted by differing amounts for each participant. Behavior very closely matched the correlation of the modulations after a phase lag was applied to one of the modulation signals. This phase lag could be adjusting for different processing times and abilities of participants' auditory and visual systems. It's known that oscillations entrain to rhythmic auditory stimuli at different phase lags across listeners (Henry & Obleser, 2012). It is possible that visual entrainment occurs in a similar manner and that these phase lags differ between the auditory and visual systems, though we are not aware of such data. Interestingly, phase lag of the entrained oscillations can be calibrated to the particular temporal structure of an audiovisual stimulus (Kösem, Gramfort, & Van Wassenhove, 2014). Thus, the phase lags reported in the current study are likely a "preferred" or "natural" phase that can be easily manipulated depending on context (e.g., attending an event that is near or far from the body which would result in different temporal relationships between auditory and visual

representations in the brain) in a manner similar to the phenomenon of recalibration of the perception of audiovisual simultaneity (Fujisaki et al., 2004; Van der Burg, Alais, & Cass, 2013).

During multisensory decisions, temporal correlation between the features of the component stimuli modulates behavior. It does so by changing the nature of the sensory evidence that is evaluated by the sensory system. The strength of the sensory evidence is proportional to the strength of the correlation of the signal. Finally, the physical correlations present in stimuli are transformed, via a phase shift, into "perceptual" correlations that are unique to an individual. This process likely occurs through differences in unisensory temporal processing. This was confirmed by a dynamic-stochastic model in which the drift rate was related to physical or to perceived correlations between the auditory and visual signals in the audiovisual presentation. These results motivate several fundamental questions. Is binding truly stochastic? Can cross-modal correlation embedded in one feature (e.g., intensity) have the same proportional effect on behavioral performance reported here in tasks utilizing orthogonal stimulus features (e.g., frequency or timbre)? What are the neural signatures of this proportional change and their relation to behavior? Finally, does the perception of naturalistic audiovisual stimuli such as speech benefit in the same way with changes in audiovisual correlation?

Materials and Methods

Participants

Twelve individuals (age = 26.4 ± 5.1, seven females) participated in the current study. All participants reported normal or corrected-to-normal vision and normal hearing and were right handed. The study was conducted in accordance with the declaration of Helsinki, and informed written consent was obtained from all participants. All procedures were approved by the Vanderbilt University Institutional Review Board. When applicable, participants were given monetary compensation for participation.

Apparatus and stimuli

All stimuli were generated in MATLAB (The MathWorks, Inc., Natick, MA) and presented using PsychToolbox version 3 (Brainard, 1997; Kleiner et al., 2007). Auditory stimuli were digitized at 44.1 kHz, and presented through calibrated open-back circumaural headphones (Sennheisser HD480). Visual stimuli were centered about a red fixation dot in the center of a dark (0.15 cd/m^2) viewing screen (Samsung Sync Master 2233rz, 120 Hz refresh rate).

Auditory stimuli were frozen tokens of white noise (generated by the *randn* function) at moderate baseline level (48 dB SPL, A-weighted). Visual stimuli consisted of a moderately

bright ring (24 cd/m^2 at baseline; inner diameter: 1.8°, outer diameter: 3.6° visual angle). Both stimuli were presented simultaneously, lasted 500 ms, and were gated by a linear 10 ms onset and offset ramp. Stimulus timing was confirmed with a Hameg 507 oscilloscope, photodiode, and microphone.

For each stimulus, auditory intensity and visual luminance, *y*, could be modulated around their baseline over time, *t*, such that

$$y(t) = [1 + m(t)] \times c(t)$$

where

$$m(t) = M \times sin\left(2\pi f_m t + \varphi_{0,j}\right)$$

and *c(t)* is the time series of the carrier stimulus (auditory: noise; visual: ring). The form of the amplitude modulation (AM) signal *m(t)* is defined by a modulation depth *M* which represents the amplitude of the modulation signal as a proportion of the amplitude of the carrier signal and ranged from 0 (no AM) to 1 (full AM), frequency f_m in Hz, and starting phase $\varphi_{0,j}$ in degrees.

On any given trial, the AM signal could be present in the auditory channel alone, the visual channel alone, both channels (audiovisual trials), or neither (catch trials; Figure 2.1b). If present, modulation depth was set to individual unisensory thresholds (see below for

thresholding procedures). Unisensory signals (AM was present in auditory stimulus only or visual stimulus only) were always presented in cosine phase such that the modulation began at the trough ($\varphi = 0^{\circ}$) and at the same frequency ($f_{m,\,visual} = 6$ Hz). When AM was present in both stimuli, visual modulation was always 6 Hz and cosine starting phase while auditory signals could be presented at various frequencies ($f_{m,\,auditory} = \{6, 6.25, 6.5, 6.75, 7\ \text{Hz}\}$) and initial phases ($\varphi_0 = \{-135, -90, -45, 0, 45, 90, 135, 180^{\circ}\}$, with $\varphi_{0,j} \in \varphi_0$). This structure results in a total of 40 (5 × 8) different audiovisual stimulus conditions.

Because we are interested in the temporal correlation between the two signals, the Pearson correlation between the auditory and visual envelopes (r_{av}) was computed for each of the 40 audiovisual conditions (Figure 2.1c). For example, when the auditory and visual envelopes were characterized by the same frequency and phase, correlation was 1. Conversely, stimuli of the same frequency but presented anti-phase resulted in a correlation of -1. The parameters chosen resulted in a representation of correlations between -1 and 1. A stimulus correlation matrix ($r_{av} \mid \varphi_0$) was constructed for all audiovisual conditions by organizing the correlation values according to their frequency and phase relationship between auditory and visual signals (Δ frequency × Δ phase; Figure 2.1d).

Procedure

Participants were seated comfortably inside an unlit WhisperRoom™ (SE 2000 Series) with their forehead placed against a HeadSpot™ (University of Houston Optometry) with the forehead rest locked in place such that a participant's primary eye position was centered with respect to the fixation point at the center of the viewing screen. Chinrest height and chair height were adjusted to the comfort of the participant.

Prior to the main experiment, each participant completed two separate 3-down 1-up staircase procedures to obtain 79.4% modulation depth thresholds for auditory and visual AM at 6 Hz. For these staircase procedures, on a given trial (Figure 2.1a), the red fixation dot appeared at the center of the screen. Participants were instructed to fixate the dot for its entire duration. After a variable time, either an auditory or visual stimulus was presented in which the presence of modulation was determined at random for each trial. Participants were instructed to report the presence of amplitude modulation (described as "flutter") after the stimulus presentation by pressing "1" on the number pad of a computer keyboard if the modulation was present or pressing "0" if the modulation was absent. The modulation depth decreased after three successive correct responses and increased after one incorrect response. At the beginning of each staircase,

the step size was set to increase or decrease modulation depth by 0.05. After two reversals (correct to incorrect response or incorrect to correct response), step size was reduced to 0.025. Finally, after eight reversals, step size became 0.01 in order to arrive at an accurate estimate of modulation depth threshold. Each staircase terminated after 20 reversals. Threshold was determined to be the average of the modulation depth at the last 10 reversals. Instructions included an example of a stimulus with AM at the initial starting modulation depth ($M = 0.5$) and an example of a stimulus with no AM. So that there was no ambiguity in cases where the first trial did not include a modulation signal, participants were informed that the first trial would have the same modulation depth as the example if present. To control for "runs" of trials with no modulation during the staircase (which could result in erroneously low threshold estimates), a sequence of two trials containing no modulation was always followed by a trial with modulation. The auditory staircase was always completed first and served as a period of dark adaptation prior to the visual staircase.

The main experiment consisted of four blocks lasting approximately 30 minutes each. Each block consisted of 10 trials of each stimulus condition (420 signal trials per block). Additionally, there were catch (no signal) trials included to make up 10% of total trials for that

block (47 catch trials per block). Therefore, each block was identical in trial composition (467 total trials per block) but with individual trials presented in a predetermined, pseudorandom order. Each participant completed a total of 1868 trials over the four blocks. Breaks were offered frequently (every 100 trials) to prevent fatigue. Participants completed the full experiment in 2-4 sessions, never completing more than 2 blocks during a session. If a participant completed two blocks in a single session, they were given the opportunity to stretch and walk around while the experimenter set up the second block. Before each block and after any break where the participant was exposed to normal light levels, participants were dark adapted for five minutes. Trials during the main experiment were identical to staircase trials with three exceptions. First, in each trial, both auditory and visual stimuli were presented. Modulation signals could be present in the visual channel alone (V_{signal}), auditory channel alone (A_{signal}), in both (AV_{signal}; with frequency and phase configuration discussed above), or neither channel (no signal). Second, modulation depth was set to a participant's unique auditory and visual modulation depth thresholds. These threshold values are shown in Table 5. Last, participants were told that they should respond as soon as they had made their decision and were instructed to respond as quickly and accurately as possible. In addition to the participant's choice, response times were recorded

for each trial, sampling every 2.2 μs (4.6 kHz). Response window was terminated after 1.5 seconds. Subsequent responses were censored. This ended up being 2% of trials or less for most participants.

Table 2.5. Participant modulation depth thresholds

Ptc.	Aud.	Vis.
1	0.041	0.047
2	0.081	0.059
3	0.028	0.049
4	0.104	0.076
5	0.051	0.042
6	0.087	0.062
7	0.068	0.043
8	0.048	0.040
9	0.060	0.058
10	0.063	0.043
11	0.072	0.050
12	0.072	0.070

Behavioral analysis

Discriminability (*d'*; a measure of sensitivity) for each of the 40 audiovisual conditions and two unisensory conditions was computed from the relative frequencies of the respective responses,

$$d' = z(H_i) - z(F)$$

where H_i is the proportion of hits ("1" | modulated stimulus) for the i^{th} condition, F is the proportion of false alarms ("1" | no modulated stimulus), and z is the inverse of the normal distribution function (MATLAB's *norminv* function) and converts the hit rates and false alarm rates into units of standard deviation of a standard normal distribution. d' was organized into a matrix in the same manner as the stimulus correlation matrix. Because the proportion of catch trials was held low and errors had no associated cost (Green & Swets, 1966), participants could potentially adopt a strategy of simply pressing "1" which would result in a correct choice more often than not. To account for this, criterion (c; a measure of bias) for each participant was computed in a similar manner such that

$$c = z(H) + z(F)$$

where H is the proportion of hits across all conditions. A single criterion was computed for each participant.

To account for individual differences, which became apparent in assessing the phase shift in the d' matrices, a series of correlation matrices based on the stimulus correlation matrix (r_{av} | φ_0) were computed after iteratively applying a single degree phase lag to one stimulus (i.e., φ_1 = {-134, -89, -44, 1, 46, 91, 136, -179°}, φ_2 = {-133, -88, -43, 2, 47, 92, 137, -178°}, in general φ_i

= {-135 + *i*, -90 + *i*, -45 + *i*, 0 + *i*, 45 + *i*, 90 + *i*, 135 + *i*, 180 + *i* } with *i* = -180, …, 180, resulting in a total of 360 different matrices). A phase-shifted correlation matrix ($r_{av} \mid \varphi_i$) could be conceptualized as the "internal" or "perceived" correlation of the signals given a particular phase lag, *i*, of one of the signals. Each of the phase-shifted correlation matrices (Figure 2.1e, nine examples shown) was in turn evaluated for correlation ($r_{d'}$) with the discriminability matrix of each participant. The resulting correlation values ($r_{d'} \mid \varphi$) were then fit to a sine wave using the nonlinear least-squares method. The phase shift value of the fitted sine wave was recorded for each participant (φ'). The CircStat toolbox (Berens, 2009) was used to describe the nature of the phase shifts and compute the directional statistics across the sample of participants. The "perceptual" correlation matrix corresponding to each participant's unique phase shift ($r_{av} \mid \varphi'$) was used to measure the dependence of behavior on perceived correlation ($r_{d'} \mid \varphi'$).

To show that phase shift is related to a central mechanism (e.g., a relative difference in processing latencies between auditory and visual systems), we tested whether the phase shift occurred systematically across all Δ frequencies within each participant. First, a predicted discriminability matrix was calculated from phase-shifted correlations. Phase-shifted correlation matrices were normalized to each participant's discriminability range by scaling and shifting each

unique correlation matrix such that the correlation values at the maximum and minimum correlation matched the *d'* values at the corresponding locations in the discriminability matrix. Next, the values in the predicted discriminability matrix were subtracted from the actual discriminability matrix, resulting in a matrix of residual errors. Then, a linear model was used to determine the relationship (i.e., slope) between Δ frequency and the magnitude and variability (standard deviation) of errors. To calculate significance of variability slope across Δ frequency, a permutation test was used that shuffled the Δ frequency label of errors before calculating standard deviation within each Δ frequency and then fitting a line to the shuffled standard deviations.

We sought to demonstrate that accounting for phase shift improved the measured correlation between behavior and stimulus correlation. Therefore, we computed this dependence on stimulus correlation ($r_{d'} \mid \varphi_0$) and subtracted it from the dependence on perceived correlation discussed above ($r_{d'} \mid \varphi'$) which yielded a score of improvement (Δr). Because of the nature of the phase shift fitting process described above, ($r_{d'} \mid \varphi'$) ≥ ($r_{d'} \mid \varphi_0$) with the difference growing to a maximum when $\varphi = \pm 180°$ even for data with no effect (random numbers). Therefore, we accounted for this statistical effect by running a simulation where we computed the phase shift

(same process described in Figure 2.1e) of 1000 matrices of shuffled data from participants chosen at random. For each matrix, we measured ($r_{d'} \mid \varphi'$) and ($r_{d'} \mid \varphi_0$) and subtracted them as above so that we had 1000 pairs of φ' and Δr. These data, along with our observed data, were fit to the function

$$\Delta r = a \times sin(\varphi') + a$$

which returned *a*, the amplitude of the function. We then bootstrapped (10000 samples of 20 randomly drawn pairs of simulated φ' and Δr chosen with replacement) fits to the simulated data to obtain a distribution of *a* for these null data. From this distribution, we computed a z-score for the observed amplitude parameter as

$$z = \frac{a_{obs}}{(u - l)/(2 \times 1.96)}$$

where a_{obs} is the amplitude parameter of the fit to the observed data and *u* and *l* are the upper and lower 95% confidence bounds from the bootstrapped fits to the shuffled data, respectively.

Diffusion model analysis

For binary choices, sequential-sampling models assume that upon presentation of the stimulus, the decision maker sequentially samples information from the stimulus display over

time, which provides sensory evidence to a decision process. It also assumes that the decision process accumulates this evidence in a noisy manner for choosing one option over the other, here "modulation present" or "modulation absent." Sequential-sampling models account simultaneously for choice frequency and choice response times. However, the focus here will be on choice frequencies. Let X(t) denote the random variable representing the numerical value of the accumulated evidence at time t. A bias, β, (i.e., prior beliefs about the stimulus before it is presented) can influence the initial starting position of the decision process, X(0). This initial state may either favor choice option "modulation present" (X(0) >0) or choice option "modulation absent" (X(0) <0). X(0)=0 reflects an unbiased response. (The initial states can also be given a probability distribution). The participant then samples small increments of evidence at any moment in time, which either favor response "modulation present" (dX(t) > 0) or response "modulation absent" (dX(t) < 0). The evidence is incremented according to a diffusion process. In particular, we apply a Wiener process with drift, lately called drift-diffusion model (Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006) with

$$dX(t) = \delta + \sigma dW(t)$$

The drift rate, δ, describes the expected value of evidence increments per unit time. The diffusion rate, σ, in front of the standard Wiener process, $W(t)$, relates to the variance of the increments. Here we set $\sigma = 1$. The small increments of evidence sampled at any moment in time are such that they either favor response "modulation present" ($dX(t) > 0$) or response "modulation absent" ($dX(t) < 0$). This process continues until the magnitude of the cumulative evidence exceeds a threshold criterion, θ. That is, the process stops and response "modulation present" is initiated as soon as the accumulated evidence reaches a criterion value for choosing response "modulation present" (here, $X(t) = \theta > 0$), or it stops and a "modulation absent" response is initiated as soon as the accumulated evidence reaches a criterion value for choosing response "modulation absent" (here, $X(t) = \theta < 0$). The probability of choosing the response "modulation present" over "modulation absent" is determined by the accumulation process reaching the threshold for response "modulation present" before reaching the threshold for response "modulation absent". The criterion is assumed to be set by the decision maker prior to the decision task. The drift rate may be related to the quality of the stimuli (i.e., the better the quality the higher the drift rate). For instance, stimuli that are easier to discriminate are reflected in a higher drift rate. In the following we consider two models. In Model 1 we assume that the

physical correlation between the auditory and visual stimuli, ($r_{av} \mid \varphi_0$), weighted by the decision maker drives the evidence accumulation process for initiating a "modulation present" or "modulation absent" response. That is, the drift rate is defined as

$$\delta = w \times (r_{av}|\phi_0)$$

Of the 40 correlation coefficients several of them were identical (for instance, a 6Hz auditory stimulus with starting phases of +45° and -45° both resulted in a correlation of .7075) resulting in 21 unique correlation coefficients and by that in 21 different drift rates.

In Model 2 we assume that the physical correlation between the auditory and visual stimuli is distorted by a shift in phase as perceived by the decision maker. That is, the drift rate is defined by

$$\delta = w \times (r_{av}|\phi_i)$$

where i is a free parameter of the model estimated from the data and its returned value corresponds to a phase shift that is unique to each participant (ϕ'). The model term ϕ_i relates to the initial phase term φ_i introduced earlier and follows the same naming conventions. A phase shift unequal to 0, ±45, ±90, ±135, or ±180 results in 40 different correlation coefficients which in turn results in 40 drift rates.

Diffusion model parameters

We assume for both models that the observed response time is the sum of the decision time, modeled by the diffusion process, and a residual time, T_r, which includes the time for processes other than the decision, e.g., sensory encoding and motor components. Here, T_r, is a constant for each participant. Because correlation coefficients varied between 1 and -1 but none of the participants showed perfect performances (e.g. 100% of correct responses to either a perfectly positively correlated stimulus pair or a perfectly negatively correlated stimulus pair), we allow an adjustment by including a weight for the correlations $0 \leq w \leq 1$. We also allow for an a priori response bias, β, in favor of one response (present/absent). The decision criteria are $\theta = |-\theta|$.

In addition to these parameters, Model 2 returns a parameter ϕ' to account for perceived correlations based on individual phase shifts (rather than correlations based on the physical stimuli only) to be estimated from the data. To summarize: For Model 1 four parameters (w, β, θ, T_r) are estimated from 63 data points (21 relative frequencies for correct responses, 21 mean response times for correct responses, 21 mean response times for incorrect responses. Trials with identical correlations were collapsed.) For Model 2 five parameters (ϕ', w, β, θ, T_r) are estimated

from 120 data points (40 relative frequencies for correct responses, 40 mean response times for correct responses, 40 mean response times for incorrect responses).

The model was implemented in terms of the matrix approach (Diederich & Busemeyer, 2003) and parameters were estimated by minimizing the chi-square function (Smith & Vickers, 1988),

$$\chi^2 = \sum \left(\frac{RT_{obs} - RT_{pred}}{SE_{RT_{obs}}} \right)^2 + \sum \left(\frac{Pr_{obs} - Pr_{pred}}{SE_{Pr_{obs}}} \right)^2$$

using the optimization routine *fminsearchbnd* in MATLAB. The *fminsearchbnd* routine is similar to the standard *fminsearch* routine except that the range of the parameters of the parameters can be predetermined, for instance, positive real numbers for the residuals, or real numbers between 0 and 1 for the weights. The *fminsearch* uses the Nelder-Mead simplex search method (Lagarias, Reeds, Wright, & Wright, 1998). $SE_{RT_{obs}}$ and $SE_{Pr_{obs}}$ refer to the standard error of the observed mean response times and relative choice frequencies, respectively. Note that mean response times and relative choice frequencies are conditioned on the stimulus presented. Here we consider only the trials in which a modulation was present.

For both models, the following procedures/restrictions to parameter values were imposed in the estimation procedure: The decision criteria (absorbing boundaries) were estimated using a

search grid. This was done because it quickens the estimation procedure when boundaries are integers (matrix approach). θ ranged from 3 to 20 in steps of 1. The residual time, T_r, was restricted to 100 ms ≤ T_r ≤ 800 ms and the weight to 0.0001 ≤ w ≤ 1. For the Model 2 parameter ϕ_i, the value of i was restricted to integers ranging from -180 to 180 in steps of 1. For each value of i in Model 2, a different set of correlations were computed.

References

Alais, D., & Burr, D. (2004). Ventriloquist Effect Results from Near-Optimal Bimodal Integration. Current Biology, 14(3), 257–262. https://doi.org/10.1016/S0960-9822(04)00043-0

Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., Lee, A. K. C., & Bizley, J. K. (2018). Integration of Visual Information in Auditory Cortex Promotes Auditory Scene Analysis through Multisensory Binding. Neuron, 97(3), 640–655.e4. https://doi.org/10.1016/j.neuron.2017.12.034

Berens, P. (2009). CircStat : A MATLAB Toolbox for Circular Statistics. Journal of Statistical Software, 31(10). https://doi.org/10.18637/jss.v031.i10

Bishop, G. H. (1933). Cyclic changes in excitability of the optic pathway of the rabbit. American Journal of Physiology--Legacy Content, 103(1), 213–224.

Bizley, J. K., Jones, G. P., & Town, S. M. (2016). Where are multisensory signals combined for perceptual decision-making? Current Opinion in Neurobiology. https://doi.org/10.1016/j.conb.2016.06.003

Bizley, J. K., Maddox, R. K., & Lee, A. K. C. (2016). Defining Auditory-Visual Objects: Behavioral Tests and Physiological Mechanisms. Trends in Neurosciences. https://doi.org/10.1016/j.tins.2015.12.007

Blake, R., & Lee, S.-H. (2005). The role of temporal structure in human vision. Behavioral and Cognitive Neuroscience Reviews, 4(1), 21–42. https://doi.org/10.1177/1534582305276839

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. Psychological Review, 113(4), 700–765. https://doi.org/10.1037/0033-295X.113.4.700

Bolognini, N., Frassinetti, F., Serino, A., & Làdavas, E. (2005). "Acoustical vision" of below threshold stimuli: Interaction among spatially converging audiovisual inputs. Experimental Brain Research, 160(3), 273–282. https://doi.org/10.1007/s00221-004-2005-z

Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial Vision, 10, 433–436. https://doi.org/10.1163/156856897X00357

Bushara, K. O., Grafman, J., & Hallett, M. (2001). Neural correlates of auditory-visual stimulus onset asynchrony detection. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 21(1), 300–4. https://doi.org/21/1/300 [pii]

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. PLoS Computational Biology, 5(7). https://doi.org/10.1371/journal.pcbi.1000436

Chuen, L., & Schutz, M. (2016). The unity assumption facilitates cross-modal binding of musical, non-speech stimuli: The role of spectral and amplitude envelope cues. Attention, Perception, and Psychophysics, 78(5), 1512–1528. https://doi.org/10.3758/s13414-016-1088-5

Colonius, H., & Diederich, A. (2004). Multisensory Interaction in Saccadic Reaction Time: A Time-Window-of-Integration Model. Journal of Cognitive Neuroscience, 16(6), 1000–1009. https://doi.org/10.1162/0898929041502733

Diederich, A., & Busemeyer, J. R. (2003). Simple matrix methods for analyzing diffusion models of choice probability, choice response time, and simple response time. Journal of Mathematical Psychology, 47(3), 304–322. https://doi.org/10.1016/S0022-2496(03)00003-8

Diederich, A., & Busemeyer, J. R. (2006). Modeling the effects of payoff on response bias in a perceptual discrimination task: bound-change, drift-rate-change, or two-stage-processing hypothesis. Perception & Psychophysics, 68(2), 194–207. https://doi.org/10.3758/BF03193669

Diederich, A., & Colonius, H. (2015). The time window of multisensory integration: relating reaction times and judgments of temporal order. Psychological Review, 122(2), 232–41. https://doi.org/10.1037/a0038696

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

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009). Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. Neuron, 61(2), 317–329. https://doi.org/10.1016/j.neuron.2008.12.005

Engel, A. K., & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. Trends in Cognitive Sciences. https://doi.org/10.1016/S1364-6613(00)01568-0

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. Journal of Speech Language and Hearing Research, 12(2), 423. https://doi.org/10.1044/jshr.1202.423

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. Nature, 415(6870), 429–433. https://doi.org/10.1038/415429a

Frassinetti, F., Bolognini, N., & Làdavas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. Experimental Brain Research, 147(3), 332–343. https://doi.org/10.1007/s00221-002-1262-y

Frens, M. A., & Van Opstal, A. J. (1995). A quantitative study of auditory-evoked saccadic eye movements in two dimensions. Experimental Brain Research, 107(1), 103–17. https://doi.org/10.1007/BF00228022

Frens, M. A., Van Opstal, A. J., Van der Willigen, R. F., Opstal, A. J. Van, & Willigen, R. F. Van Der. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. Perception & Psychophysics, 57(6), 802–816. https://doi.org/10.3758/BF03206796

Fries, P. (2005). A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence. Trends in Cognitive Sciences. https://doi.org/10.1016/j.tics.2005.08.011

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. Nature Neuroscience, 7(7), 773–778. https://doi.org/10.1038/nn1268

Gold, J., & Shadlen, M. (2007). The neural basis of decision making. Annu. Rev. Neurosci, 30(1), 535–574. https://doi.org/10.1146/annurev.neuro.29.051605.113038

Grant, K. W., & Seitz, P. F. 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 Pt 1), 1197–1208. https://doi.org/10.1121/1.422512

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. Society, 1, 521. https://doi.org/10.1901/jeab.1969.12-475

Henry, M. J., Herrmann, B., & Obleser, J. (2014). Entrained neural oscillations in multiple frequency bands comodulate behavior. Proceedings of the National Academy of Sciences of the United States of America, 111(41), 1408741111-. https://doi.org/10.1073/pnas.1408741111

Henry, M. J., & Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. Proceedings of the National Academy of Sciences, 109(49), 20095–20100. https://doi.org/10.1073/pnas.1213390109

Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. Journal of Experimental Psychology, 63(3), 289–293. https://doi.org/10.1037/h0055703

Hipp, J. F., Engel, A. K., & Siegel, M. (2011). Oscillatory synchronization in large-scale cortical networks predicts perception. Neuron, 69(2), 387–396. https://doi.org/10.1016/j.neuron.2010.12.027

Jack, C. E., & Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. Perceptual and Motor Skills, 37(3), 967–979. https://doi.org/10.2466/pms.1973.37.3.967

Kleiner, M., Brainard, D. H., Pelli, D. G., Broussard, C., Wolf, T., & Niehorster, D. (2007). What's new in Psychtoolbox-3? Perception, 36, S14. https://doi.org/10.1068/v070821

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal Inference in Multisensory Perception. PLoS ONE, 2(9), e943. https://doi.org/10.1371/journal.pone.0000943

Kösem, A., Gramfort, A., & Van Wassenhove, V. (2014). Encoding of event timing in the phase of neural oscillations. NeuroImage, 92, 274–284. https://doi.org/10.1016/j.neuroimage.2014.02.010

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence Properties of the Nelder--Mead Simplex Method in Low Dimensions. SIAM Journal on Optimization, 9(1), 112–147. https://doi.org/10.1137/S1052623496303470

Laming, D. R. J. (1968). Information theory of choice-reaction times. Information theory of choice-reaction times. Oxford: Academic Press. https://doi.org/10.1002/bs.3830140408

Leite, F. P. (2012). A comparison of two diffusion process models in accounting for payoff and stimulus frequency manipulations. Attention, Perception, & Psychophysics, 74(6), 1366–1382. https://doi.org/10.3758/s13414-012-0321-0

Macaluso, E., Frith, C. D., & Driver, J. (2002). Crossmodal spatial influences of touch on extrastriate visual areas take current gaze direction into account. Neuron, 34(4), 647–658. https://doi.org/10.1016/S0896-6273(02)00678-5

Macaluso, E., George, N., Dolan, R., Spence, C., & Driver, J. (2004). Spatial and temporal factors during processing of audiovisual speech: a PET study. Neuroimage, 21(2), 725–732. https://doi.org/10.1016/j.neuroimage.2003.09.049

Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. ELife, 2015(4), 1–11. https://doi.org/10.7554/eLife.04995.001

Magnotti, J. F., Ma, W. J., & Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech. Frontiers in Psychology, 4, 798. https://doi.org/10.3389/fpsyg.2013.00798

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), 678–685. https://doi.org/10.1121/1.392336

McKendrick, A. M., Denniss, J., & Turpin, A. (2014). Response times across the visual field: Empirical observations and application to threshold determination. Vision Research, 101, 1–10. https://doi.org/10.1016/j.visres.2014.04.013

Mégevand, P., Molholm, S., Nayak, A., & Foxe, J. J. (2013). Recalibration of the Multisensory Temporal Window of Integration Results from Changing Task Demands. PLoS ONE, 8(8). https://doi.org/10.1371/journal.pone.0071608

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. The Journal of Neuroscience, 7(10), 3215–29. https://doi.org/citeulike-article-id:409430

Meredith, M. A., & Stein, B. E. (1986). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. Brain Research, 365(2), 350–354. https://doi.org/10.1016/0006-8993(86)91648-3

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. Perception & Psychophysics, 58(3), 351–362. https://doi.org/10.3758/BF03206811

Nozaradan, S., Peretz, I., & Mouraux, A. (2012). Steady-state evoked potentials as an index of multisensory temporal binding. NeuroImage, 60(1), 21–28. https://doi.org/10.1016/j.neuroimage.2011.11.065

Odegaard, B., Wozny, D. R., & Shams, L. (2015). Biases in Visual, Auditory, and Audiovisual Perception of Space. PLOS Computational Biology, 11(12), e1004649. https://doi.org/10.1371/journal.pcbi.1004649

Otto, T. U., & Mamassian, P. (2012). Noise and correlations in parallel perceptual decision making. Current Biology, 22(15), 1391–1396. https://doi.org/10.1016/j.cub.2012.05.031

Parise, C. V., & Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. Nature Communications, 7(12), 364. https://doi.org/10.1038/ncomms11543

Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. Current Biology, 22(1), 46–49. https://doi.org/10.1016/j.cub.2011.11.039

Parise, C. V, Harrar, V., Ernst, M. O., & Spence, C. (2013). Cross-correlation between Auditory and Visual Signals Promotes Multisensory Integration. Multisensory Research, 26, 1–10. https://doi.org/10.1163/22134808-00002417

Rach, S., Diederich, A., & Colonius, H. (2011). On quantifying multisensory interaction effects in reaction time and detection rate. Psychological Research, 75(2), 77–94. https://doi.org/10.1007/s00426-010-0289-0

Ratcliff, R., & McKoon, G. (2008). The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks. Neural Computation, 20(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Ratcliff, R., & Smith, P. L. (2004). A Comparison of Sequential Sampling Models for Two-Choice Reaction Time. Psychological Review, 111(2), 333–367. https://doi.org/10.1037/0033-295X.111.2.333

Schroeder, C. E., & Lakatos, P. (2008). Low-frequency neuronal oscillations as instruments of sensory selection. Trends in Neurosciences, 32(1), 9–18. https://doi.org/10.1016/j.tins.2008.09.012

Senkowski, D., Schneider, T. R., Foxe, J. J., & Engel, A. K. (2008). Crossmodal binding through neural coherence: implications for multisensory processing. Trends in Neurosciences. https://doi.org/10.1016/j.tins.2008.05.002

Senkowski, D., Talsma, D., Grigutsch, M., Herrmann, C. S., & Woldorff, M. G. (2007). Good times for multisensory integration: Effects of the precision of temporal synchrony as revealed by gamma-band oscillations. Neuropsychologia, 45(3), 561–571. https://doi.org/10.1016/j.neuropsychologia.2006.01.013

Simon, D. M., Nidiffer, A. R., & Wallace, M. T. (in press). Rapid Recalibration to Asynchronous Audiovisual Speech Modulates the Rate of Evidence Accumulation. Scientific Reports.

Singer, W., & Gray, C. (1995). Visual feature integration and the temporal correlation hypothesis. Annual Review of Neuroscience, 18, 555–586. https://doi.org/10.1146/annurev.ne.18.030195.003011

Smith, P. L., & Vickers, D. (1988). The accumulator model of two-choice discrimination. Journal of Mathematical Psychology, 32(2), 135–168. https://doi.org/10.1016/0022-2496(88)90043-0

Stein, B. E. (Ed.). (2012). The New Handbook of Multisensory Processes. Cambridge, MA: MIT Press.

Stone, J. V, Hunkin, N. M., Porrill, J., Wood, R., Keeler, V., Beanland, M., … Porter, N. R. (2001). When is now? Perception of simultaneity. Proceedings. Biological Sciences / The Royal Society, 268(1462), 31–38. https://doi.org/10.1098/rspb.2000.1326

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. The Journal of the Acoustical Society of America, 26(2), 212–215. https://doi.org/10.1121/1.1907309

Thut, G., Schyns, P. G., & Gross, J. (2011). Entrainment of perceptually relevant brain oscillations by non-invasive rhythmic stimulation of the human brain. Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2011.00170

Tononi, G., & Koch, C. (2008). The neural correlates of consciousness: An update. Annals of the New York Academy of Sciences. https://doi.org/10.1196/annals.1440.004

Van der Burg, E., Alais, D., & Cass, J. (2013). Rapid Recalibration to Audiovisual Asynchrony. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 33(37), 14633–7. https://doi.org/10.1523/JNEUROSCI.1182-13.2013

Vatakis, A., & Spence, C. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. Perception & Psychophysics, 69(5), 744–756. https://doi.org/10.3758/BF03193776

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: an empirical validation. Memory & Cognition, 32(7), 1206–1220. https://doi.org/10.3758/BF03196893

Wallace, M. T., Wilkinson, L. K., & Stein, B. E. (1996). Representation and integration of multiple sensory inputs in primate superior colliculus. Journal of Neurophysiology, 76(2), 1246–1266.

Chapter 3. Multisensory binding is proportional to stimulus correlation

Introduction

Our conscious perception depends crucially on our ability to appropriately piece together the features of events in our environment. We do this effortlessly across auditory and visual domains in processes called auditory scene analysis (Bregman, 1990) and visual grouping (Wertheimer, 1923). Each unisensory domain has its own set of cues (e.g., spectral, spatial, and temporal) that influence the binding of features into objects. One cue that is common to both unisensory modalities, temporal correlation, has also been shown to be a robust cue that binds features *across* modalities (Bizley et al., 2016). Indeed correlation across time is a property of cross-modal stimuli that originate from a common source, such as the mouth movements and vocal intensity of audiovisual speech (Chandrasekaran et al., 2009).

Temporal correlation is an important cue that leads to the binding and integration of multisensory signals (Grant & Seitz, 2000; Maddox et al., 2015; Parise & Ernst, 2016; Parise et al., 2012, 2013). Despite demonstrating the importance of correlation on perception, these

studies have been limited to a restricted set of correlations. Because of this, it is unclear whether our perceptual system evaluates stimuli for the presence of correlation or evaluate the strength of correlation between stimuli. This distinction may be important in the context of multiple competing stimuli. For example, the cocktail party problem ambiguous mouth movements can have strong, errant correlations with unrelated acoustic speech envelopes (Chandrasekaran et al., 2009). Evaluating the strength of correlation rather than the presence of sufficient correlation could possibly mitigate this problem.

Recently, we found that behavioral performance does vary with the magnitude of correlation imbedded in dynamic audiovisual stimuli (Nidiffer, Diederich, Ramachandran, & Wallace, 2018), showing that the brain does perform a computation on the strength of audiovisual correlation. However, in that study, participants were making judgements on the presence of amplitude modulation, which was the correlated feature. It has been suggested that to unambiguously dissociate multisensory integration—which encapsulates a large number of neuronal, perceptual, and decisional processes involving the convergence of sensory information—from binding (i.e., object formation), judgements should be made on a stimulus feature orthogonal to the feature that induces the binding (Bizley et al., 2016). This assertion is

rooted in principles of object-based attention (Blaser, Pylyshyn, & Holcombe, 2000; Desimone & Duncan, 1995; Shinn-Cunningham, 2008; Treisman, 1998) where attention, when directed toward an object, enhances all features of that object. For example, to demonstrate audiovisual binding, Maddox (2015) manipulated correlation of the amplitude envelopes of auditory and visual streams while asking participants to discriminate a deviation in the frequency of the auditory stream.

In the current study, we leveraged these principles to determine whether our previous findings extend to multisensory binding. Participants were to make judgements on a frequency modulation feature while changing the strength of correlation in an audiovisual stream. We find that behavioral performance on the orthogonal feature does change with temporal correlation in a linear fashion, though to a lesser degree than when judgements are made on the correlated feature. These results suggest that multisensory integration and binding both depend on the magnitude of similarity between events in the environment.

Results

Multisensory binding is modulated proportional to stimulus correlation

In the first experiment, participants detected near threshold frequency modulation (FM) events imbedded in amplitude modulated (AM) audiovisual streams (Figure 3.1). The phase of the visual AM was adjusted across conditions, thereby changing the correlation between the auditory and visual streams. If the AM correlation is inducing binding of the audiovisual streams, we expect the change in stimulus correlation to induce changes in the perception of the FM event. Figure 3.2 shows discriminability (d') data from two participants plotted across visual phase. There is a clear cyclic pattern in these data (Figure 3.2a). Moreover, there seems to be a phase shift whereby the best discrimination did not occur when the auditor and visual AM streams perfectly overlapped, a finding that was evident and thoroughly tested in our previous report (Nidiffer et al., 2018).

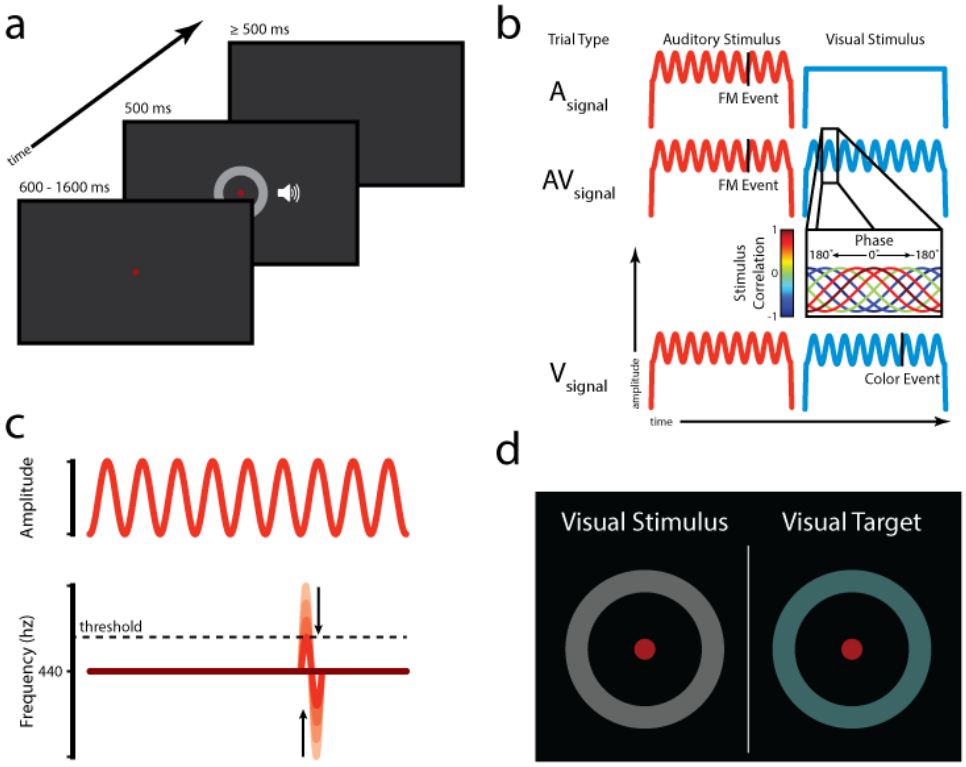


Figure 3.1 (previous page): Stimulus and task. ***(a)*** *Participants fixated a red dot at the center of the display. After a variable time, an amplitude modulated (AM) audiovisual stimulus lasting 1.5 s appeared. Participants were asked to report the presence of a frequency (c) or color (d) deviation target.* ***(b)*** *AM could be present in the auditory stimulus (by modulating the intensity) or both auditory and visual stimuli (by modulating the luminance). Targets consisting of a frequency deviation or color deviation could be present at 1 s after stimulus onset. During audiovisual AM with the frequency deviation, the phase of the visual modulation was varied to generate a range of stimulus correlations (inset).* ***(c)*** *The auditory stimulus consisted of a single pure tone of 440 Hz which could be briefly (100 ms) frequency modulated (FM). FM depth was adjusted to the threshold of each participant via 3-down-1-up staircase procedure prior to the main experiment.* ***(d)*** *The visual stimulus was a grey ring set upon a black background. During a visual event, its color would briefly (100 ms) change to blue.*

In an effort to account for this phase shift, we fit these data to a sine wave and extracted the phase parameter (see methods) to measure the deviation of the peak from zero. This is necessary for calculating each participant's perceptual correlation. Previously, we have shown that this phase shift can be accounted for by iteratively fitting phase-shifted matrices (8 phases × 5 frequencies) to discriminability matrices, consistent with a timing difference between unisensory systems. However, since we only measured discriminability at one frequency, we sought to compare these two methods using data from the previous study (Nidiffer et al., 2018). Discriminability data from one frequency were fit to the same sine wave used here. Figure 3.2b shows this measure of phase shift as plotted against the phase shifts obtained previously. These two methods produce comparable results when considering participants that had reliable estimates of their phase shift. Not only are they strongly correlated (rho = 0.95, p = 0.02), but their values are very consistent (slope = 0.92, CI: [0.73, 1.13]).

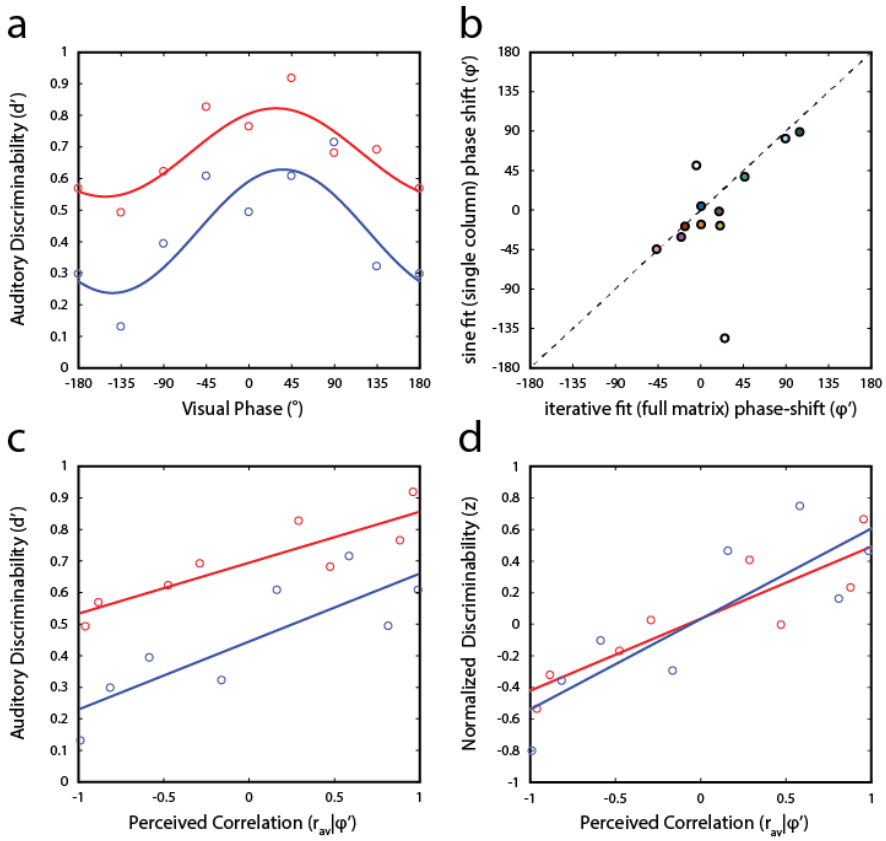


Figure 3.2: Individual participant data. ***(a)*** *Discriminability (d')—a measure of sensitivity—of the frequency event was calculated and is shown plotted against the visual phase. Each phase maps onto a different stimulus correlation. A cyclic relationship between visual phase and auditory discriminability is evident. As shown previously, best performance is not aligned to the best correlated stimuli in the environment (0°) but is instead shifted along phase. We fit this curve with a sine wave and extracted the phase shift parameter to recompute correlation in perceptual space.* ***(b)*** *Might change to phase shift distribution.* ***(c)*** *Auditory discriminability plotted against perceived audiovisual correlation which accounts for the phase shift in (a). Discriminability is clearly modulated by correlation in a linear manner. Although this relationship is evident across the sample, individual participants performed at different levels.* ***(d)*** *Normalized discriminability was calculated in order to account for individual differences in performance. The z-score of d' was taken for each condition and within each participant and is shown plotted here as a function of perceived audiovisual correlation.*

This phase shift was applied to the audiovisual AM envelope and the correlations based on these new envelopes were computed. Participants' discriminability is plotted against these stimulus correlations in Figure 3.2c. So that we can directly compare participants who are performing at different levels, we normalized each participants' discriminability data (Figure 3.2d). In line with our hypothesis, discriminability was significantly modulated by stimulus correlation across participants (Figure 3.3a; r = 0.39 CI: [0.18 0.57], p = 0.00033). This relationship was positive such that stronger correlations in the AM feature was associated with better discriminability of the orthogonal FM feature, and thus influenced binding in a proportional manner.

In the previous work (Nidiffer et al., 2018), participants detected AM in similar stimuli. In that study, participants similarly showed a linear increase in discriminability of the AM, an effect that can't be unambiguously attributed to the binding of the stimuli (Bizley et al., 2016). In both studies, the dependence on correlation was measured by finding the slope of the best fitting line between stimulus correlation and discriminability. The effect sizes for individual participants were significantly different between the current (r = 0.40±0.20) and previous investigation (r = 0.62±0.25; t_{27} = -2.65, p = 0.013; Figure 3.2b), the dependence of binding on

correlation (slope = 0.09±0.04) was significantly smaller than the dependence of integration on correlation (slope = 0.37±0.22; t_{27} = -5.15, p = 2.0×10^{-5}; Figure 3.2c). The difference across the most positive and negative correlation (Δd' = 0.18) is comparable to a study that used only two extremes of similarity (Δ d' ≈ 0.2; Maddox et al., 2015).

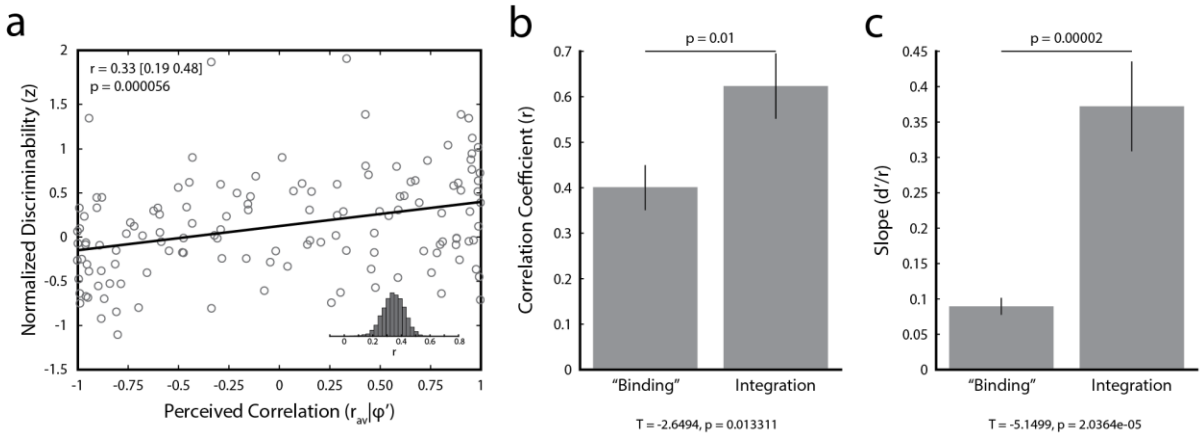


Figure 3.3: Group binding data. ***(a)*** *Normalized discriminability across all participants is significantly correlated to perceived temporal correlations. A bootstrapped distribution of correlations (inset) does not cross below zero.* ***(b)*** *Correlation coefficients (mean ± s.e.m.) from each participant's non-normalized data in the current study where a stronger correlation of audiovisual amplitude modulations leads to an increase in discriminability of the orthogonal frequency event ("Binding") and a previous study where stronger correlation of audiovisual AM leads to increased detection of those envelopes (Integration) are significantly different, indicating the effect of correlation is stronger on integration than binding.* ***(c)*** *Slope (mean ± s.e.m.) of the line of best fit to non-normalized data in the current and previous study are different, indicating a lower dependence on correlation in the current study.*

AM phase shift is related to unisensory temporal processing

Previously, we found that individuals show unique differences in their perception of multisensory correlation that could be accounted for by lagging one modality with respect to the other (Nidiffer et al., 2018). We hypothesized that the resultant measure, phase shift, is related to individual differences in unisensory processing speeds. There are several known phenomena that are related to unisensory differences. First, reaction speeds have been shown to be different for auditory and visual stimuli and lead to different asynchronies that result in peak multisensory gain across participants (Hershenson, 1962). We had an *a priori* expectation that measures taken from the RT task would correlate with phase shift due to the similarity in the tasks and probably similarity in neural architecture underlying the two behaviors (Bizley, Jones, et al., 2016). Second, point of subjective simultaneity and binding window width which accounts for temporal differences in our perception of audiovisual synchrony and relate to sensory differences and differences in propagation in the environment (Zampini, Guest, Shore, & Spence, 2005). Due to the difference between this task, synchrony judgement, and our task, detection, we expected the correlation between these measures and ours to be weak at best. Two additional experiments

were conducted in order to quantify measures of unisensory and multisensory temporal processing and to relate them to phase shifts obtained in the current experiment.

In Experiment 2, participants reacted as quickly as possible to auditory, visual, and audiovisual stimuli. Audiovisual stimuli were presented a different stimulus onset asynchronies (SOAs), including objective synchrony. Figure 3.4a shows reaction times for auditory (red line), visual (blue line), and audiovisual (black line) stimuli in a single subject. The multisensory reaction times across SOAs were subtracted from a prediction based on the unisensory reaction times while accounting for stimulus lag for that condition (Figure 3.4b-c). These curves were fit to a Gaussian function (R^2 = 0.59±0.22). Unisensory RT differences (Figure 3.4a, blue minus red lines) and peak RT gain (Figure 3.4b, mean parameter of Gaussian fit) were computed for each participant. Experiment 3 consisted of a simultaneity judgement task. Audiovisual stimuli were presented with a larger range of SOAs and participants were asked to report whether they perceived the stimuli to be synchronous or asynchronous, emphasizing accuracy over speed. The probability of the perception of synchrony was calculated across SOAs (Figure 3.4d-e). These curves were fit to the same Gaussian function (R^2 = 0.95±0.03; Figure 3.4d) and we measured the point of subjective simultaneity (mean) and window width (2×standard deviation).

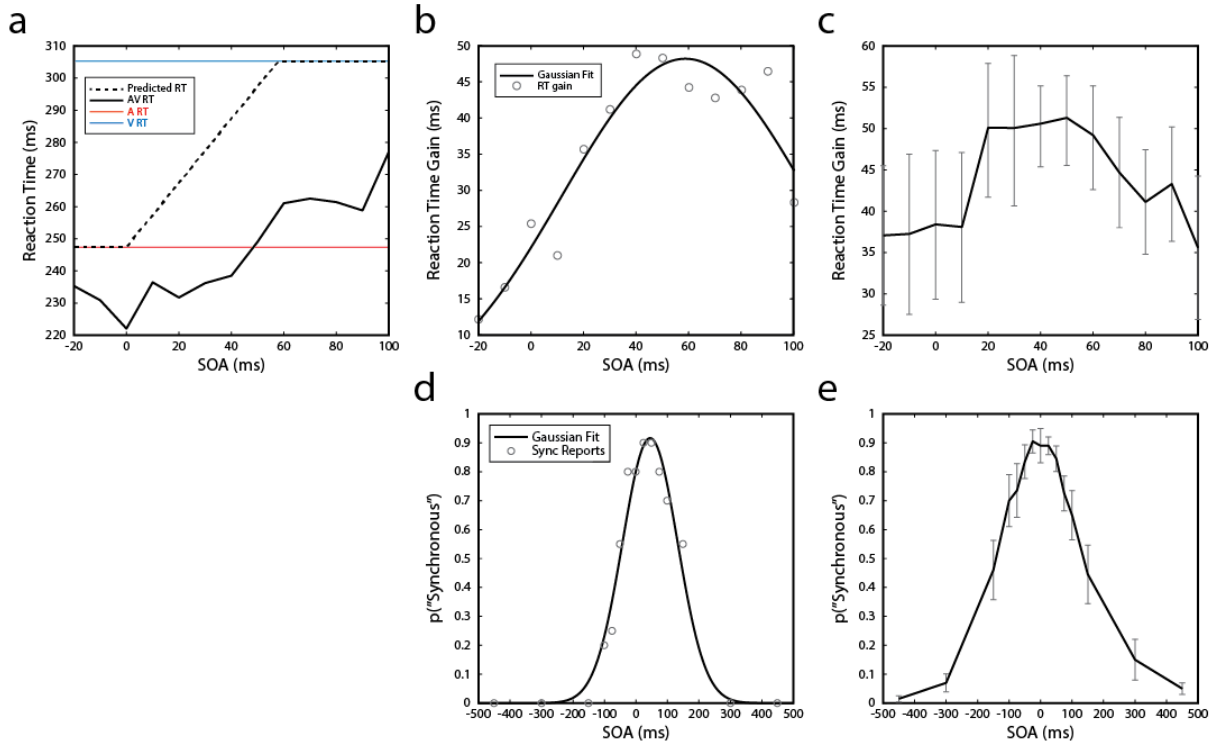


Figure 3.4. Individual and group reaction time and simultaneity judgement data. ***(a)*** *Reaction time data from an individual participant in Experiment 2. Auditory (red line), visual (blue line) and audiovisual (solid black line) reaction times are plotted against SOA. Unisensory reaction time lines are extended across SOAs for illustrative purposes. Auditory reaction time is faster than visual reaction times, which was typical for most participants. A prediction (dashed line) of audiovisual reaction time computed from the unisensory data while accounting for lag imposed by SOA is presented for comparison with empirical audiovisual reaction times. Audiovisual reaction times typically were faster than this prediction.* ***(b)*** *Reaction time gain for a single participant from (a) was calculated by subtracting obtained audiovisual reaction times from their predictions. This curve was fit with a Gaussian function for each participant.* ***(c)*** *Reaction time gain (mean ± s.e.m.) across all participants show gains across typical SOAs.* ***(d)*** *Simultaneity judgement data from the same participant in (a) is shown across SOAs. This curve was fit with a Gaussian function for each participant.* ***(e)*** *Simultaneity judgements (mean ± s.e.m.) across all participants exhibit typical temporal binding windows and the typical rightward shift.*

Each of these measures, unisensory RT difference, peak RT gain, PSS, and window width are plotted against phase shift in Figure 3.5. The difference between auditory and visual reaction times (50.1±34.9 ms) along with the delay corresponding to peak RT gain (42.6±48.6 ms) were in line with those reported previously (Hershenson, 1962). Although the mean PSS (7.4±29.3 ms) was slightly lower that what has been previously reported (Zampini et al., 2005), the distributions overlap considerably. In comparison, window size in the current study (429.9±203.1) was considerably larger than was reported in that study (Zampini et al., 2005). However, the stimuli used in that study were very punctate (9 ms vs. 166 ms in the current study) and longer stimuli are known to produce wider temporal windows (Vroomen & Stekelenburg, 2011; Wallace & Stevenson, 2014). When comparing to phase shift, neither temporal window measures, PSS and window width, were correlated with phase shift. However, both RT difference ($r = 0.63$, $p = 0.031$) and peak RT ($r = 0.62$, $p = 0.032$) were significantly correlated with phase shift, indicating that they likely share a common mechanism.

In summary, we have extended our previous findings, showing that in addition to integration, multisensory binding is linearly related to stimulus correlation. We have further

demonstrated that a phase shift, which was demonstrated here and previously, is related to other forms of uni- and multisensory temporal processing.

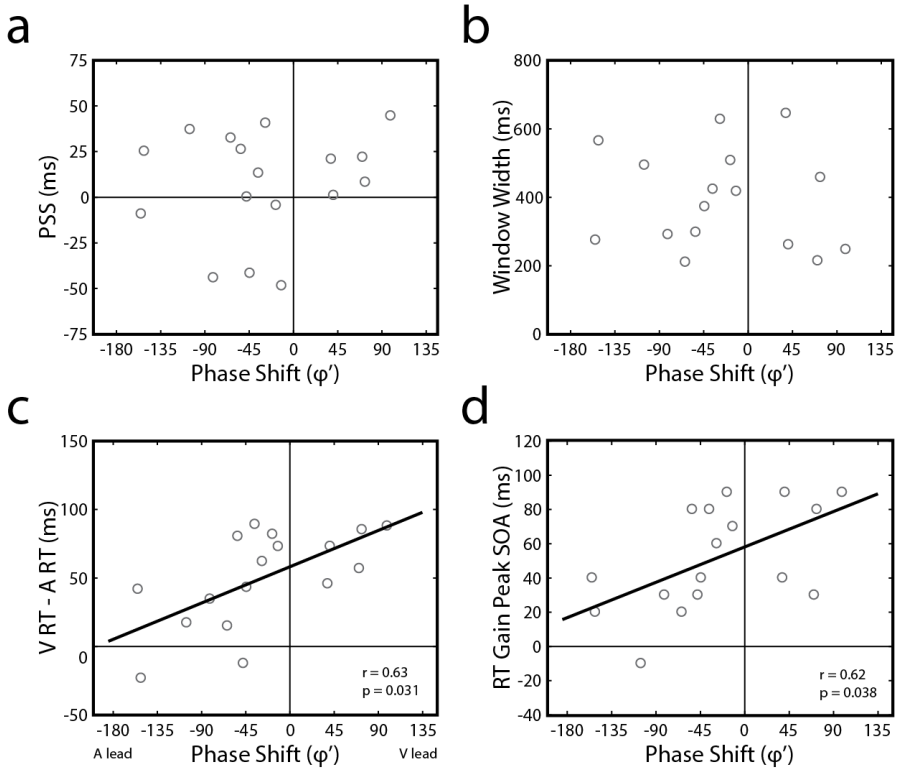


Figure 3.5. Phase shift correlates with unisensory reaction time differences. *(a) Point of subjective simultaneity (PSS) is plotted against phase shifts obtained in Experiment 1.* ***(b)*** *Temporal binding window width is plotted against phase shifts obtained in Experiment 1.* ***(c)*** *Unisensory RT differences are correlated phase shifts obtained in Experiment 1.* ***(d)*** *RT peak gain SOA is plotted against phase shifts obtained in Experiment 1.*

Discussion

Temporal correlation has been shown to influence a variety of multisensory processes. Detection is improved (Nidiffer et al., 2018), cue combination is statistically optimal (Parise et al., 2012), spatial and temporal segregation is more difficult (Chuen & Schutz, 2016; Jack & Thurlow, 1973; Parise et al., 2013; Vatakis & Spence, 2007), and selective attention is improved (Maddox et al., 2015). Previously, we showed that this relationship scales with the magnitude of the correlation, which could potentially provide a means for selection of appropriate signals in challenging environments (Nidiffer et al., 2018). Here, we build up on this foundation, using tenets of object-based attention (Desimone & Duncan, 1995; Shinn-Cunningham, 2008), to show that multisensory binding occurs proportional to the level of correlation.

The fact that multisensory interaction scales with the strength of correlation requires the neural computation of the degree of similarity between environmental signals. Further, that this process can influence the perception of other features is evidence that cross-modal binding is stochastic and dependent on the strength of correlation. Given that our environment is composed of auditory and visual signals such as speech that span a range of correlations (Chandrasekaran et al., 2009), this computation may underlie the appropriate association of

those signals in complex acoustical environments where a binary determination of correlation may be insufficient.

Griffiths and Warren (2004) proposed four principles that define perceptual objecthood, specifically for auditory objects. Among those principles is the notion that information in the sensory world related to an object is separable from other sensory information. Temporal correlation (referred to as coherence) has been shown to be a strong cue for the binding and segregation of sound sources (Elhilali et al., 2009). This process has further been shown to depend on the degree of coherence (O'Sullivan et al., 2015; Teki et al., 2013). In short, we can form an auditory object based on the temporal correlation among its features and segregate that object from an uncorrelated background. Our ability to form a perceptual object and segregate a background is enhanced by a correlated visual stimulus (Maddox et al., 2015). Since multisensory binding is proportional to the strength of the correlation, a natural extension of the current work is to directly test whether auditory stream segregation is proportionally enhanced by a visual stimulus in the same way by leveraging stimulus competition in a stream segregation task while manipulating the strength of the correlation of a visual stream.

Previously, it was demonstrated that rhythmic audiovisual stimuli with the strongest correlations (i.e., those with the same frequency and phase) do not necessarily drive the best behavioral performance (Nidiffer et al., 2018). Instead, performance was found to be shifted along the phase dimension. A likely candidate for the mechanism underlying this process is differences in stimulus processing time between the sensory systems, especially those involving the entrainment of oscillations. Although there is no direct evidence of such a process, there are several pieces of tangential evidence. First, unisensory entrainment is different across individuals (Henry & Obleser, 2012; Simon & Wallace, 2017), suggesting some variability in the entrainment process that might also manifest across modalities. Second, entrainment timing is malleable under conditions involving audiovisual onset timing differences (Kösem et al., 2014). Here we provide further support for this hypothesis in showing that phase shift is related to unisensory reaction time differences.

Although detection speed and oscillatory entrainment are largely separate phenomena, it is possible that their differences across the sensory systems are a manifestation of general timing differences. It is unclear whether the phase relationship discussed here is preserved across frequencies or is simply a timing difference between the sensory systems. Further, whether this

phenomenon has any bearing on audiovisual binding of more complex, arrhythmic stimuli remains to be seen. One study relating stimulus correlation to multisensory integration found that correlation lead to optimal cue combination, irrespective of onset timing (i.e., cross-correlation; Parise et al., 2012). However, the stimuli used in that study were arrhythmic auditory and visual events, rather than continuous fluctuations, and the authors made no attempt to analyze any potential effect of sensory delay.

Our findings and those of others (Atilgan et al., 2018; Maddox et al., 2015), that audiovisual temporal correlation enhances the discrimination of a frequency feature, suggest a pathway in which visual speech enhances the discrimination of acoustic speech. The cortical representation of an auditory stream, both speech and non-speech, is enhanced by a congruent visual stream (Atilgan et al., 2018; Crosse, Butler, & Lalor, 2015). Visual congruence also enhances the representation of orthogonal frequency features embedded in the auditory stream (Atilgan et al., 2018). Frequency features and their cortical representation are important for speech perception (Elliott & Theunissen, 2009; Mesgarani & Chang, 2012) and frequency discrimination ability has been linked to speech perception (Nan et al., 2018).

Materials and Methods

Participants

17 individuals (age = 25.4 ± 4.9, 10 females) participated in the current study. All participants reported normal or corrected-to-normal vision and normal hearing and were right handed. The study was conducted in accordance with the declaration of Helsinki, and informed written consent was obtained from all participants. All procedures were approved by the Vanderbilt University Institutional Review Board. When applicable, participants were given monetary compensation or course credit for participation.

Experiment 1

Apparatus and stimuli

All stimuli were generated in MATLAB (The MathWorks, Inc., Natick, MA) and presented using PsychToolbox version 3(Brainard, 1997; Kleiner et al., 2007). Auditory stimuli were digitized at 44.1 kHz, and presented through calibrated open-back circumaural headphones (Sennheisser HD480). Visual stimuli were centered about a red fixation dot at the center of a dark (0.15 cd/m^2) viewing screen (Samsung Sync Master 2233rz, 120 Hz refresh rate).

The auditory stimulus was a single frequency (440 Hz) tone presented at moderate level (48 dB SPL, A-weighted). Visual stimuli consisted of a moderately bright ring (24 cd/m^2; inner diameter: 1.8°, outer diameter: 3.6° visual angle) over a black background. Both stimuli were presented simultaneously, lasted 1.5 s, and were gated by a linear 10 ms onset and offset ramp. Stimulus timing was confirmed with a Hameg 507 oscilloscope, photodiode, and microphone.

The amplitude of each stimulus, *y*, was modulated over time, *t*, such that

$$y(t) = [1 + m(t)] \times c(t)$$

where

$$m(t) = M \times \cos\left(2\pi f_{am} t + \varphi_{0,j}\right)$$

and *c(t)* is the time series of the carrier stimulus (auditory: tone; visual: ring). The form of the amplitude modulation (AM) signal *m(t)* is defined by a modulation depth *M* which represents the amplitude of the modulation signal as a proportion of the amplitude of the carrier signal, modulation frequency f_{am} in Hz ($f_{am} = 6$), and starting phase $\varphi_{0,j}$ in degrees. The auditory stimulus could be frequency modulated (FM) such that

$$c(t) = \cos\left(2\pi f_c t + \frac{f_\Delta}{f_{fm}} \sin\left(2\pi f_{fm} t\right)\right)$$

where f_c is the frequency of the tone in Hz ($f_c = 440$), f_Δ is the deviation of the frequency modulation in Hz, and f_{fm} is the frequency of the FM in Hz ($f_{fm} = 10$). A FM event was implemented using the *fmmod()* in MATLAB and briefly (100 ms, one full FM cycle) occurred 1 second after stimulus onset so that it occurred at exactly the same phase of the AM to obviate any behavioral dependence on AM phase (Henry & Obleser, 2012).

The AM signal was always present in the auditory channel with modulation depth set to M = 0.5 and starting phase set so that the modulation begins at the trough ($\varphi_{0,j} = 0°$). Visual AM could be present (M = 0.5) or absent (M = 0). Thus, this configuration produced four separate conditions based on a factorial design included Auditory AM (no visual AM) vs Audiovisual AM and FM event present (go trials) vs no FM event (catch trials). Visual modulation during audiovisual go trials occurred with various starting phases ($\varphi_0 = \{-135, -90, -45, 0, 45, 90, 135, 180°\}$, with $\varphi_{0,j} \in \varphi_0$). Because we were interested in the interactions between auditory and visual stimuli on an orthogonal feature, but the task could be carried out with only the auditory stimuli, we included a visual target condition (go trial) to ensure participants were observing both auditory and visual stimuli. In this condition the color of the visual stimulus changed gradually

from its base color (RGB = {100,100,100}) to blue-green (RGB = {60,100,100}) and back over 100ms. The visual and auditory target never occurred in the same trial.

Procedure

Participants were seated comfortably inside an unlit WhisperRoom™ (SE 2000 Series) with their forehead placed against a HeadSpot™ (University of Houston Optometry) with the forehead rest locked in place such that a participant's primary eye position was centered with respect to the fixation point at the center of the viewing screen. Chinrest height and chair height were adjusted to the comfort of the participant.

Prior to the main experiment, each participant completed a 3-down 1-up staircase procedure to obtain an estimate of their FM deviation (f_Δ) threshold. For the staircase procedure, on a given trial (Figure 3.1a), the red fixation dot appeared at the center of the screen. Participants were instructed to fixate the dot for its entire duration. After a variable time, the auditory stimulus was presented. The FM event was presented at random for each trial. Participants were instructed to report the presence of the FM event (described as a "frequency deviant") after the stimulus presentation by pressing "1" on the number pad of a computer keyboard if the modulation was present or pressing "0" if the modulation was absent. The f_Δ

decreased after three successive correct responses and increased after one incorrect response. At the beginning of each staircase, the step size was set to increase or decrease f_Δ by 5Hz. After two reversals (correct to incorrect response or incorrect to correct response), step size was reduced to 2Hz. Finally, after eight reversals, step size became 1Hz in order to arrive at an accurate estimate threshold. Each staircase terminated after 20 reversals. Threshold was determined to be the average of the modulation depth at the last 10 reversals.

Instructions included an example of a stimulus with FM at the initial starting modulation depth ($f_\Delta = 30$) and an example of a stimulus with no FM. So that there was no ambiguity in cases where the first trial did not include a modulation signal, participants were informed that the first trial would have the same modulation depth as the example if present. To control for "runs" of trials with no modulation during the staircase (which could result in erroneously low threshold estimates), a sequence of two trials containing no modulation was always followed by a trial with modulation. At the conclusion of the staircase, the experimenter visually inspected the staircase for its typical asymptotic form and had participants repeat the procedure if necessary.

The main experiment consisted of two blocks lasting approximately 25 minutes each. Each block consisted of 20 trials of each stimulus condition (160 signal trials per block).

Additionally, there were catch (no signal) trials (10% of total trials, 23 trials per block), visual (visual color change) trials (10%, 23) auditory only (no visual AM) trials (5%, 11), and auditory only catch trials (5%, 11). Therefore, each block was identical in trial composition (229 total trials per block) but with individual trials presented in a predetermined, pseudorandom order. Each participant completed a total of 458 trials over the two blocks. Breaks were offered frequently (every 100 trials) to prevent fatigue. Participant were given the opportunity to stretch and walk around while the experimenter set up the next block. Trials during the main experiment were identical to staircase trials with two exceptions. First, in each trial, both auditory and visual stimuli were presented. Modulation signals could be present in the auditory channel alone (auditory only and auditory catch), in both (with phase configuration discussed above), or neither channel (no signal catch trials). Second, participants were told that they should respond as soon as they had made their decision and were instructed to respond as quickly and accurately as possible. In addition to the participant's choice, response times were recorded for each trial, sampling every 2.2 μs (4.6 kHz).

Analysis

Discriminability (*d'*; a measure of sensitivity) for each of the 8 audiovisual conditions and two unisensory conditions was computed from the relative frequencies of the respective responses,

$$d' = z(H_i) - z(F)$$

where H_i is the proportion of hits ("1" | FM stimulus) for the ith condition, F is the proportion of false alarms ("1" | no FM stimulus) from the corresponding catch trial condition, and z is the inverse of the normal distribution function (MATLAB's *norminv* function). *d'* was organized into a matrix in the same manner as the stimulus correlation matrix. Because the proportion of catch trials was held low and errors had no associated cost (Green & Swets, 1966), participants could potentially adopt a strategy of simply pressing "1" which would result in a correct choice more often than not. To account for this, criterion (*c*; a measure of bias) for each participant was computed in a similar manner such that

$$c = z(H) + z(F)$$

where H is the proportion of hits across all conditions. A single criterion was computed for each participant.

Because we have previously shown that perception of rhythmic audiovisual stimuli is shifted across phase, we fit d' data for the audiovisual AM conditions to the function

$$d' = \mathrm{a} \times \cos\left(\varphi_{0,j} + \varphi'\right) + b$$

where $\varphi_{0,j}$ is the starting phase of the visual stimulus. The fitting returns parameters φ', which is the phase shift and a and b influence the magnitude and shift of the function, respectively. We recalculated the correlations using the phase shift parameter as described previously(Nidiffer et al., 2018) and these phase-shifted correlations were used to measure the dependence of behavior on stimulus correlation. To normalize d', which was on different scales for different participants, the z-score of each d' value was calculated for each participant individually. All participant data were pooled and a Pearson correlation was computed for stimulus correlation and discriminability. Bootstrapped correlation distribution by computing the Pearson correlation on pairs of stimulus correlation and its corresponding discriminability, sampled with replacement.

Experiment 2 and 3

Apparatus and stimuli

The stimuli were brief (166 ms) and consisted of the same visual ring described above and a frozen token of broadband auditory noise. The visual stimulus was presented about a fixation

dot that remained on the screen, uninterrupted, for the entirety of the experiments. The noise was generated by MATLAB's *randn()* function and presented diotically at moderate level (48 dB SPL, A-weighted). The amplitude envelope of both stimuli was modulated as described in Experiment 1, but because the duration was 166ms, only one trough-to-trough cycle of AM was present. Thus, the stimuli did not appear to flutter, but had the same envelope characteristics.

Auditory and visual stimuli could be present individually (only in Experiment 2) or together during a trial. When presented together, they were presented synchronously or with a stimulus onset asynchrony (SOA) by delaying one stimulus relative by a short interval. For Experiment 2 SOA = {-±10, ±20, 30, 40, 50, 60, 70, 80, 90, and 100ms}. For Experiment 3, SOA = {±25, ±50, ±75, ±100, ±150, ±300, ±450ms}. Negative values indicate that the visual stimulus occurred after the auditory stimulus.

Procedure

Experiments 2 and 3 occurred between the two blocks of Experiment 1 and their order was randomized across participants. Both experiments took 12-15 minutes to complete. Participants were seated comfortably in the experiment room as detailed above. Experiment 2 was a speeded response task (Hershenson, 1962). A trial consisted of the presentation of a visual,

auditory, or audiovisual stimulus with their temporal relationship as described above. Participants were asked to react as quickly as possible to any stimulus, auditory or visual, with a button press. Reaction time was recorded as the interval of time between stimulus onset (the first stimulus in the case of asynchronous audiovisual stimuli) and the button press. Experiment 3 was a simultaneity judgement task (Zampini et al., 2005). Each trial consisted of the presentation of an audiovisual stimulus with temporal relationships as described above. After each trial, a response screen appeared, prompting participants to indicate on the number pad of a keyboard whether each audiovisual pair were presented synchronously (by pressing "1") or not (by pressing "0"). Participants were asked to take their time and answer as accurately as possible. For both experiments, 20 trials of each condition were presented and trials were separated by a variable time from $U(1,3)$.

Analysis

For Experiment 2, median reaction times (RTs) for auditory, visual, and audiovisual presentations were calculated. The difference of median auditory and visual RTs was calculated. A prediction of audiovisual RTs across SOA was computed by taking the minimum of each auditory or visual RT while accounting for the lag imposed by the SOA(Hershenson, 1962).

Audiovisual RTs were subtracted from this prediction to yield a measure of RT facilitation. These data were then fit to a Gaussian curve:

$$f(x) = ae^{\frac{(x-\mu)^2}{2\sigma^2}}$$

where μ represents the SOA at peak multisensory gain. For experiment 3, the proportion of synchronous judgement was computed at each SOA. These data were fit to a Gaussian curve as above, where the μ represents the point of subjective simultaneity (PSS) and 2σ is a measure of the width of the so-called temporal binding window.

References

Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., Lee, A. K. C., & Bizley, J. K. (2018). Integration of Visual Information in Auditory Cortex Promotes Auditory Scene Analysis through Multisensory Binding. Neuron, 97(3), 640–655.e4. https://doi.org/10.1016/j.neuron.2017.12.034

Bizley, J. K., Jones, G. P., & Town, S. M. (2016). Where are multisensory signals combined for perceptual decision-making? Current Opinion in Neurobiology. https://doi.org/10.1016/j.conb.2016.06.003

Bizley, J. K., Maddox, R. K., & Lee, A. K. C. (2016). Defining Auditory-Visual Objects: Behavioral Tests and Physiological Mechanisms. Trends in Neurosciences. https://doi.org/10.1016/j.tins.2015.12.007

Blaser, E., Pylyshyn, Z. W., & Holcombe, A. O. (2000). Tracking an object through feature space. Nature, 408(6809), 196–199. https://doi.org/10.1038/35041567

Brainard, D. H. (1997). The Psychophysics Toolbox. Spatial Vision, 10, 433–436. https://doi.org/10.1163/156856897X00357

Bregman, A. S. (1990). Auditory scene analysis : the perceptual organization of sound. MIT Press.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. PLoS Computational Biology, 5(7). https://doi.org/10.1371/journal.pcbi.1000436

Chuen, L., & Schutz, M. (2016). The unity assumption facilitates cross-modal binding of musical, non-speech stimuli: The role of spectral and amplitude envelope cues. Attention, Perception, and Psychophysics, 78(5), 1512–1528. https://doi.org/10.3758/s13414-016-1088-5

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. Journal of Neuroscience, 35(42), 14195–14204. https://doi.org/10.1523/JNEUROSCI.1829-15.2015

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. Annual Review of Neuroscience, 18(1), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009). Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. Neuron, 61(2), 317–329. https://doi.org/10.1016/j.neuron.2008.12.005

Elliott, T. M., & Theunissen, F. E. (2009). The Modulation Transfer Function for Speech Intelligibility. PLoS Computational Biology, 5(3), e1000302. https://doi.org/10.1371/journal.pcbi.1000302

Grant, K. W., & Seitz, P. F. 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 Pt 1), 1197–1208. https://doi.org/10.1121/1.422512

Green, D. M., & Swets, J. A. (1966). Signal detection theory and psychophysics. Society, 1, 521. https://doi.org/10.1901/jeab.1969.12-475

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? Nature Reviews Neuroscience, 5(11), 887–892. https://doi.org/10.1038/nrn1538

Henry, M. J., & Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. Proceedings of the National Academy of Sciences, 109(49), 20095–20100. https://doi.org/10.1073/pnas.1213390109

Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. Journal of Experimental Psychology, 63(3), 289–293. https://doi.org/10.1037/h0055703

Jack, C. E., & Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. Perceptual and Motor Skills, 37(3), 967–979. https://doi.org/10.2466/pms.1973.37.3.967

Kleiner, M., Brainard, D. H., Pelli, D. G., Broussard, C., Wolf, T., & Niehorster, D. (2007). What's new in Psychtoolbox-3? Perception, 36, S14. https://doi.org/10.1068/v070821

Kösem, A., Gramfort, A., & Van Wassenhove, V. (2014). Encoding of event timing in the phase of neural oscillations. NeuroImage, 92, 274–284. https://doi.org/10.1016/j.neuroimage.2014.02.010

Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. ELife, 2015(4), 1–11. https://doi.org/10.7554/eLife.04995.001

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature, 485(7397), 233–236. https://doi.org/10.1038/nature11020

Nan, Y., Liu, L., Geiser, E., Shu, H., Gong, C. C., Dong, Q., … Desimone, R. (2018). Piano training enhances the neural processing of pitch and improves speech perception in Mandarin-speaking children. Proceedings of the National Academy of Sciences of the United States of America, 201808412. https://doi.org/10.1073/pnas.1808412115

Nidiffer, A. R., Diederich, A., Ramachandran, R., & Wallace, M. T. (2018). Multisensory perception reflects individual differences in processing temporal correlations. BioRxiv, 264457. https://doi.org/10.1101/264457

O'Sullivan, J. A., Shamma, S. A., & Lalor, E. C. (2015). Evidence for Neural Computations of Temporal Coherence in an Auditory Scene and Their Enhancement during Active Listening. Journal of Neuroscience, 35(18), 7256–7263. https://doi.org/10.1523/JNEUROSCI.4973-14.2015

Parise, C. V., & Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. Nature Communications, 7(12), 364. https://doi.org/10.1038/ncomms11543

Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. Current Biology, 22(1), 46–49. https://doi.org/10.1016/j.cub.2011.11.039

Parise, C. V, Harrar, V., Ernst, M. O., & Spence, C. (2013). Cross-correlation between Auditory and Visual Signals Promotes Multisensory Integration. Multisensory Research, 26, 1–10. https://doi.org/10.1163/22134808-00002417

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. Trends in Cognitive Sciences, 12(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Simon, D. M., & Wallace, M. T. (2017). Rhythmic Modulation of Entrained Auditory Oscillations by Visual Inputs. Brain Topography, 30(5), 565–578. https://doi.org/10.1007/s10548-017-0560-4

Teki, S., Chait, M., Kumar, S., Shamma, S., & Griffiths, T. D. (2013). Segregation of complex acoustic scenes based on temporal coherence. ELife, 2, e00699. https://doi.org/10.7554/eLife.00699

Treisman, A. (1998). Feature binding, attention, and object perception. Essent. Sources Sci. Study Conscious., 8, 226. https://doi.org/10.1016/j.ejpn.2004.03.003

Vatakis, A., & Spence, C. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. Perception & Psychophysics, 69(5), 744–756. https://doi.org/10.3758/BF03193776

Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. Cognition, 118(1), 78–86. https://doi.org/10.1016/j.cognition.2010.10.002

Wallace, M. T., & Stevenson, R. A. The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities, 64 Neuropsychologia § (2014). https://doi.org/10.1016/j.neuropsychologia.2014.08.005

Wertheimer, M. (1923). Untersuchungen zur Lehre yon der Gestalt II. Psychologische Forschung, 4, 301–350.

Zampini, M., Guest, S., Shore, D. I., & Spence, C. (2005). Audio-visual simultaneity judgments. Perception & Psychophysics, 67(3), 531–544. https://doi.org/10.3758/BF03193329

Chapter 4. How similarity and proximity shape two multisensory processes: binding and integration

The Multisensory World Revisited

Events (and objects) in our environment produce a rich diversity of sensory energies which correspond to different features of those events. Somehow, our sensory systems effortlessly collect, transduce, and process this information about the features of the environment, often from unreliable signals. One of the great challenges our brains overcome in constructing our perceptual experience is sorting through these neural signals and deciding which correspond to the same event. Often, signals across the sensory modalities provide redundant or complementary information about single events or objects in the environment. The brain is able to combine these multisensory signals synergistically, and thus enhance the representation of and our interaction with the environment (Murray & Wallace, 2012; Stein, 2012).

From the earliest descriptions, it's clear that the benefits of multisensory integration lie in the enhancement of unisensory representations that are weak (Crosse et al., 2016; Meredith &

Stein, 1986b; Sumby & Pollack, 1954) or the complementary combinations of strengths across unisensory systems (e.g., the speed of audition and spatial accuracy of vision; Corneil et al., 2002). These interactions are beneficial typically when sensory signals are informative of the same event. The presence of a cross-modal signal that comes from a source that is different from a target often interferes with behavioral performance (Amlôt et al., 2003; Corneil & Munoz, 1996; c.f., Murray et al., 2005). Therefore, critical in the integration process is determining whether two signals come from a common event, a process termed causal inference (Shams & Beierholm, 2010). In this way, multisensory integration and the benefits it confers can be thought of as consequence of multisensory binding, but as we'll discuss in this chapter, this relationship isn't as simple as this.

Several stimulus factors have been demonstrated to contribute to multisensory binding and causal inference. These include the spatial and temporal proximity between unisensory stimuli (Körding et al., 2007; Magnotti et al., 2013) and their similarity, based on temporal correlation (Bizley et al., 2016; Parise & Ernst, 2016). Even though we make the case in this chapter that proximity lacks specificity for multisensory binding on its own, it is undoubtedly a feature of stimuli that belong together and plays a key role in multisensory integration. The

question remains as to the relationship between proximity and similarity cues and their relationship to multisensory binding. In this chapter we will build on the idea that integration and binding are two distinct multisensory processes (Bizley et al., 2016). We argue that proximity and similarity are stimulus properties that drive integration and binding, respectively, albeit nonexclusively. Further, we propose a connection where the manifestations of multisensory integration related to stimulus proximity (i.e., the spatial and temporal principles) are shaped by stimuli the brain has learned to bind via stimulus similarity during development. We frame this connection through a set of predictions that can be tested and we outline a series of experiments aimed at testing them.

Two Multisensory Worlds: Proximity and Similarity

One of the hallmarks of sensory signals that originate from a common event is their proximity in space and time. In line with this, multisensory interactions in neurons become more prevalent and powerful with increasing spatially and temporally proximity of unisensory stimuli (e.g., the so-called spatial and temporal principles; Meredith et al., 1987; Meredith & Stein, 1986a). Multisensory behaviors are constrained by these same principles that govern interactions in neurons (Stein et al., 1988). These principles were shown to influence human behaviors as

well (Bolognini et al., 2005; Frassinetti et al., 2002; Frens et al., 1995). Together, these phenomena are loosely interpreted as evidence that spatial and temporal proximity serve as cues to our multisensory system during causal inference (Körding et al., 2007; Magnotti et al., 2013).

However, stimulus proximity and the resultant multisensory enhancements are often at odds with binding and even occurs following stimuli that do not belong to a common physical event. When asked to indicate whether spatially and temporally disparate multisensory stimuli appear "unified," participants report a perception of unity even for stimuli with low spatial and temporal proximity (up to 15° and 800 ms of disparity; Wallace et al., 2004). The spatial and temporal principles are commonly demonstrated with stimuli that are highly artificial and have arbitrary pairings across modalities, such as simple flashes and noise or tone bursts (Hershenson, 1962) or moving visual bars of light with stationary noise bursts (Meredith et al., 1987; Meredith & Stein, 1986a). Multisensory interactions are even observed when the relationship between the unisensory stimuli is explicitly dissociated (e.g., during focused attention; Bolognini et al., 2005; Colonius, 2010; Van Wanrooij, Bremen, & John Van Opstal, 2010). To our knowledge there has been no attempt to account for this apparent inconsistency.

Another multisensory stimulus factor—similarity (e.g., temporal correlation)—and its yet unexplored relationship to proximity might represent the key to accounting for this discrepancy. Correlation over time is known to be important in unisensory feature binding (Bizley & Cohen, 2013; Blake & Lee, 2005; Elhilali et al., 2009; S. H. Lee & Blake, 1999; Shinn-Cunningham, 2008) and more recently, it has been demonstrated as a cue that can be utilized to bind multisensory signals (Maddox et al., 2015; Parise et al., 2012, 2013). Multisensory temporal correlation reflects changes that are shared between sensory signals when a single dynamic event produces energy across multiple modalities, such as audiovisual speech (Chandrasekaran et al., 2009). In line with this, improvements in speech in noise (Ross et al., 2007; Sumby & Pollack, 1954) have been attributed to a temporal envelope present in visible mouth movements of a speaker that is shared with the amplitude envelope of acoustic speech (Bernstein et al., 2004; Grant & Seitz, 2000; Kim & Davis, 2004; Munhall et al., 1996). The presence of a speaking face alone is not sufficient for these enhancements. Furthermore, the presentation of a simple visual stimulus where the size changes with the envelope of an acoustic speech signal confers a small but reliable benefit in detecting that speech in noise (Bernstein et

al., 2004), illustrating the importance of temporal correlation independent of other factors related to the face on our multisensory perception.

There appears to be a trade-off between the effects of similarity and proximity on multisensory perception. Temporal similarity between multisensory stimuli can compensate for a lack of proximity in time or space, resulting in larger just-noticeable differences in detecting spatial and temporal conflict compared to uncorrelated audiovisual stimuli (Chuen & Schutz, 2016; Parise et al., 2013; Vatakis, Navarra, Soto-Faraco, & Spence, 2007). In fact, the phenomenon known as the ventriloquism effect, where the binding of a dummy's mouth movements and the ventriloquist's voice can cause the perception that the dummy is the source of the voice, is dependent on the temporal similarity between the auditory and visual streams (Jack & Thurlow, 1973; Thurlow & Jack, 1973). Similar effects have been reported for other forms of similarity, such as the learned association between a steaming kettle and a whistle sound (C. V. Jackson, 1953), indicating that binding and the compensation of the lack of proximity can be driven by influences that are not based on temporal correlation.

Findings from a recent experiment (Nidiffer et al., 2018) were reanalyzed in order to empirically disentangle proximity and similarity as they relate to another principle of

multisensory integration: inverse effectiveness (Meredith & Stein, 1983; Sumby & Pollack, 1954). In brief, participants were asked to detect amplitude modulation (AM; e.g., "flutter") in audiovisual stimuli. The AM could be present in unisensory channels independently or in both channels. When present in both channels, the temporal correlation was manipulated through changes in frequency and/or phase of the auditory stimulus. Figure 4.1a-f presents a reanalysis of multisensory data and unreported unisensory data from the experiment. We fit a line relating multisensory discriminability to stimulus correlation (Figure 4.1a). The intercept term of this line was taken as a representation of general multisensory performance related to the presentation of the two AM signals independent of their similarity. This value was compared to the best unisensory performance to derive a proportional measure of multisensory enhancement analogous to integrative index (Stevenson et al., 2014; Figure 4.1a, straight arrows). We compared this to a measure of dependence on similarity: the slope of the same line (Figure 4.1a, curved arrows). In an initial test of their independence, we found no significant relationship between these measures (Figure 4.1b). Further, the magnitude of the enhancement from stimulus proximity in our task was dependent on the unisensory performance (viz. inverse effectiveness), but the benefit afforded by similarity was not (Figure 4.1c-d). Finally, in

agreement with the principle of congruent effectiveness presented by Otto and colleagues (2013) general multisensory benefits were associated with the equivalence of unisensory performance whereas similarity benefits were higher when auditory performance was greater (Figure 4.1e-f). These data, along with those presented earlier demonstrate the separate influences of similarity and proximity in three major principles that guide multisensory processes: space, time, and effectiveness.

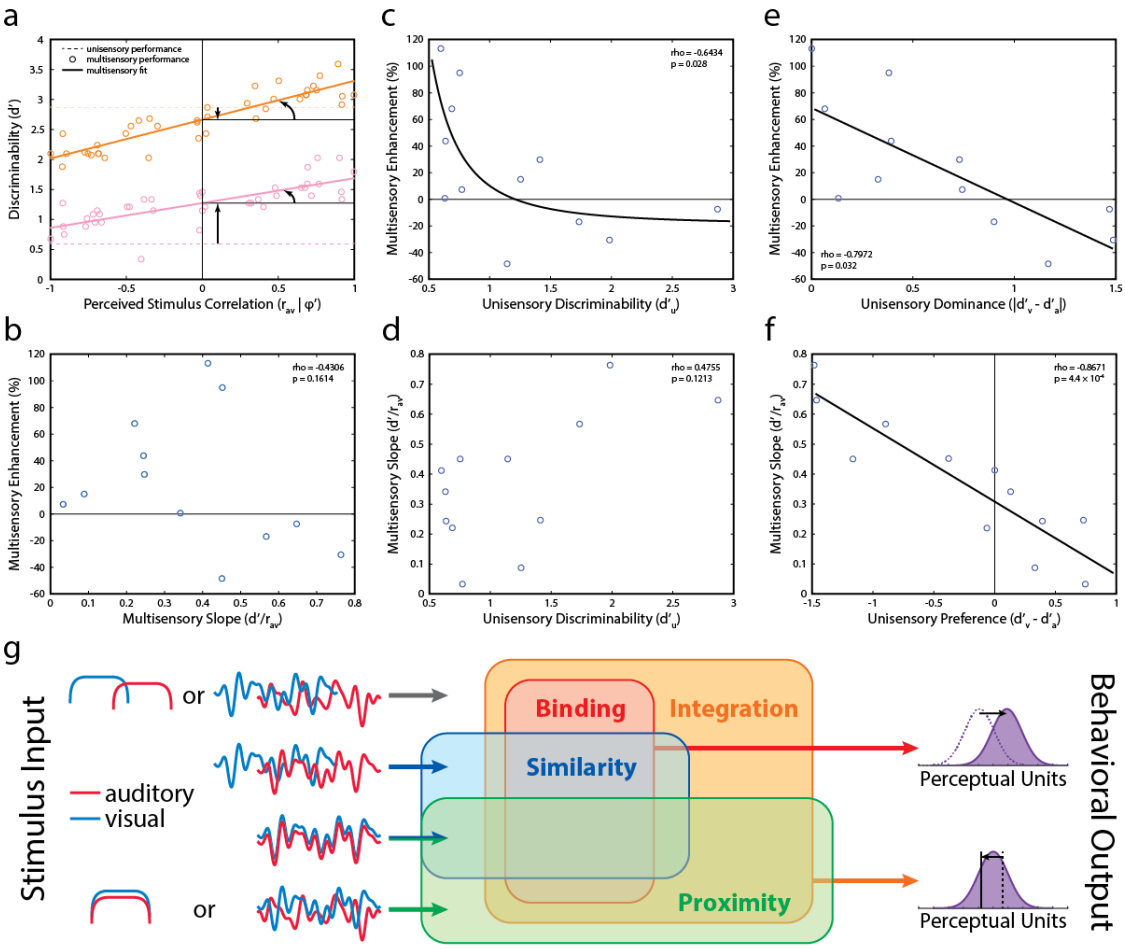


Figure 4.1 (previous page): The relationship between binding/integration and stimulus similarity/proximity. ***(a)*** *To show that the effects of proximity and similarity can be empirically dissociated, behavioral sensitivity for two participants plotted against audiovisual stimulus correlation. A line (solid) was fit to multisensory conditions. The slope of the line (curved arrow) was taken to represent the effect of similarity. The intercept of this line was compared to unisensory performance (dashed line) proportionally (straight arrow) and was taken to represent the effect of proximity (enhancement).* ***(b)*** *The slopes were plotted against enhancements for each participant. No relationship was found between the two measures.* ***(c)*** *Multisensory enhancement is maximal when unisensory performance is low, consistent with inverse effectiveness* ***(d)*** *No relationship between slope and unisensory performance was found.* ***(e)*** *Multisensory enhancement is greatest for participants who exhibited equal auditory and visual performance.* ***(f)*** *Slope is greatest for participants who performed better on the auditory task.* ***(g)*** *Schematic representing the relationship between similarity/proximity (left) and binding/integration (right). Binding specifically results in a perceptual shift while integration often results in a criterion shift (adapted from Bizley et al., 2016).*

The distinction between two processes—integration and binding—that shape multisensory behavior and perception has grown in recent years. Integration constitutes any form of convergence or interaction across the senses. Binding is more restrictive, referring to the process of grouping cross-modal stimulus features into a unified object (Bizley et al., 2016). Multisensory binding is a form of integration, but the reverse is not necessarily true. Figure 4.1g describes the relationship between binding and integration proposed by Bizley and colleagues, which we extend to include stimulus features similarity and proximity. Typically, binding is

driven by some consistency or similarity between dynamically changing unisensory features (e.g., matching amplitude envelopes; Maddox et al., 2015; c.f., Mishra, Martinez, & Hillyard, 2013). We posit that integration, on the other hand, requires only the presence and proximity of stimuli. It makes sense that binding require more constrained stimulus features since it's considered a more restrictive process.

It is clear is that stimulus similarity and proximity are important in shaping our multisensory perceptions and in processes such as feature binding. But due to seemingly conflicting information in the literature, the nature of the relationships between these stimulus features is unclear. Further, their role in binding and causal inference is also unclear. Here we hypothesize a link to explain their relationship whereby low-level spatial and temporal proximity filters which are observable at the level of single neurons (Meredith et al., 1987; Meredith & Stein, 1986a; Stein & Meredith, 1993) and behavior (Bolognini et al., 2005; Frassinetti et al., 2002) are scaffolded and refined by stimulus similarity—which is a feature that is more closely tied to cross-modal signals originating from a single event.

Thus, we propose a developmental link between similarity and proximity. According to this framework, stimulus similarity is a robust cue for which stimuli should be bound and shapes

the brain's current estimate of causal inference. Low-level proximity filters reflect the brain's learned estimate of causal inference and can, in the absence of dynamic similarity between stimuli, drive binding. This learned estimate is built upon stimulus similarity present in environmental stimuli during development. During the inference of a common cause, the proximity and similarity of multisensory stimuli (as well as other factors such as perceptual priors) contribute to this inference differently (e.g., with different weights).

This framework can begin to explain empirical findings related to similarity and proximity that are seemingly discrepant. For example, in a multisensory stimulus where similarity is high, but proximity is low (e.g., a ventriloquist's voice and a dummy's mouth), the brain's judgement of causal inference and the subsequent binding weights similarity over proximity. The framework also is agnostic to potential top-down factors that may influence the binding of simple stimuli (e.g., Mishra et al., 2013; Wallace et al., 2004). This relationship would then become more important in shaping multisensory filters in complex and noisy environments where spatial and temporal proximity are less informative of a common source underlying multisensory signals. Practically, this hypothesis can develop a set of predictions based on the environmental relationship between cross-modal signal onset and the correlation between them.

A Developmental Scaffold—the Relationship Between Similarity and Proximity

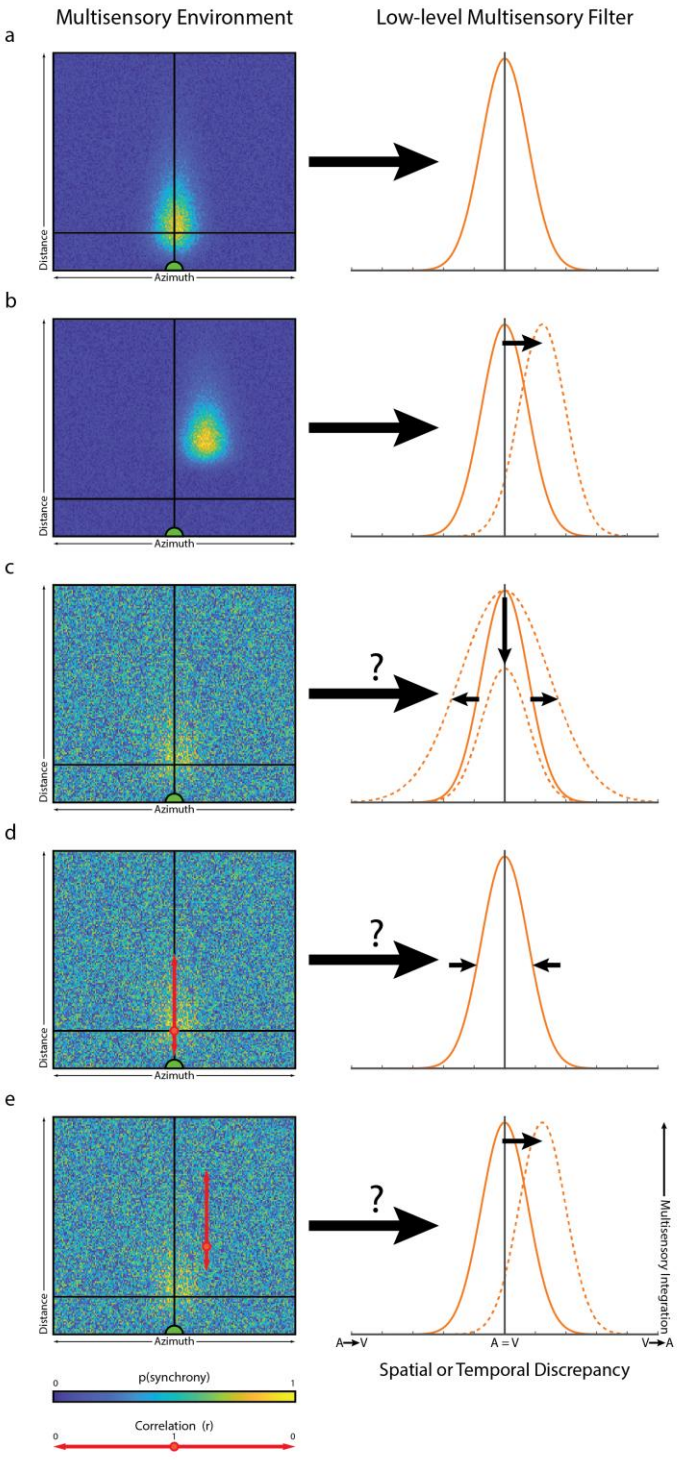
During development, our sensory systems undergo profound changes (Hensch, 2004) including tuning and refinement of neuronal receptive fields. This process is shaped by and depend on experiences in the developmental environment. Importantly, sensory representations tend to adapt to represent the sensory structure that is present in the external world (Rabinowitz, Willmore, Schnupp, & King, 2011; Schwartz & Simoncelli, 2001; Simoncelli & Olshausen, 2001). When any orderly structure is present and there are no sensory features to represent, development is delayed (Chang & Merzenich, 2003). Conversely, an over-abundance of a feature results in its over-represented in the brain (de Villers-Sidani, Chang, Bao, & Merzenich, 2007; Hirsch & Spinelli, 1971). Finally, development in an enriched sensory environment speed the onset of development and enhance the representation of the sensory environment (Cancedda et al., 2004; Engineer et al., 2004).

This experience-dependent refinement is not just restricted to unisensory processing. Multisensory processes are also shaped to represent the environment during development through a process of perceptual narrowing (Lewkowicz & Ghazanfar, 2009). Early in development, multisensory temporal filters are broad (Lewkowicz, 1996) and narrow over the

course of maturation (Hillock, Powers, & Wallace, 2011). Spatial receptive fields of multisensory neurons follow a similar trajectory during development: newborn receptive fields are large and can extend the entire surface of the body or visual or auditory field but are gradually refined over the course of maturation (Wallace, Carriere, Perrault, Vaughan, & Stein, 2006; Wallace & Stein, 1997, 2001). Similar to unisensory neurons, multisensory neurons develop the ability to integrate their inputs during maturation, tuning their spatial and temporal filters to match the statistical structure of their developmental environment (Figure 4.2a-b ; Polley et al., 2008; Wallace & Stein, 2007).

Figure 4.2 (next page). Proposed link between proximity and similarity during multisensory development. *These plots represent the synchronous occurrences of multisensory stimuli in the environment across azimuth (spatial proximity) and distance (temporal proximity). One sensory signal is indicated by the crossed black lines directly in front of an observer (green dot) and the other is represented by a probability cloud. This probability cloud is blurred across the spatial axis due to the variability of our sensory estimates of space and further blurred across the distance/time axis due to the increased discrepancy between sensory timing with increased distance.* ***(a)*** *In a quiet environment with little stimulus competition (left), most coincident multisensory signals originate from the same event and thus come from the same location in space. The likelihood that synchronous signal come from separate regions of space is quite low and thus simple stimuli and their onset times are sufficient to shape and tune multisensory proximity filters (right).* ***(b)*** *When the multisensory spatial and temporal statistics of the environment are changed, multisensory filters shift to reflect that change.* ***(c)*** *We predict that an increase in stimulus competition, with nothing to anchor the binding of auditory and visual stimuli in the environment will broaden multisensory filters, reduce the magnitude of multisensory integration, or both.* ***(d)*** *We predict further that introducing correlation in stimuli should provide a sufficient anchor for*

binding and thus refine multisensory filters. ***(e)*** *If correlation serves as this anchor, shifting the correlation should result in shifted multisensory filters.*



When one event produces multimodal energies, those energies are spatially and temporally coincident. And when there are few events occurring in the environment, the likelihood that separate events occur simultaneously is low and thus coincident signals most often originate from the same location in space and converge on neurons within a narrow time window. In this environment, these spatial and temporal relationships between these coincident unisensory signals may provide sufficient information as to whether a set of events belong together. As a result, low-level multisensory filters built on proximity alone can accurately reflect a common source for these signals (Figure 4.2a). These filters can also shift when spatial and temporal discrepancies are introduced during development (Figure 4.2b). When more events occur, the environment becomes crowded with sensory signals. The chance of separate events producing coincident but spatially disparate energies increases. The result is a degraded or noisy spatial and temporal structure in the environment. In the absence of a more robust cue for appropriate cross-modal association, may decrease the selectivity (i.e., width) of low-level spatial and temporal filters (Figure 4.2c), the overall benefit (i.e., amplitude of enhancement) of multisensory stimulus combination, or both. Empirically, this hypothesis could be tested by measuring multisensory spatial and temporal filters (after Meredith et al., 1987; Meredith &

Stein, 1986a) in animals reared in a simple environment where multisensory coincidence occurs only with spatial and temporal proximity (Figure 4.2a) or environments with increasing levels of sensory complexity driving increased spurious stimulus overlap (Figures 4.2c).

If increasing the stimulus competition in the environment reduces the spatial and temporal specificity of multisensory filters, a separate cue, distinct from spatial and temporal ambiguity, must be responsible for the refinement of spatial and temporal processing. As alluded above, a strong candidate is temporal correlation. In a complex developmental environment, two stimuli might often occur at the same time but come from separate locations. Despite their common onset, their dynamic features (e.g., amplitude envelope) are far less likely to be correlated. Following this prediction, adding correlation to spatially and temporally proximal multisensory stimuli would restore the specificity of spatial and temporal receptive fields in multisensory neurons (Figure 4.2d). This prediction can be tested experimentally by rearing animals in the same noisy environment as before, but where signals have complex amplitude envelopes (after Maddox et al., 2015) with environmentally coincident stimuli (simulated "events") sharing a common fluctuation(Figure 4.2d, red line).

This developmental relationship can be further tested by probing the relative strength of correlation in shaping low-level filters. As above, animals can be reared in a complex sensory environment. If correlation serves as a true anchor for appropriately associating stimuli and if spatial and temporal filters are reflective of the brains prediction of that association, then if the correlation is applied to spatially and/or temporally disparate signals during development, we would observe a shift in those filters (Figure 4.2e). If present, the magnitude of this shift relative to the disparity that is present during development would provide a measure of the strength of the scaffold that correlation provides to the low-level filters (after Jay, Martha F, Sparks, 1984). We can further probe this by varying the noisiness of the environment and measuring the resultant filter shift. Further, the limits can be tested by correlating stimuli with different levels of spatial and temporal disparity.

A mechanism that could potentially be responsible for establishing low-level filters through correlation is correlated spike-timing-dependent plasticity (STDP). STDP strengthens synaptic inputs when those inputs occur just before the post-synaptic neuron fires (Sjöström, Rancz, Roth, & Häusser, 2008) and in itself is a related to the degree of correlation between pre and postsynaptic activity (olde Scheper, Meredith, Mansvelder, van Pelt, & van

Ooyen, 2018). This logic has been extended to account for how multisensory neurons are tuned through development. A neural network model that has been shown to produce physiological features of mature multisensory neurons (Cuppini, Ursino, Magosso, Rowland, & Stein, 2010; Magosso, Cuppini, Serino, Di Pellegrino, & Ursino, 2008; Ursino, Cuppini, Magosso, Serino, & di Pellegrino, 2009) is shaped during a developmental "training phase" where Hebbian rules refine and align spatial receptive fields of the simulated neurons according to the unisensory inputs (Cuppini, Magosso, Rowland, Stein, & Ursino, 2012; Cuppini, Stein, Rowland, Magosso, & Ursino, 2011). Importantly, when developmental exposure consisted of misaligned unisensory inputs, the mature model produced integrated responses only when inputs were appropriately misaligned. Despite the simple nature of the input "stimuli" used during the development of these neural networks, it is plausible that correlation across dynamic inputs could induce similar development of multisensory neuronal features.

Concluding Remarks

Neural and behavioral enhancements conferred by multisensory integration are likely the result of the brain inferring a common origin of multisensory stimuli. Indeed, these interactions are most useful when they involve the combination of sensory inputs from a single

event or object. However, as illustrated in this chapter, these interactions can occur in response to paired multisensory stimuli that do not strictly belong together (e.g., artificial stimuli such as flashes and beeps), as long as they occur in close spatial and temporal proximity. Spatial and temporal proximity become less important in guiding multisensory enhancements when stimuli are correlated (i.e., similar) over time.

We've presented a framework that attempts to bridge this disparity. Patterns of multisensory integration based on the proximity of multisensory stimuli (e.g., spatial and temporal filters) reflect the statistics of multisensory stimuli that the brain has learned to associate with a common origin. Stimuli that originate from a single event, because of their spatial and temporal proximity, *should* be bound. Our framework posits that this bridge is built during development while the brain is adapting to the statistics of the environment. We predict that this bridge is guided by stimulus similarity when stimulus proximity becomes an ineffective cue for linking signals. This is possible because similarity is more selective of a common event than proximity.

Following others (Bizley et al., 2016), we acknowledge that the process of multisensory integration is separate from binding. We extend this logic to relate two stimulus features—

similarity and proximity—to each other and ascribe each (loosely) to a multisensory process. It's further likely that these stimulus features (and by extension their ascribed processes) tap into different multisensory operations, again loosely. Multisensory operations associated with proximity are mostly "hard-wired" (Alvarado, Rowland, Stanford, & Stein, 2008; Magosso et al., 2008; Rowland, Stanford, & Stein, 2007; Ursino et al., 2009) but still susceptible to the statistics of the environment during development (Cuppini et al., 2012) and . This operation likely bestows enhancements based on a special case of neural integration that depends on the dendritic morphology of excitatory inputs and those inputs being balanced by parallel local inhibitory interneurons which synapse close to the soma (Rowland et al., 2007). Binding, on the other hand, represents an instantiation of a more flexible architecture involving synchronization across neural populations (Engel, Roelfsema, Fries, Brecht, & Singer, 1997; Fries, 2005; Schroeder & Lakatos, 2008; Womelsdorf et al., 2007). Such a division of multisensory labor has been proposed before. Proximity—being a "classical" integration cue—may rely on the divisive normalization operation (Ohshiro et al., 2011) and similarity may drive the oscillatory operations (Lakatos, Karmos, Mehta, Ulbert, & Schroeder, 2008) in the scheme proposed by Van Atteveldt

and colleagues (2014). Taking this into account, there exist at least two "flavors" of multisensory processes which depend on two distinct operations and two separate stimulus features.

References

Alvarado, J. C., Rowland, B. A., Stanford, T. R., & Stein, B. E. (2008). A neural network model of multisensory integration also accounts for unisensory integration in superior colliculus. Brain Research, 1242, 13–23. https://doi.org/10.1016/J.BRAINRES.2008.03.074

Amlôt, R., Walker, R., Driver, J., & Spence, C. (2003). Multimodal visual-somatosensory integration in saccade generation. Neuropsychologia, 41(1), 1–15. https://doi.org/10.1016/S0028-3932(02)00139-2

Bernstein, L. E., Auer, E. T., & Takayanagi, S. (2004). Auditory speech detection in noise enhanced by lipreading. Speech Communication, 44(1–4), 5–18. https://doi.org/10.1016/J.SPECOM.2004.10.011

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. Nature Reviews Neuroscience. https://doi.org/10.1038/nrn3565

Bizley, J. K., Maddox, R. K., & Lee, A. K. C. (2016). Defining Auditory-Visual Objects: Behavioral Tests and Physiological Mechanisms. Trends in Neurosciences. https://doi.org/10.1016/j.tins.2015.12.007

Blake, R., & Lee, S.-H. (2005). The role of temporal structure in human vision. Behavioral and Cognitive Neuroscience Reviews, 4(1), 21–42. https://doi.org/10.1177/1534582305276839

Bolognini, N., Frassinetti, F., Serino, A., & Làdavas, E. (2005). "Acoustical vision" of below threshold stimuli: Interaction among spatially converging audiovisual inputs. Experimental Brain Research, 160(3), 273–282. https://doi.org/10.1007/s00221-004-2005-z

Cancedda, L., Putignano, E., Sale, A., Viegi, A., Berardi, N., & Maffei, L. (2004). Acceleration of visual system development by environmental enrichment. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 24(20), 4840–4848. https://doi.org/10.1523/JNEUROSCI.0845-04.2004

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. PLoS Computational Biology, 5(7). https://doi.org/10.1371/journal.pcbi.1000436

Chang, E. F., & Merzenich, M. M. (2003). Environmental noise retards auditory cortical development. Science (New York, N.Y.), 300(5618), 498–502. https://doi.org/10.1126/science.1082163

Chuen, L., & Schutz, M. (2016). The unity assumption facilitates cross-modal binding of musical, non-speech stimuli: The role of spectral and amplitude envelope cues. Attention, Perception, and Psychophysics, 78(5), 1512–1528. https://doi.org/10.3758/s13414-016-1088-5

Colonius. (2010). The optimal time window of visual–auditory integration: a reaction time analysis. Frontiers in Integrative Neuroscience, 4, 1–8. https://doi.org/10.3389/fnint.2010.00011

Corneil, B. D., & Munoz, D. P. (1996). The Influence of Auditory and Visual Distractors on Human Orienting Gaze Shifts. The Journal of Neuroscience, 16(24), 8193–8207.

Corneil, B. D., Van Wanrooij, M. M., Munoz, D. P., & Van Opstal, A. J. (2002). Auditory-visual interactions subserving goal-directed saccades in a complex scene. Journal of Neurophysiology, 88(1), 438–454. https://doi.org/10.1038/377059a0

Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration. Journal of Neuroscience, 36(38), 9888–9895. https://doi.org/10.1523/JNEUROSCI.1396-16.2016

Cuppini, C., Magosso, E., Rowland, B. A., Stein, B. E., & Ursino, M. (2012). Hebbian mechanisms help explain development of multisensory integration in the superior colliculus: a neural network model. Biological Cybernetics, 106, 691–713. https://doi.org/10.1007/s00422-012-0511-9

Cuppini, C., Stein, B. E., Rowland, B. A., Magosso, E., & Ursino, M. (2011). A computational study of multisensory maturation in the superior colliculus (SC). Experimental Brain Research, 213(2–3), 341–349. https://doi.org/10.1007/s00221-011-2714-z

Cuppini, C., Ursino, M., Magosso, E., Rowland, B. A., & Stein, B. E. (2010). An emergent model of multisensory integration in superior colliculus neurons. Frontiers in Integrative Neuroscience, 4(6). https://doi.org/10.3389/fnint.2010.00006

de Villers-Sidani, E., Chang, E. F., Bao, S., & Merzenich, M. M. (2007). Critical period window for spectral tuning defined in the primary auditory cortex (A1) in the rat. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 27(1), 180–189. https://doi.org/10.1523/JNEUROSCI.3227-06.2007

Elhilali, M., Ma, L., Micheyl, C., Oxenham, A. J., & Shamma, S. A. (2009). Temporal Coherence in the Perceptual Organization and Cortical Representation of Auditory Scenes. Neuron, 61(2), 317–329. https://doi.org/10.1016/j.neuron.2008.12.005

Engel, A. K., Roelfsema, P. R., Fries, P., Brecht, M., & Singer, W. (1997). Role of the temporal domain for response selection and perceptual binding. Cerebral Cortex, 7(6), 571–582. https://doi.org/10.1093/cercor/7.6.571

Engineer, N. D., Percaccio, C. R., Pandya, P. K., Moucha, R., Rathbun, D. L., & Kilgard, M. P. (2004). Environmental Enrichment Improves Response Strength, Threshold, Selectivity, and Latency of Auditory Cortex Neurons. Journal of Neurophysiology, 92(1), 73–82. https://doi.org/10.1152/jn.00059.2004

Frassinetti, F., Bolognini, N., & Làdavas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. Experimental Brain Research, 147(3), 332–343. https://doi.org/10.1007/s00221-002-1262-y

Frens, M. A., Van Opstal, A. J., Van der Willigen, R. F., Opstal, a J. Van, & Willigen, R. F. Van Der. (1995). Spatial and temporal factors determine auditory-visual interactions in human saccadic eye movements. Perception & Psychophysics, 57(6), 802–816. https://doi.org/10.3758/BF03206796

Fries, P. (2005). A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence. Trends in Cognitive Sciences. https://doi.org/10.1016/j.tics.2005.08.011

Grant, K. W., & Seitz, P. F. 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 Pt 1), 1197–1208. https://doi.org/10.1121/1.422512

Hensch, T. K. (2004). Critical Period Regulation. Annual Review of Neuroscience, 27(1), 549–579. https://doi.org/10.1146/annurev.neuro.27.070203.144327

Hershenson, M. (1962). Reaction time as a measure of intersensory facilitation. Journal of Experimental Psychology, 63(3), 289–293. https://doi.org/10.1037/h0055703

Hillock, A. R., Powers, A. R., & Wallace, M. T. (2011). Binding of sights and sounds: Age-related changes in multisensory temporal processing. Neuropsychologia, 49(3), 461–467. https://doi.org/10.1016/j.neuropsychologia.2010.11.041

Hirsch, H. V. B., & Spinelli, D. N. (1971). Modification of the distribution of receptive field orientation in cats by selective visual exposure during development. Experimental Brain Research, 12(5), 509–527. https://doi.org/10.1007/BF00234246

Jack, C. E., & Thurlow, W. R. (1973). Effects of degree of visual association and angle of displacement on the "ventriloquism" effect. Perceptual and Motor Skills, 37(3), 967–979. https://doi.org/10.2466/pms.1973.37.3.967

Jackson, C. V. (1953). Visual Factors in Auditory Localization. Quarterly Journal of Experimental Psychology, 5(2), 52–65. https://doi.org/10.1080/17470215308416626

Jay, M. F., & Sparks, D. L. (1984). Auditory receptive fields in primate superior colliculus shift with changes in eye position. Nature, 309(24), 345–347. https://doi.org/10.1038/309345a0

Kim, J., & Davis, C. (2004). Investigating the audio–visual speech detection advantage. Speech Communication, 44(1–4), 19–30. https://doi.org/10.1016/J.SPECOM.2004.09.008

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal Inference in Multisensory Perception. PLoS ONE, 2(9), e943. https://doi.org/10.1371/journal.pone.0000943

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of Neuronal Oscillations as a Mechanism of Attentional Selection. Science, 320, 110–113.

Lee, S. H., & Blake, R. (1999). Visual form created solely from temporal structure. Science, 284(5417), 1165–1168. https://doi.org/10.1126/science.284.5417.1165

Lewkowicz, D. J. (1996). Perception of Auditory-Visual Temporal Synchrony in Human Infants. Journal of Experimental Psychology : Human Perception and Performance, 22(5), 1094–1106. https://doi.org/10.1037/0096-1523.22.5.1094

Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through perceptual narrowing. Trends in Cognitive Sciences, 13(11), 470–478. https://doi.org/10.1016/j.tics.2009.08.004

Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. ELife, 2015(4), 1–11. https://doi.org/10.7554/eLife.04995.001

Magnotti, J. F., Ma, W. J., & Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech. Frontiers in Psychology, 4, 798. https://doi.org/10.3389/fpsyg.2013.00798

Magosso, E., Cuppini, C., Serino, A., Di Pellegrino, G., & Ursino, M. (2008). A theoretical study of multisensory integration in the superior colliculus by a neural network model. Neural Networks, 21(6), 817–829. https://doi.org/10.1016/j.neunet.2008.06.003

Meredith, M. A., Nemitz, J. W., & Stein, B. E. (1987). Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. The Journal of Neuroscience, 7(10), 3215–3229. https://doi.org/citeulike-article-id:409430

Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. Science, 221(4608), 389–391. https://doi.org/10.1126/science.6867718

Meredith, M. A., & Stein, B. E. (1986a). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. Brain Research, 365(2), 350–354. https://doi.org/10.1016/0006-8993(86)91648-3

Meredith, M. A., & Stein, B. E. (1986b). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. Journal of Neurophysiology, 56(3), 640–662. https://doi.org/citeulike-article-id:844215

Mishra, J., Martinez, A., & Hillyard, S. A. (2013). Audition influences color processing in the sound-induced visual flash illusion. Vision Research, 93, 74–79. https://doi.org/10.1016/J.VISRES.2013.10.013

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. Perception & Psychophysics, 58(3), 351–362. https://doi.org/10.3758/BF03206811

Murray, M. M., Molholm, S., Michel, C. M., Heslenfeld, D. J., Ritter, W., Javitt, D. C., … Foxe, J. J. (2005). Grabbing your ear: Rapid auditory-somatosensory multisensory interactions in low-level sensory cortices are not constrained by stimulus alignment. Cerebral Cortex, 15(7), 963–974. https://doi.org/10.1093/cercor/bhh197

Murray, M., & Wallace, M. (Eds.). (2012). The Neural Bases of Multisensory Processes. Boca Raton: CRC Press. https://doi.org/10.1201/b11092

Nidiffer, A. R., Diederich, A., Ramachandran, R., & Wallace, M. T. (2018). Multisensory perception reflects individual differences in processing temporal correlations. BioRxiv, 264457. https://doi.org/10.1101/264457

Ohshiro, T., Angelaki, D. E., & DeAngelis, G. C. (2011). A normalization model of multisensory integration. Nature Neuroscience, 14(6), 775–782. https://doi.org/10.1038/nn.2815

olde Scheper, T. V., Meredith, R. M., Mansvelder, H. D., van Pelt, J., & van Ooyen, A. (2018). Dynamic Hebbian Cross-Correlation Learning Resolves the Spike Timing Dependent Plasticity Conundrum. Frontiers in Computational Neuroscience, 11, 119. https://doi.org/10.3389/fncom.2017.00119

Otto, T. U., Dassy, B., & Mamassian, P. (2013). Principles of Multisensory Behavior. Journal of Neuroscience, 33(17), 7463–7474. https://doi.org/10.1523/JNEUROSCI.4678-12.2013

Parise, C. V., & Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. Nature Communications, 7(12), 364. https://doi.org/10.1038/ncomms11543

Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. Current Biology, 22(1), 46–49. https://doi.org/10.1016/j.cub.2011.11.039

Parise, C. V, Harrar, V., Ernst, M. O., & Spence, C. (2013). Cross-correlation between Auditory and Visual Signals Promotes Multisensory Integration. Multisensory Research, 26, 1–10. https://doi.org/10.1163/22134808-00002417

Polley, D. B., Hillock, A. R., Spankovich, C., Popescu, M. V, Royal, D. W., & Wallace, M. T. (2008). Development and plasticity of intra- and intersensory information processing. Journal of the American Academy of Audiology, 19(10), 780–798.

Rabinowitz, N. C., Willmore, B. D. B., Schnupp, J. W. H., & King, A. J. (2011). Contrast Gain Control in Auditory Cortex. Neuron, 70(6), 1178–1191. https://doi.org/10.1016/J.NEURON.2011.04.030

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cerebral Cortex, 17(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

Rowland, B. A., Stanford, T. R., & Stein, B. E. (2007). A Model of the Neural Mechanisms Underlying Multisensory Integration in the Superior Colliculus. Perception, 36(10), 1431–1443. https://doi.org/10.1068/p5842

Schroeder, C. E., & Lakatos, P. (2008). Low-frequency neuronal oscillations as instruments of sensory selection. Trends in Neurosciences, 32(1), 9–18. https://doi.org/10.1016/j.tins.2008.09.012

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. Nature Neuroscience, 4(8), 819–825. https://doi.org/10.1038/90526

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. Trends in Cognitive Sciences, 14(9), 425–432. https://doi.org/10.1016/j.tics.2010.07.001

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. Trends in Cognitive Sciences, 12(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural Image Statistics and Neural Representation. Annual Review of Neuroscience, 24, 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193

Sjöström, P. J., Rancz, E. A., Roth, A., & Häusser, M. (2008). Dendritic Excitability and Synaptic Plasticity. Physiological Reviews, 88(2), 769–840. https://doi.org/10.1152/physrev.00016.2007

Stein, B. E. (Ed.). (2012). The New Handbook of Multisensory Processes. Cambridge, MA: MIT Press.

Stein, B. E., Huneycutt, S. W., & Alex Meredith, M. (1988). Neurons and behavior: the same rules of multisensory integration apply. Brain Research, 448(2), 355–358. https://doi.org/10.1016/0006-8993(88)91276-0

Stein, B. E., & Meredith, M. A. (1993). The Merging of the Sense. Cognitive Neuroscience Series. https://doi.org/10.3389/neuro.01.019.2008

Stevenson, R. A., Ghose, D., Fister, J. K., Sarko, D. K., Altieri, N. A., Nidiffer, A. R., … Wallace, M. T. (2014). Identifying and Quantifying Multisensory Integration: A

Tutorial Review. Brain Topography, 27(6), 707–730. https://doi.org/10.1007/s10548-014-0365-7

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. The Journal of the Acoustical Society of America, 26(2), 212–215. https://doi.org/10.1121/1.1907309

Thurlow, W. R., & Jack, C. E. (1973). Certain determinants of the "ventriloquism effect". Perceptual and Motor Skills, 36, 1171–1184. https://doi.org/10.2466/pms.1973.36.3c.1171

Ursino, M., Cuppini, C., Magosso, E., Serino, A., & di Pellegrino, G. (2009). Multisensory integration in the superior colliculus: a neural network model. Journal of Computational Neuroscience, 26(1), 55–73. https://doi.org/10.1007/s10827-008-0096-4

Van Atteveldt, N., Murray, M. M., Thut, G., & Schroeder, C. E. (2014). Multisensory integration: Flexible use of general operations. Neuron. https://doi.org/10.1016/j.neuron.2014.02.044

Van Wanrooij, M. M., Bremen, P., & John Van Opstal, A. (2010). Acquired prior knowledge modulates audiovisual integration. European Journal of Neuroscience, 31(10), 1763–1771. https://doi.org/10.1111/j.1460-9568.2010.07198.x

Vatakis, A., Navarra, J., Soto-Faraco, S., & Spence, C. (2007). Temporal recalibration during asynchronous audiovisual speech perception. Experimental Brain Research, 181(1), 173–181. https://doi.org/10.1007/s00221-007-0918-z

Wallace, M. T., Carriere, B. N., Perrault, T. J., Vaughan, J. W., & Stein, B. E. (2006). The development of cortical multisensory integration. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 26(46), 11844–11849. https://doi.org/10.1523/JNEUROSCI.3295-06.2006

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. Experimental Brain Research, 158(2), 252–258. https://doi.org/10.1007/s00221-004-1899-9

Wallace, M. T., & Stein, B. E. (1997). Development of Multisensory Neurons and Multisensory Integration in Cat Superior Colliculus. The Journal of Neuroscience, 17(7), 2429–2444.

Wallace, M. T., & Stein, B. E. (2001). Sensory and multisensory responses in the newborn monkey superior colliculus. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 21(22), 8886–8894. https://doi.org/10.1523/JNEUROSCI.21-22-08886.2001

Wallace, M. T., & Stein, B. E. (2007). Early experience determines how the senses will interact. Journal of Neurophysiology, 97(1), 921–926. https://doi.org/10.1152/jn.00497.2006

Womelsdorf, T., Schoffelen, J.-M., Oostenveld, R., Singer, W., Desimone, R., Engel, A. K., & Fries, P. (2007). Modulation of Neuronal Interactions Through Neuronal Synchronization. Science, 316(5831), 1609–1612. https://doi.org/10.1126/science.1139597

Chapter 5. General Discussion

> *"But it is impossible to perceive two objects coinstantaneously in the same sensory act unless they have been mixed, [when, however, they are no longer two], for their amalgamation involves their becoming one, and the sensory act related to one object is itself one, and such act, when one, is, of course, coinstantaneous with itself. Hence, when things are mixed we of necessity perceive them coinstantaneously: for we perceive them by a perception actually one."*
>
> — Aristotle, De Sensu et Sensibilibus

Summary and Implications of Results

The findings reported herein are the first pieces of evidence that the correlation between auditory and visual streams, over time, can influence multisensory feature integration and perceptual binding commensurate with the strength of that correlation. As correlation between the features [i.e., amplitude modulation (AM) envelopes] of the signals increased, so did behavioral performance in tasks designed to quantify integration and binding. First, despite holding the depth, and thus the detectability, of unisensory AM constant across conditions, multisensory configurations in which the two AM envelopes were better overlapped (i.e., were more correlated) improved that detectability of the AM. Unlike many previous findings which

relate multisensory behavioral changes to unisensory differences, the current finding cannot be attributed to changes in the unisensory domains (e.g., inverse effectiveness) but is rather a product of a multisensory comparison. Second, the benefits of (de)correlating amplitude envelopes extended to a separate feature of the stimulus. When the correlation between auditory and visual AM envelopes was increased, the detectability of a frequency modulation (FM) event, improved. According to principles of object-based attention (Desimone & Duncan, 1995; Shinn-Cunningham, 2008), when our attention is directed toward an object based on one feature (in our case, the amplitude envelope), other features (here, the frequency) of that object are enhanced. Therefore, the enhancement of the FM feature is important in showing that correlation in AM is inducing the perceptual binding of the auditory and visual streams (Bizley et al., 2016). Together, these findings are in agreement with the temporal correlation hypothesis (Gray, 1999; Singer & Gray, 1995) and the hypothesis that our multisensory perceptual experience is driven in part by temporal coherence across uni- and multisensory brain regions (Senkowski, Schneider, et al., 2008).

Common across both tasks, we found that behavioral performance was not necessarily best when the stimuli were objectively correlated. Instead, both tasks revealed a transform

between stimulus correlation and participant performance that occurred along the phase dimension and was unique to each participant. This transformation could be accounted for by simulating a delay in one of the sensory signals prior to computing correlation across conditions. This delay simulated differences in temporal processing between the sensory systems. Indeed, a tangential experiment found that differences between auditory and visual reaction times (RTs) correlate with phase shift across participants. Although this phase shift stands up to several tests of its validity, several aspects of its instantiation remain unclear and constitute a direction for future work.

Finally, we presented a perspective on these two multisensory processes, binding and integration, which we argued are loosely shaped by separate stimulus features: similarity and proximity and we proposed a developmental link to bridge the two. We compared two metrics from our data: one measuring the benefit of combining of auditory and visual signals and another measuring the benefit of temporal correlation. We related these to integration and binding, respectively; an integration conforms to known principles of integration [i.e., inverse effectiveness (Meredith & Stein, 1983) and congruent effectiveness (Otto et al., 2013)] while the binding metric does not. The finding that binding can be dissociated from the principles of

integration is in agreement with previous reports which were not specifically aimed at differentiating multisensory integration from binding (Chuen & Schutz, 2016; Denison, Driver, & Ruff, 2013; Parise et al., 2012, 2013; Vatakis & Spence, 2007). Admittedly, this collection of evidence is tangential at best, and so we suggested a series of experiments aimed at testing the relationship between integration and binding directly during development.

Integration versus binding

This issue concerning the differences in binding and integration is a major issue in multisensory research (Bizley et al., 2016; Vatakis & Spence, 2007; Wallace et al., 2004) and in this dissertation. Multisensory integration typically describes the convergence and interaction of sensory information in the brain and often leads to measurable changes in neural activity, perception, decisional processes, or overt behavior (Stein et al., 2010; Stein, Stanford, & Rowland, 2009; Stein & Stanford, 2008). There seems to be a hardline argument that "integration" necessarily involves a differences between multisensory and unisensory responses (Stein et al., 2010), however, the definition used in this dissertation does not hold this as an absolute requirement. The definition is broad as it involves a large and widely varied set of neural mechanisms, experimental techniques, assumptions, and dependent measures. A set of three

general principles—space, time, and effectiveness—have been used to describe multisensory integration and have been repeatedly demonstrated across a variety of domains (see Chapter 1 for a short review of these principles).

Multisensory binding, on the other hand, is a specific integrative process whereby multisensory features are linked to form a single, unified representation of an external event or object (Bizley et al., 2016). Being a specific process, binding carries a restrictive definition. Typically, binding involves some similarity between features (e.g., temporal correlation) of multisensory stimuli and the enhancement of other features present in that object, a finding replicated in the current work. This last component, the enhancements of the features of an object other than the ones carrying the correlation, are also not an absolute requirement for binding. Rather, this is a means to exclude interactions in which the underlying process is ambiguous. For example, if responses are made on the feature that induces binding (as was the case in Chapter 2 where participants detected the AM that carried the temporal correlation), it is difficult to differentiate between a true perceptual enhancement that results from binding and a shift in the decisional criterion.

These similarities that drive binding, especially temporal correlation, appear to operate independently of—and sometimes bend—core multisensory principles. It was shown previously that temporally correlated stimuli can be separated further in time and space than uncorrelated stimuli while still being perceived as spatially and temporally coincident (Chuen & Schutz, 2016; Denison et al., 2013; Parise et al., 2012, 2013; Vatakis & Spence, 2007). Our results presented in Chapter 4 round out the list of core multisensory principles by showing that integration follows the principle of inverse effectiveness but a measure likely reflecting binding by temporal correlation does not. Thus, temporal correlation, especially when it induces multisensory binding, appears to act as a perceptual "glue" that binds features between the senses.

This metaphor of "glue" has been invoked previously in the context of binding. In the Feature-Integration Theory of attention (Treisman & Gelade, 1980), features are encoded in parallel along a number of dimension (e.g., color, orientation, brightness). These features are represented separately until bound together into a perceptual object at a later stage of processing. In this stage, stimulus locations are processed serially by focused attention and features falling in the same locus of attention are combined into a single object. In this model, the "glue" that binds

features together is that attention, which seems to conflict with the notion of temporal correlation as that glue.

Attention can also influence the processing of features across modalities, especially when they are temporally correlated. When participants are listening to auditory speech, directing attention toward matched visual speech increases neural activity compared to attending unmatched visual speech (Fairhall & Macaluso, 2009). Further, enhanced neural activity related to deploying attention toward irrelevant visual speech is associated with poorer recognition of audiovisual speech targets (Senkowski, Saint-Amour, Gruber, & Foxe, 2008). This apparently discrepancy between bottom-up (e.g., temporal correlation as the "glue") and top-down (e.g., attention as the "glue") processing was addressed more recently (Talsma, Senkowski, Soto-Faraco, & Woldorff, 2010). Briefly, Talsma proposed that in the absence of stimulus competition, integration (and by extension, binding) of a temporally-correlated target such as audiovisual speech may be automatic. But when other stimuli or tasks are introduced that capture attention, neural resources are diverted toward the extra stimulus and away from the target. Focusing attention on the target then becomes necessary to properly integrate that target, a process that is enhanced when that target is temporally correlated (Maddox et al., 2015;

O'Sullivan & Lalor, 2017). The glue metaphor can then be extended. If attention provides the "glue" that binds features, then temporal correlation [or some other form of consistency (Bizley et al., 2016)] provides the structure on which the glue can adhere. When the structure is better matching (e.g., has increased temporal correlation), the bond becomes stronger which provides a more salient signal for attentional capture.

Although it may be possible to differentiate binding and integration theoretically and empirically, it is only logical that the two are linked. Enhancements bestowed by simple multisensory integration help guide our action only in the context of appropriate causal inference (Körding et al., 2007; Magnotti et al., 2013; Schutz & Kubovy, 2009; Shams & Beierholm, 2010). Therefore, multisensory enhancements that can be measured with stimuli with no environmental relationship, such as the spatial and temporal principles (Bolognini et al., 2005; Frassinetti et al., 2002) must be shaped somehow by the binding process. In Chapter 4, we propose that temporal correlation is this scaffold. A natural environment is noisy with stimuli that often overlap in onset time but not in space. A multisensory system built only on stimulus timing would have spatial and temporal filters that are broad and non-specific. These filters

ought to be refined during development by a cue that is less ambiguous as to the relationship between two sensory signals. We proposed that this cue is temporal correlation.

Binding through neural synchrony

It has been suggested that the coordination of activity across brain regions is accomplished by the synchronization of rhythmic fluctuations in those regions (Fries, 2005; Gray, 1999; Senkowski, Schneider, et al., 2008). Neural synchrony and oscillatory entrainment has been identified or implicated as important factor in sensory awareness (Engel & Singer, 2001), attentional selection (Lakatos et al., 2008; Womelsdorf & Fries, 2007), flexible routing of information in the brain (Fries, 2005; Womelsdorf et al., 2007), and sensory feature binding (Hipp et al., 2011; Singer & Gray, 1995), which was the focus of this dissertation.

Multisensory binding, and the subsequent improvement in perception, is driven in part by oscillatory synchronization (Engel, Senkowski, & Schneider, 2007; Senkowski, Schneider, et al., 2008). Low-frequency neural oscillations reflect the pattern of neural excitability over time (Bishop, 1933) and therefore influence the probability of neural firing and even stimulus-related activity (Buzsaki & Draguhn, 2004; Lakatos et al., 2005). Behavioral performance has been linked to oscillatory phase in both auditory (Henry et al., 2014; Henry & Obleser, 2013;

Neuling, Rach, Wagner, Wolters, & Herrmann, 2012; Ng, Schroeder, & Kayser, 2012; Okada, 1994) and visual (Busch, Dubois, & VanRullen, 2009; Cravo, Rohenkohl, Wyart, & Nobre, 2013; Mathewson, Gratton, Fabiani, Beck, & Ro, 2009) tasks. Oscillations appears to be particularly important in the context of temporally fluctuating stimuli (Henry et al., 2014; Schroeder, Wilson, Radman, Scharfman, & Lakatos, 2010) where endogenous neuronal oscillations readily entrain to low-frequency fluctuations of sensory inputs (Thut et al., 2011), and offer a dynamic prediction about upcoming signals (Engel, Fries, & Singer, 2001).

Oscillations are important in the processing and binding of audiovisual speech. Endogenous oscillations entrain to the rhythmic structure of acoustic speech (Luo & Poeppel, 2007; Peelle, Gross, & Davis, 2013; Zoefel, Archer-Boyd, & Davis, 2018) and these oscillations control local spiking activity through hierarchical organization (Giraud & Poeppel, 2012; Lakatos et al., 2005). The ability of the auditory cortex to encode the structure of acoustic speech (Pasley et al., 2012) is correlated with its comprehension (Ahissar et al., 2001). Visual speech is known to also entrain oscillations in cortex (O'Sullivan, Crosse, Di Liberto, & Lalor, 2017). Access to temporally congruent audiovisual speech cues enhances the synchrony of brain activity to that speech in both auditory and visual cortices (Park, Kayser, Thut, & Gross, 2016;

Schroeder, Lakatos, Kajikawa, Partan, & Puce, 2008). Thus, the cortical representation of on-going speech is enhanced by seeing the face of the speaker (Crosse et al., 2015) and this benefit is larger in noisy conditions where the representation is degraded (Crosse et al., 2016).

The utility of oscillations in multisensory binding becomes apparent in the presence of competing stimuli. The entrainment of neural oscillations has been proposed as a mechanism attentional selection (Lakatos et al., 2008; Schroeder & Lakatos, 2008) which is of particular importance when there are competing streams, for example when trying to focus on one speaker among a group of speakers (Zion-Golumbic & Schroeder, 2012; Zion Golumbic et al., 2012). When listeners are listening to a mixture of competing speech streams, neural activity in auditory cortex is able to encode and reconstruct the attended speech stream as if the listener heard that speaker alone (Mesgarani & Chang, 2012). The representation of attended speech in a multi-speaker environment is enhanced during the presentation of a congruent visual stream (O'Sullivan & Lalor, 2017). Consistent with theories of object-based attention (Bizley et al., 2016; Desimone & Duncan, 1995; Shinn-Cunningham, 2008), this process may be mediated by enhancements in features (e.g., frequency or timbre) orthogonal to those that are correlated in audiovisual speech (e.g., amplitude; Maddox et al., 2015). Similarly, during the presentation of

competing auditory speech-like streams, a visual stream that is temporally congruent with one of the auditory streams enhances the representation of that stream and its other features in single neurons in auditory cortex (Atilgan et al., 2018).

Interestingly, oscillatory synchrony across large cortical networks can mediate binding in the absence of rhythmically fluctuating stimuli. In the stream-bounce illusion, two identical visual objects—typically disks or bars of light—appear at the left and right sides of a display and travel toward the opposite side of the display. Their paths coincide mid-way through the presentation which is most often interpreted perceptually as the objects passing through each other. However, on rare presentations, some observers report the perception of the disks colliding and then bouncing apart (Metzger, 1934). The introduction of an auditory token at the moment of coincidence of the visual objects biases reports in favor of a "bounce" percept (R. Sekuler et al., 1997). The auditory stimulus can be decoupled from the visual stimulus by presentation of identical auditory tokens before and after the auditory token, thus forming and auditory stream that is overall unmatched from the visual stream. When this happens, report the bounce percept less often (Watanabe & Shimojo, 2001). Critically, when the frequency of the middle auditory token is different from the flanking tokens, thus ungrouping the auditory stream

(Darwin, 1992), the perception of a bounce is recovered. Therefore, the perception of a "bounce" is dependent on the binding of the visual streams with the auditory event. During identical multisensory presentations of this illusion, two long-range cortical networks involving frontal, parietal, and occipital regions and across multiple frequency bands become more synchronized when observers bind the auditory token with the visual streams and report the perception of the visual stimuli bouncing (Hipp et al., 2011), illustrating the importance of neural synchrony on audiovisual binding, even in the absence of temporal correlation.

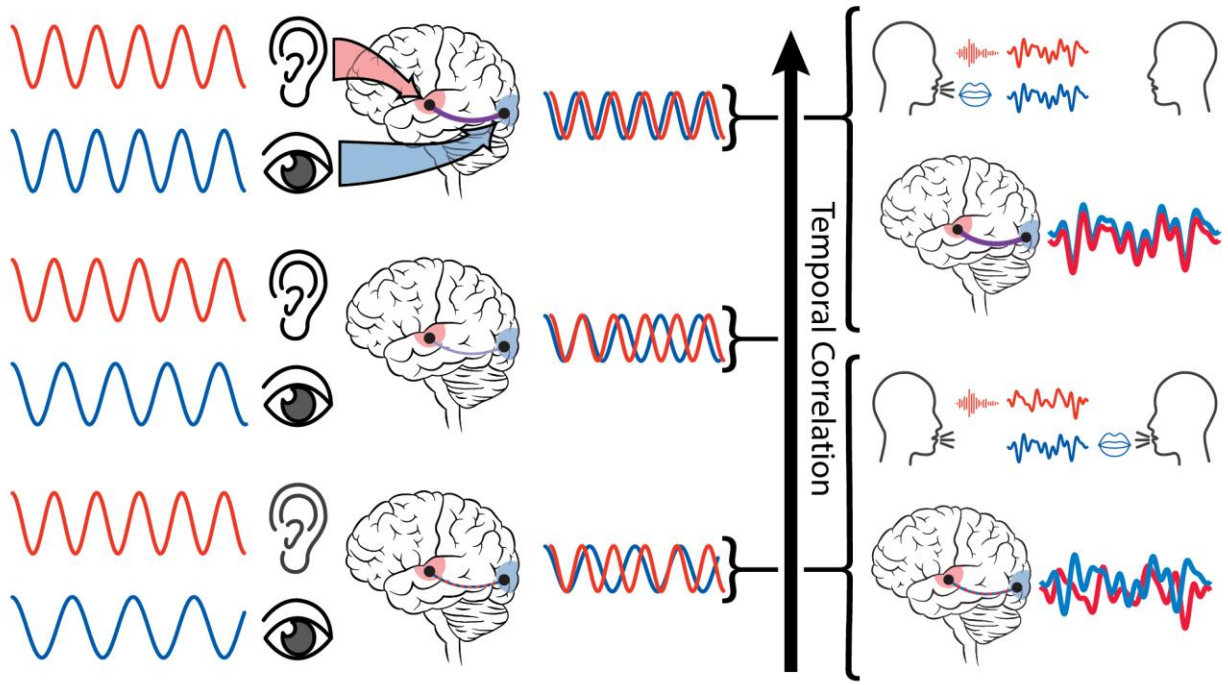


Figure 5.1 (previous page): Representation of stimulus correlation and neural synchrony. *Rhythmic auditory (red) and visual (blue) stimuli are known to entrain oscillations in their respective sensory cortices. When played together, auditory and visual cortex may become (de)synchronized based on the*

parameters of the stimulus rhythms (frequency and phase). Along the vertical axis we show a decrease in frequency of the visual stimulus which leads to a decrease in the frequency of the entrained oscillations in visual cortex. These changes desynchronize the neural activity between auditory and visual cortices which ultimately decouples the activity. This can be likened to audible speech signals and mouth movements generated by one speaker (high temporal correlation, high synchrony) or by two speakers (low temporal correlation and synchrony; adapted from Beker et al 2018).

The signals that were used in the work described here, sinusoidal AM, were chosen because of their known ability to entrain ongoing neural oscillations (Henry et al., 2014; Henry & Obleser, 2012; Schroeder et al., 2010; Thut et al., 2011). Rhythmic uni- and multisensory stimuli are able to entrain oscillations in multiple frequency bands at once (Henry et al., 2014; Nozaradan et al., 2012). Moreover, behavioral sensitivity is modulated by both oscillatory patterns (Henry et al., 2014). We leveraged this ability with the assumption that modifying the parameters of the AM in one stimulus would change the oscillations that follow that stimulus. Thus, manipulating the magnitude of correlation of the stimuli should change the degree of synchrony of the underlying oscillatory activity and their functional connection (Figure 5.1), as described in the examples above. In turn, this connection should drive binding. The work presented in Chapter 3 confirms parts of this hypothesis.

The work presented here shows that behavioral performance in two orthogonal tasks depends on a linear relationship between two multisensory AM signals (viz., correlation). Because of the nature of the stimuli used here, behavioral performance likely also depends on the strength of neural synchrony that is established by correlated oscillatory activity. First, when low-frequency fluctuations in auditory and visual signals are congruent, both auditory and visual neural activity that follows a higher-frequency feature is enhanced across trials (Nozaradan et al., 2012). Second, in a similar experiment that used non-rhythmic stimuli, the ability of neurons in auditory cortex to follow an auditory stream and encode events in that stream are enhanced by a congruent visual stream (Atilgan et al., 2018). Whether these or other neural processes scale with the strength of the correlation is a large focus for future research endeavors (see below).

Flexible binding

Audiovisual stimuli, such as speech, that come from a common event are typically well correlated with each other (Chandrasekaran et al., 2009). However, the sensory world is very complex. Even though the correlations between stimuli that go together are typically strong, spurious correlations do exist between unrelated, randomly-paired auditory and visual speech signals. In a situation where multiple individuals are speaking, a mechanism that simply indexes

whether signals are correlated or uncorrelated could lead to obligatory and errant binding. Flexible binding of the appropriate signals would require a mechanism based on the strength of the correlation. So far, most studies have tested the effect of only two levels of correlation on behavioral and neural processes (Atilgan et al., 2018; Denison et al., 2013; Maddox et al., 2015; Nozaradan et al., 2012; Parise et al., 2012, c.f. 2013), making inference about the measurement of correlation impossible. However, a recent report of a general multisensory correlation detector model (Parise & Ernst, 2016), utilized the detection of correlation along a continuum to predict a number of multisensory behaviors.

By manipulating the frequency and phase relationships between the two AM stimuli, we were able to generate stimuli with a range of temporal correlations over which to test binding and integration. The first main finding, that multisensory integration depends on the strength of correlation, is evidence that the brain can represent of the degree of similarity between two streams. Furthermore, the second finding is that this incremental process appears to influence multisensory object formation (Bizley et al., 2016). Thus, we are increasingly more likely to form multisensory objects as the underlying correlation between sensory streams increases. This allows for appropriate inference of multisensory associations in complex sensory environments.

The incremental dependence on correlation could also be involved in a multisensory process similar to hierarchical image segmentation (Ullman, 2007; Ullman & Sali, 2000; Ullman, Vidal-Naquet, & Sali, 2002). Image segmentation involves the representation of objects by groups of common image fragments. Visual features of increasing complexity are combined hierarchically and flexibly to promote object invariance. It was proposed that intermediately complex features, which convey more information about an object than simpler or more complex ones (Ullman et al., 2002), could be grouped by a hierarchy of similarity (Kubilius, Wagemans, & Op de Beeck, 2014).

In a version of this process, different levels of objects could potentially be bound by different degrees of correlation. Take for example a forest of trees. One could focus attention to the forest, to a cluster of trees, to a single tree, or to a leaf from a single tree. Within a single leaf, its uniform surface is likely to have a high degree of correlation across space and, if the wind is blowing, across time as all parts of the leaf will move together. A single tree contains a multitude of leaves. The individual leaves are more correlated with themselves than other leaves and so the correlations across the entire tree are diminished relative to a single leaf. This logic can be extended to account for different correlations among a cluster of trees or the entire forest. We are

left with four levels of objects with three levels of correlations. As a general principle, correlation decreases with increased object hierarchy. Focusing attention on each of these levels possibly draws from our ability to process different levels of correlation as demonstrated in the current work.

Modeling temporal correlation

One of the novel approaches to the current work is the use of a decisional model to explain the evidential nature of correlation. Evidence accumulation models, especially the class of diffusion model used here, can disentangle participant- and stimulus-related phenomena by fitting choice probabilities and RTs simultaneously (Voss et al., 2004). From these two measures, decision models can reliably estimate differences in participant bias, speed/accuracy trade-off, the time it takes to process stimuli, and the amount of sensory information in that stimuli relevant for the decision.

While analyzing the data from Chapter 2, we discovered an interesting feature of our stimuli that we discussed briefly in the Discussion section of that chapter. Just by changing the duration of the audiovisual stimulus we can change its overall temporal correlation. For example, when the auditory AM is 7 Hz/90° and the visual AM is 6 Hz/0°, at 500 ms the correlation

coefficient is -0.69 (Figure 5.2a). When that duration is extended out to 1 s, the correlation becomes ~0 (Figure 5.2b). Since observer will typically report the presences of AM after ~800 ms, it is impossible that the decisional system is utilizing correlation present in the full 1 second of duration. Considering this observation, we attempted an experiment aimed at dissecting how this correlation unfolding in the stimulus over time, in turn, unfolded in behavior. We hypothesized that we could measure the interval over which participants accumulated correlation by varying stimulus duration (and thus the magnitude of correlation) prior to a feature (an auditory gap) that participants were detecting.

These experiments ultimately did not pan out. However, a slight modification of our model might be able to measure the build-up of correlation over time. In our model, we constrained drift rate to be proportional to stimulus correlation. Specifically, we took the *average* correlation over the entire 500 ms stimulus duration. Although this assumes the participants are using the entire correlation, behavioral data was fit nicely by this assumption. For each condition, this resulted in a drift rate that was a constant term which was added to the random fluctuations (diffusion) in the evidence accumulation process. This constant drift term is a feature of many accumulator models (Ratcliff et al., 2016). However, a subset of accumulator models—such as

the Ornstein-Uhlenbeck, interactive race, and leaky-competing models—employ non-constant drift rates. In many of these models, drift rates simply decay over time but it is possible to constrain drift rate to any shape across a single trial (Diederich & Oswald, 2014).

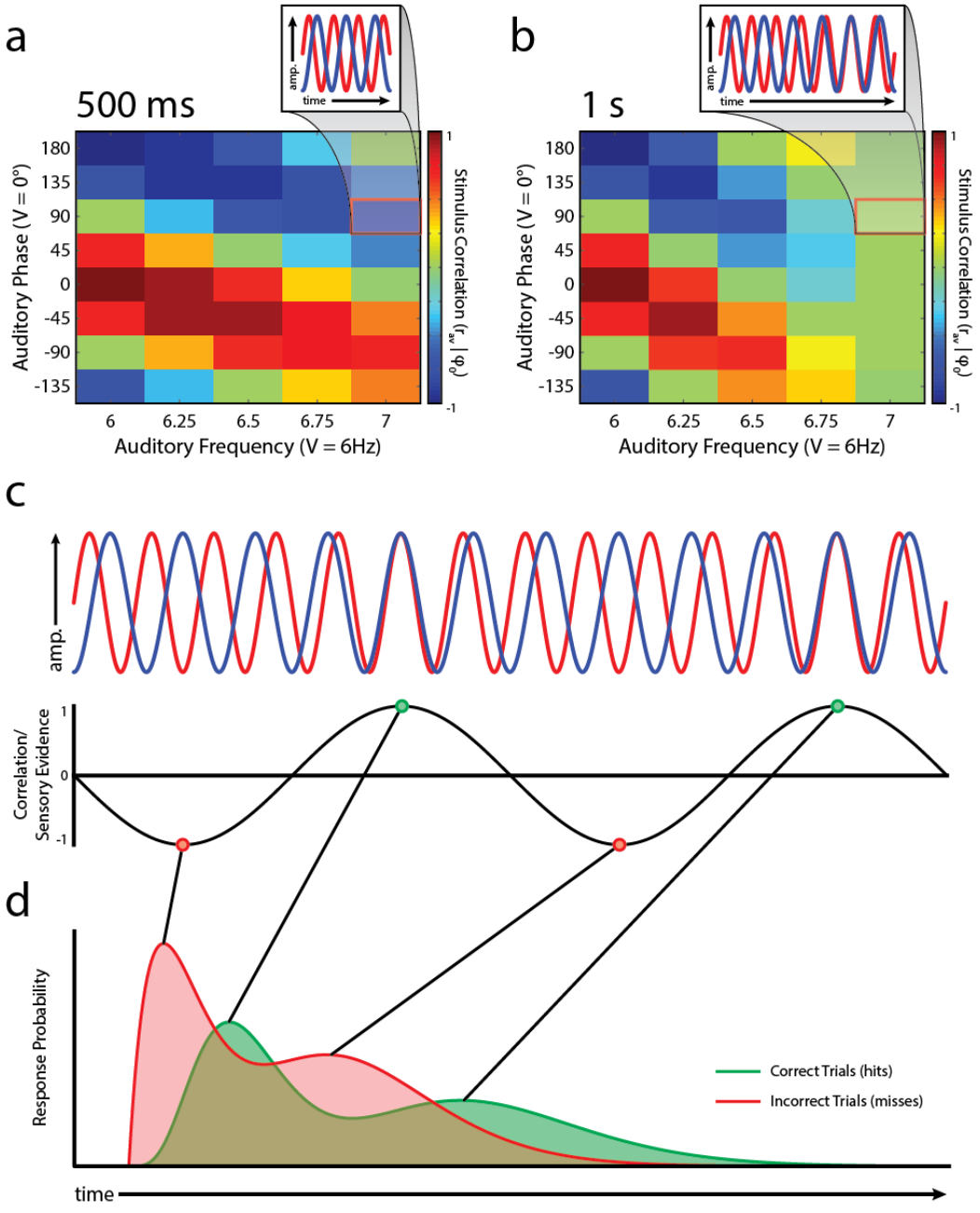


Figure 5.2 (previous page): Representation of stimulus correlation across durations. ***(a)*** *Stimulus correlation matrix for 500 ms stimulus. The condition in which auditory AM parameters are 7Hz/90° is highlighted and the auditory (red) and visual (blue) AM envelopes are shown above the matrix.* ***(b)*** *The same information is shown as in (a) except for the stimulus duration is now 1 s. By simply changing duration, the general pattern of the matrix, and the correlation of the example condition have changed.* ***(c)*** *A longer duration stimulus ~2 s is shown for AM envelopes (red and blue) having the same parameters as in (a) and (b). Below, the fluctuations of correlation are depicted. High correlations align with in-phase envelopes while low correlations occur when envelopes are out of phase.* ***(d)*** *Probability density functions (PDF) of correct (green) and incorrect (red) responses. Note the bi-modality of each PDF and the out-of-phase nature of the modes. For this condition, there are more incorrect trials due to the stimulus initially having a negative correlation. In the "opposite" condition in which the auditory has a starting phase of -90°, we would expect the correlation to begin positive and the two RT distributions to reverse (red becomes green and vice-versa). The time axis has been compressed relative to (c).*

A more complex model could be constructed where temporal fluctuations in the correlation (Figure 5.2c) are used to constrain drift rate. Given that correlations fluctuate between positive and negative which is a product of the auditory and visual AM moving in and out of phase, we might expect RT distributions to be multimodal with modes aligning to the waveform of the correlation fluctuations (Figure 5.2d). Because positive correlation should induce correct responses (hits) and negative correlation should lead to errors (misses), we would predict that the multiple modes be out of phase across correct and error RT distributions.

Because of these intricacies, we would not be able to employ model fitting on mean RTs as we did in Chapter 2 and would instead need to fit the entire distribution, and thus the procedure would require many correct *and* error trials, which can take as many as 200 trials to get a satisfactory estimate of both distributions (Wagenmakers, 2009).

Future Experiments

A practical extension of the utility of temporal correlation is in the processing of audiovisual speech. Visual access to the face improves speech intelligibility (Erber, 1969; Ross et al., 2007; Sumby & Pollack, 1954). Given the correlations embedded in audiovisual speech and their importance to speech processing and perception (Crosse et al., 2015; Grant & Seitz, 2000; Munhall et al., 1996; O'Sullivan & Lalor, 2017; Park et al., 2016; Venezia et al., 2016), the question of how correlation strength affects these processes naturally emerges. Does correlation bestow benefits on speech perception in the same incremental manner as described in this dissertation?

Another question provoked by the current findings involves the neural substrates of temporal correlation. What is the underlying computation of correlation and where (in terms of both physical location and along the processing hierarchy) can we find it? Does multisensory

correlation incrementally enhance activity in both respective sensory cortices? Is there another region of the brain that is tasked with measuring the strength of correlation? With regard to the ascent along the auditory processing axis (Marslen-Wilson & Warren, 1994), does correlation incrementally affect the representation of speech envelope (Crosse et al., 2015) or spectrogram (Pasley et al., 2012), their intermediate phonemic representation (Di Liberto, O'Sullivan, & Lalor, 2015), the transform into linguistic information (Brodbeck, Hong, & Simon, 2018), and their modulation by attention (Brodbeck et al., 2018; Mesgarani & Chang, 2012; O'Sullivan & Lalor, 2017)?

The questions and experiments posed here surround the temporal structure of the stimuli and its relevance to human speech. Electroencephalography (EEG) or magnetoencephalography (MEG) represent excellent tools given their temporal precision and non-invasive nature. Moreover, the questions posed involve questions about neural encoding and tracking of stimuli which are easily posed using time-series data from M/EEG (Crosse et al., 2015; Ding, Melloni, Zhang, Tian, & Poeppel, 2016; Haynes & Rees, 2006; Keitel, Gross, & Kayser, 2018). Although similar experiments have been performed while recording from neurons *in vivo* (Atilgan et al., 2018), the work proposed here will focus on M/EEG methods. One set of

proposed experiments would benefit from single neuron recordings, but those have been described in Chapter 4.

Neural underpinnings of correlation

In one experiment, participants would be presented with stimuli like those used in the current work, AM visual and auditory stimuli where one AM signal changes along frequency and/or phase. Time-frequency analysis would be used instead of more traditional event-related potential (ERP) analysis. Spectrograms of the M/EEG signal recorded during audiovisual conditions would be compared to predicted spectrograms based on a model combining unisensory auditory (A) and visual (V) spectrograms [e.g., A + V, max(A,V)]. Here, we would predict that the power difference at the frequency of AM (auditory, visual, or average frequency) would improve with correlation. Correlation can be regressed against these measures across electrodes and the regression coefficients can be plotted topographically. One might expect these power differences to vary with correlation over temporal and occipital electrodes, as these scalp locations have been reported to be active during the processing of similar stimuli (Crosse et al., 2015, 2016). Following an experiment that showed changes in synchronization during binding (Hipp et al., 2011), this experiment could also benefit from functional connectivity analysis

(Vinck, Oostenveld, van Wingerden, Battaglia, & Pennartz, 2011) to explore how correlation changes network dynamics throughout the brain and whether those changes correspond to those shown previously.

Another experiment might be aimed at testing how well one stimulus aids in the neural tracking of another stimulus, which describes how well a particular feature is represented in the brain. These features can be stimulus related (e.g., envelope, spectrogram, coherence; Crosse et al., 2015; O'Sullivan et al., 2015; Pasley et al., 2012). or even related to information extracted from the stimulus (e.g., semantic information; Ding et al., 2016). In these experiments, stimuli are typically long (>3 s) and usually require an arrhythmic temporal form (c.f., Ding et al., 2016) and so using sinusoidal AM stimuli is impractical. However, more complex envelopes can be constructed in the frequency domain by selecting desired frequencies with a uniform amplitude and assigning each a random phase and computing the inverse Fourier transform (Maddox et al., 2015). Visual stimuli with varying degrees of correlation can be constructed in the same way by specifying a correlation in the phase component (R. K. Maddox, personal communication, March 6, 2018). M/EEG recorded during the presentation of these stimuli can be used to define a temporal filter, called the temporal response function (TRF; Crosse, Di Liberto, Bednar, &

Lalor, 2016; Lalor & Foxe, 2010), which represents the mapping between some time-varying sensory feature(s) in a stimulus—in this case the amplitude envelope—and the resultant M/EEG response. The TRF can be used to reconstruct the time course of the auditory envelope signal by convolving it with the neural signal across different listening conditions via cross-validation (Figure 5.3a; Crosse et al., 2015; Crosse & Lalor, 2014). Frequency features could be added to the stimuli and spectrograms can be reconstructed from the underlying neural activity (Pasley et al., 2012). Reconstruction accuracy can be measured by correlating the stimulus envelope with the reconstruction of that envelope (or the actual and reconstructed spectrograms). Reconstruction accuracy can then be measured with the addition of temporally coherent visual stimuli (Figure 5.3a-b). Predictably, increasing the coherence should proportionally improve the neural tracking (Figure 5.3c, right side). A topography of channel contributions to the reconstruction (i.e., which channels carry the most information) can be used to localize where the information used in the reconstruction comes from.

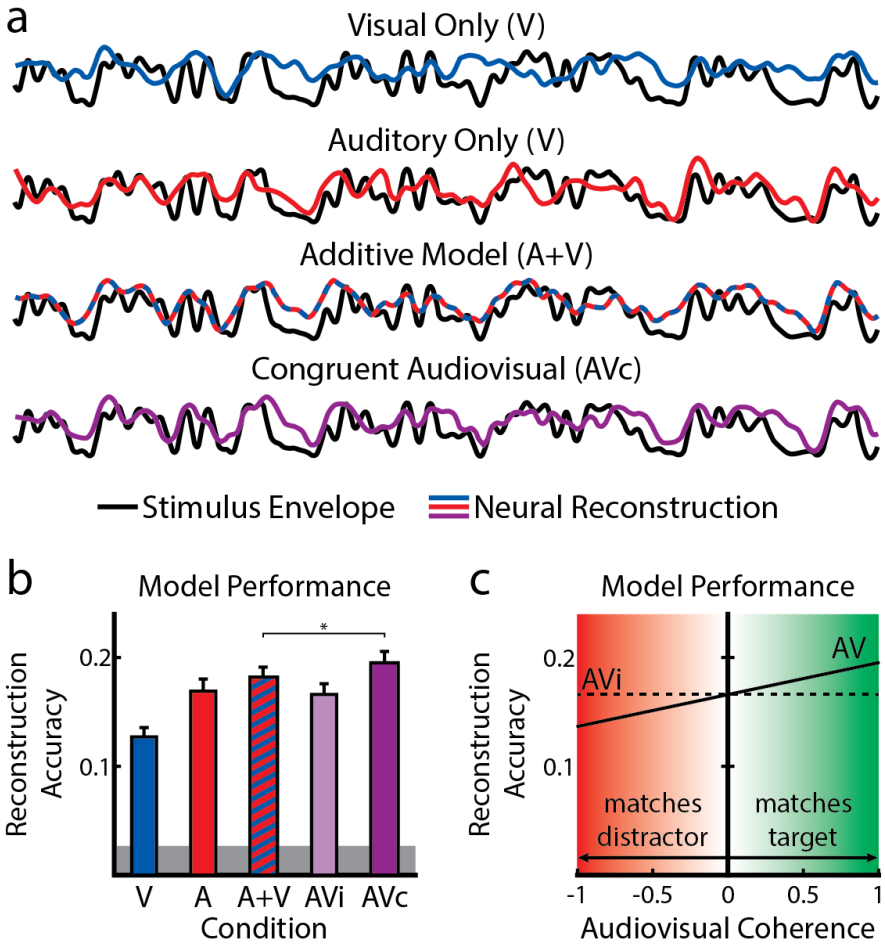


Figure 5.3: Stimulus reconstruction with M/EEG (previous page): A. *Auditory stimulus envelopes (black) are reconstructed from neural activity during auditory (red), visual (blue), and audiovisual (purple) presentations. The accuracy is measured as the Pearson's correlation between stimulus envelope and the reconstruction.* ***B.*** *Reconstruction accuracy is improved by the presentation of a congruent audiovisual stimulus (adapted from Crosse et al., 2015).* ***C.*** *Hypothesized linear effect of correlation on reconstruction accuracy on a selective attention experiment. When visual coherence matches a distractor, reconstruction accuracy suffers.*

In a version of the experiment, participants can be asked to detect a target feature imbedded in the stimuli. Presumably, the sensitivity of detecting the target should vary with the

ability of the neural activity to encode the stimulus and the target. This could be measured by correlating behavior with neural tracking performance (e.g., reconstruction accuracy) or with the performance of a classifier that chooses whether the feature was present or not. The assumption here is that higher reconstruction accuracy induced by stronger correlations should enhance the representation of the target feature. We can also test the effects on attention by simultaneously presenting a distractor stimulus while varying the visual stream from fully coherent with the target to incoherent and further to fully congruent with the distractor. The coherence of the visual stream and which auditory stream it is coherent with should affect both behavior and the ability to reconstruct the stimulus envelope proportionally (Figure 5.3c).

Oscillatory phase shift

One way to test whether phase shift is related to neural oscillations is to measure the neural activity in an experiment similar to the first described above. Then the benefit with different frequency and phase conditions could be fit to stimulus correlation computed with a phase lag in a manner similar to what was described in Chapter 2. Neural phase shift can be compared to behavioral phase shift across participants. A necessary experiment that might highlight any mechanistic interactions would involve taking separate and direct measures of

auditory and visual phase lag (Henry & Obleser, 2012). A prediction of phase shift can then be computed by taking the difference between auditory and visual phase lags. Any systematic difference between the predicted and fit phase shift might indicate an interaction that normalizes the stimulus timing in the brain.

Speech perception

Previous reports relating the effect of similarity on speech perception have done so indirectly. These studies have reported different conditions in which the auditory and visual speech were unmatched in gender (Vatakis & Spence, 2007), speaking speed (and thus duration; Munhall et al., 1996), and direction (forward and reverse; Kim & Davis, 2004). To more directly evaluate the contribution of correlation to our perception of speech, one behavioral experiment would measure recognition of speech in noise with and without the addition of a visible face. Performance would be titrated to a noise level where multisensory benefit to matched stimuli is maximal (Ross et al., 2007) in order to maximized the potential effect of changing the visual correlation. Mouth movements can be artificially modulated by an envelope that is correlated with the acoustic envelope along a continuum of strengths (as described above). Participants can be asked to identify spoken words. Audiovisual speech recognition can be compared across

different correlations and compared to auditory only performance. Further, performance on trials with highly-correlated artificial speech can be compared to matched audiovisual speech as a measure of the contribution of visemes and visual semantic information (P. L. Jackson, 1988).

An extension of this would be aimed at resolving the neural substrates involved in the contribution of correlated audiovisual speech to speech perception. In this task participants would listen to continuous speech signals with similar artificially modulated mouth movements. In the experiments participants would be tasked with reporting whether a target word appeared in the speech stimulus across different correlations. The underlying neural effect of correlation could be observed in a measure of reconstruction accuracy (see above; Crosse et al., 2015; Crosse & Lalor, 2014) or the accuracy of a classifier trying to classify correct and incorrect responses. Hit rate would likely be correlated with reconstruction/classifier accuracy across correlation levels and across subjects.

One of the driving predictions related to a linear representation of correlation was that it allowed for the selection of an appropriate signals to bind when the potential signals are correlated to some degree. According to the prediction, the selection should be based on the strongest correlation among audiovisual pairs. We can probe this hypothesis by introducing

competition. In one task, two differentiable speech streams (e.g., one to the left ear and one to the right ear, one begins earlier than the other) are presented in noise. A target word is displayed prior to speech onset and is present in one of the two streams. Participants are asked to making a judgement about the stream containing the target word. A visual stimulus involving a face with an artificially modulated mouth movements can be presented that is correlated to some degree with both speech envelopes. It would, of course, be more correlated with one acoustic cue. According to the prediction, participant performance should increase with difference in correlation between the visual cue and the two auditory cues.

One final experiment is aimed at deconstructing how correlation affects processing along the auditory hierarchy. For these experiments, M/EEG is recorded from participants as they watch videos of natural, continuous speech. We can use TRFs to map the temporal filter for a set of audiovisual features, across a variety of listening conditions. For example, forward models that involve increasingly complex features such as amplitude envelopes, acoustic spectra, phonemes and visemes, or even the semantic information in the speech can be generated. We can deconvolve the neural signal with correlations embedded in these features. These TRFs should give us a measure of the contribution of each feature to the neural signal. Manipulation of the

listening condition (low vs. high background noise, one speaker vs. many speakers) might provide conditions in which these features contribute differently to the processing of speech. One might predict that in high levels of noise, where the physical features of the acoustic speech (i.e., the amplitude envelope or spectrogram) are degraded, integration of correlation might occur at the phonemic/visemic level or higher. A more causal role might be established by manipulating these features just prior to a target word. By interrupting the brain's tracking of these features and their relation to word recognition, we can ascribe a causal role for each in the intelligibility of speech.

Conclusions

Previous work has characterized the role of temporal correlation in a variety of processes such as temporal processing, spatial processing, selective attention, and speech perception. These studies have forfeited a fine-grain view of correlation in favor of more natural and realistic tasks and stimuli. In this dissertation, we have simplified the task and the stimuli in favor of a deeper look into temporal correlation and its role in multisensory processes of integration and binding. We have parametrically varied correlation and found that the strength of correlation makes a contribution to both of these processes, but with different magnitudes. This granularity of

processing temporal correlation in the brain is requisite to a host of processes such as selection of stronger correlation, which may be necessary for appropriate binding in complex audiovisual environments, and hierarchical binding. The answers to the questions posed here lay a solid foundation for future work on the investigation of temporal correlation, its instantiation in the brain, and its effects on different processes involving multisensory integration and binding, and the neural underpinnings of those processes.

References

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proceedings of the National Academy of Sciences of the United States of America, 98(23), 13367–13372. https://doi.org/10.1073/pnas.201400998

Atilgan, H., Town, S. M., Wood, K. C., Jones, G. P., Maddox, R. K., Lee, A. K. C., & Bizley, J. K. (2018). Integration of Visual Information in Auditory Cortex Promotes Auditory Scene Analysis through Multisensory Binding. Neuron, 97(3), 640–655.e4. https://doi.org/10.1016/j.neuron.2017.12.034

Bishop, G. H. (1933). Cyclic changes in excitability of the optic pathway of the rabbit. American Journal of Physiology--Legacy Content, 103(1), 213–224.

Bizley, J. K., Maddox, R. K., & Lee, A. K. C. (2016). Defining Auditory-Visual Objects: Behavioral Tests and Physiological Mechanisms. Trends in Neurosciences. https://doi.org/10.1016/j.tins.2015.12.007

Bolognini, N., Frassinetti, F., Serino, A., & Làdavas, E. (2005). "Acoustical vision" of below threshold stimuli: Interaction among spatially converging audiovisual inputs. Experimental Brain Research, 160(3), 273–282. https://doi.org/10.1007/s00221-004-2005-z

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Transformation from auditory to linguistic representations across auditory cortex is rapid and attention dependent for continuous speech. BioRxiv, 326785. https://doi.org/10.1101/326785

Busch, N. A., Dubois, J., & VanRullen, R. (2009). The Phase of Ongoing EEG Oscillations Predicts Visual Perception. Journal of Neuroscience, 29(24), 7869–7876. https://doi.org/10.1523/JNEUROSCI.0113-09.2009

Buzsaki, G., & Draguhn, A. (2004). Neuronal Oscillations in Cortical Networks. Science, 304(5679), 1926–1929. https://doi.org/10.1126/science.1099745

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A., & Ghazanfar, A. A. (2009). The natural statistics of audiovisual speech. PLoS Computational Biology, 5(7). https://doi.org/10.1371/journal.pcbi.1000436

Chuen, L., & Schutz, M. (2016). The unity assumption facilitates cross-modal binding of musical, non-speech stimuli: The role of spectral and amplitude envelope cues. Attention, Perception, and Psychophysics, 78(5), 1512–1528. https://doi.org/10.3758/s13414-016-1088-5

Cravo, A. M., Rohenkohl, G., Wyart, V., & Nobre, A. C. (2013). Temporal expectation enhances contrast sensitivity by phase entrainment of low-frequency oscillations in visual cortex. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 33(9), 4002–4010. https://doi.org/10.1523/JNEUROSCI.4675-12.2013

Crosse, M. J., Butler, J. S., & Lalor, E. C. (2015). Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. Journal of Neuroscience, 35(42), 14195–14204. https://doi.org/10.1523/JNEUROSCI.1829-15.2015

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. Frontiers in Human Neuroscience, 10, 604. https://doi.org/10.3389/fnhum.2016.00604

Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2016). Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration. Journal of Neuroscience, 36(38), 9888–9895. https://doi.org/10.1523/JNEUROSCI.1396-16.2016

Crosse, M. J., & Lalor, E. C. (2014). The cortical representation of the speech envelope is earlier for audiovisual speech than audio speech. Journal of Neurophysiology, 111(7), 1400–1408. https://doi.org/10.1152/jn.00690.2013

Darwin, C. J. (1992). Listening to two Things at Once. In M.E.H. Schouten (Ed.), The Auditory Processing of Speech (pp. 133–147). Berlin: Mouton de Gruyter.

Diederich, A., & Oswald, P. (2014). Sequential sampling model for multiattribute choice alternatives with random attention time and processing order. Frontiers in Human Neuroscience, 8, 697. https://doi.org/10.3389/fnhum.2014.00697

Denison, R. N., Driver, J., & Ruff, C. C. (2013). Temporal Structure and Complexity Affect Audio-Visual Correspondence Detection. Frontiers in Psychology, 3, 619. https://doi.org/10.3389/fpsyg.2012.00619

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. Annual Review of Neuroscience, 18(1), 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205

Di Liberto, G. M., O'Sullivan, J. A., & Lalor, E. C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. Current Biology, 25(19), 2457–2465. https://doi.org/10.1016/J.CUB.2015.08.030

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. Nature Neuroscience, 19(1), 158–164. https://doi.org/10.1038/nn.4186

Engel, A. K., Fries, P., & Singer, W. (2001). Dynamic predictions: Oscillations and synchrony in top–down processing. Nature Reviews Neuroscience, 2(10), 704–716. https://doi.org/10.1038/35094565

Engel, A. K., Senkowski, D., & Schneider, T. R. (2007). Multisensory Integration through Neural Coherence. In M. M. Murray & Wallace Mark T (Eds.), The Neural Bases of Multisensory Processing (pp. 115–130). Boca Raton: CRC Press. https://doi.org/NBK92855 [bookaccession]

Engel, A. K., & Singer, W. (2001). Temporal binding and the neural correlates of sensory awareness. Trends in Cognitive Sciences. https://doi.org/10.1016/S1364-6613(00)01568-0

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. Journal of Speech Language and Hearing Research, 12(2), 423. https://doi.org/10.1044/jshr.1202.423

Frassinetti, F., Bolognini, N., & Làdavas, E. (2002). Enhancement of visual perception by crossmodal visuo-auditory interaction. Experimental Brain Research, 147(3), 332–343. https://doi.org/10.1007/s00221-002-1262-y

Fries, P. (2005). A mechanism for cognitive dynamics: Neuronal communication through neuronal coherence. Trends in Cognitive Sciences. https://doi.org/10.1016/j.tics.2005.08.011

Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. Nature Neuroscience, 15(4), 511–517. https://doi.org/10.1038/nn.3063

Grant, K. W., & Seitz, P. F. 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 Pt 1), 1197–1208. https://doi.org/10.1121/1.422512

Gray, C. M. (1999). The temporal correlation hypothesis of visual feature integration: still alive and well. Neuron, 24(1), 31–47, 111–125. https://doi.org/10.1016/S0896-6273(00)80820-X

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. Nature Reviews Neuroscience, 7(7), 523–534. https://doi.org/10.1038/nrn1931

Henry, M. J., Herrmann, B., & Obleser, J. (2014). Entrained neural oscillations in multiple frequency bands comodulate behavior. Proceedings of the National Academy of Sciences of the United States of America, 111(41), 1408741111-. https://doi.org/10.1073/pnas.1408741111

Henry, M. J., & Obleser, J. (2012). Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. Proceedings of the National Academy of Sciences, 109(49), 20095–20100. https://doi.org/10.1073/pnas.1213390109

Henry, M. J., & Obleser, J. (2013). Dissociable neural response signatures for slow amplitude and frequency modulation in human auditory cortex. PloS One, 8(10), e78758. https://doi.org/10.1371/journal.pone.0078758

Hipp, J. F., Engel, A. K., & Siegel, M. (2011). Oscillatory synchronization in large-scale cortical networks predicts perception. Neuron, 69(2), 387–396. https://doi.org/10.1016/j.neuron.2010.12.027

Jackson, P. L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. The Volta Review, 90(5), 99–115.

Keitel, A., Gross, J., & Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. PLOS Biology, 16(3), e2004473. https://doi.org/10.1371/journal.pbio.2004473

Kim, J., & Davis, C. (2004). Investigating the audio–visual speech detection advantage. Speech Communication, 44(1–4), 19–30. https://doi.org/10.1016/J.SPECOM.2004.09.008

Körding, K. P., Beierholm, U., Ma, W. J., Quartz, S., Tenenbaum, J. B., & Shams, L. (2007). Causal Inference in Multisensory Perception. PLoS ONE, 2(9), e943. https://doi.org/10.1371/journal.pone.0000943

Kubilius, J., Wagemans, J., & Op de Beeck, H. P. (2014). A conceptual framework of computations in mid-level vision. Frontiers in Computational Neuroscience, 8, 158. https://doi.org/10.3389/fncom.2014.00158

Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of Neuronal Oscillations as a Mechanism of Attentional Selection. Science, 320, 110–113.

Lakatos, P., Shah, A. S., Knuth, K. H., Ulbert, I., Karmos, G., & Schroeder, C. E. (2005). An Oscillatory Hierarchy Controlling Neuronal Excitability and Stimulus Processing in the Auditory Cortex An Oscillatory Hierarchy Controlling Neuronal Excitability and Stimulus Processing in the Auditory Cortex. Journal of Neurophysiology, 94(3), 1904–1911. https://doi.org/10.1152/jn.00263.2005

Lalor, E. C., & Foxe, J. J. (2010). Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. European Journal of Neuroscience, 31(1), 189–193. https://doi.org/10.1111/j.1460-9568.2009.07055.x

Luo, H., & Poeppel, D. (2007). Phase Patterns of Neuronal Responses Reliably Discriminate Speech in Human Auditory Cortex. Neuron, 54(6), 1001–1010. https://doi.org/10.1016/J.NEURON.2007.06.004

Maddox, R. K., Atilgan, H., Bizley, J. K., & Lee, A. K. (2015). Auditory selective attention is enhanced by a task-irrelevant temporally coherent visual stimulus in human listeners. ELife, 2015(4), 1–11. https://doi.org/10.7554/eLife.04995.001

Magnotti, J. F., Ma, W. J., & Beauchamp, M. S. (2013). Causal inference of asynchronous audiovisual speech. Frontiers in Psychology, 4, 798. https://doi.org/10.3389/fpsyg.2013.00798

Marslen-Wilson, W., & Warren, P. (1994). Levels of perceptual representation and process in lexical access: words, phonemes, and features. Psychological Review, 101(4), 653–675.

Mathewson, K. E., Gratton, G., Fabiani, M., Beck, D. M., & Ro, T. (2009). To See or Not to See: Prestimulus Alpha Phase Predicts Visual Awareness. Journal of Neuroscience, 29(9), 2725–2732. https://doi.org/10.1523/JNEUROSCI.3963-08.2009

Meredith, M. A., & Stein, B. E. (1983). Interactions among converging sensory inputs in the superior colliculus. Science, 221(4608), 389–391. https://doi.org/10.1126/science.6867718

Mesgarani, N., & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. Nature, 485(7397), 233–236. https://doi.org/10.1038/nature11020

Metzger, W. (1934). Beobachtungen über phänomenale Identität. Psychologische Forschung, 19(1), 1–60. https://doi.org/10.1007/BF02409733

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. Perception & Psychophysics, 58(3), 351–362. https://doi.org/10.3758/BF03206811

Neuling, T., Rach, S., Wagner, S., Wolters, C. H., & Herrmann, C. S. (2012). Good vibrations: Oscillatory phase shapes perception. NeuroImage, 63(2), 771–778. https://doi.org/10.1016/j.neuroimage.2012.07.024

Ng, B. S. W., Schroeder, T., & Kayser, C. (2012). A precluding but not ensuring role of entrained low-frequency oscillations for auditory perception. The Journal of Neuroscience : The Official Journal of the Society for Neuroscience, 32(35), 12268–12276. https://doi.org/10.1523/JNEUROSCI.1877-12.2012

Nozaradan, S., Peretz, I., & Mouraux, A. (2012). Steady-state evoked potentials as an index of multisensory temporal binding. NeuroImage, 60(1), 21–28. https://doi.org/10.1016/j.neuroimage.2011.11.065

O'Sullivan, A. E., Crosse, M. J., Di Liberto, G. M., & Lalor, E. C. (2017). Visual Cortical Entrainment to Motion and Categorical Speech Features during Silent Lipreading. Frontiers in Human Neuroscience, 10, 679. https://doi.org/10.3389/fnhum.2016.00679

O'Sullivan, A. E., & Lalor, E. C. (2017). Improved attentional decoding at a cocktail party for audiovisual speech. In Society for Neuroscience. Washington, DC.

O'Sullivan, J. A., Shamma, S. A., & Lalor, E. C. (2015). Evidence for Neural Computations of Temporal Coherence in an Auditory Scene and Their Enhancement during Active Listening. Journal of Neuroscience, 35(18), 7256–7263. https://doi.org/10.1523/JNEUROSCI.4973-14.2015

Okada, Y. F. H. andYamaguchi T. (1994). Multimodal responses of the nonspiking giant interneurons in the brain of the crayfish Procambarus clarkii. J Comp Physiol [A], 174(4), 411–419. https://doi.org/10.1007/BF00191707

Otto, T. U., Dassy, B., & Mamassian, P. (2013). Principles of Multisensory Behavior. Journal of Neuroscience, 33(17), 7463–7474. https://doi.org/10.1523/JNEUROSCI.4678-12.2013

Parise, C. V., & Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. Nature Communications, 7(12), 364. https://doi.org/10.1038/ncomms11543

Parise, C. V., Spence, C., & Ernst, M. O. (2012). When correlation implies causation in multisensory integration. Current Biology, 22(1), 46–49. https://doi.org/10.1016/j.cub.2011.11.039

Parise, C. V, Harrar, V., Ernst, M. O., & Spence, C. (2013). Cross-correlation between Auditory and Visual Signals Promotes Multisensory Integration. Multisensory Research, 26, 1–10. https://doi.org/10.1163/22134808-00002417

Park, H., Kayser, C., Thut, G., & Gross, J. (2016). Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. ELife, 5, e14521. https://doi.org/10.7554/eLife.14521

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., … Chang, E. F. (2012). Reconstructing Speech from Human Auditory Cortex. PLoS Biology, 10(1), e1001251. https://doi.org/10.1371/journal.pbio.1001251

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. Cerebral Cortex, 23(6), 1378–1387. https://doi.org/10.1093/cercor/bhs118

Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review, 85(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

Ratcliff, R., Smith, P. L., Brown, S. D., & McKoon, G. Diffusion Decision Model: Current Issues and History, 20 Trends in Cognitive Sciences § (2016). https://doi.org/10.1016/j.tics.2016.01.007

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. Cerebral Cortex, 17(5), 1147–1153. https://doi.org/10.1093/cercor/bhl024

Schroeder, C. E., & Lakatos, P. (2008). Low-frequency neuronal oscillations as instruments of sensory selection. Trends in Neurosciences, 32(1), 9–18. https://doi.org/10.1016/j.tins.2008.09.012

Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. Trends in Cognitive Sciences, 12(3), 106–113. https://doi.org/10.1016/j.tics.2008.01.002

Schroeder, C. E., Wilson, D. A., Radman, T., Scharfman, H., & Lakatos, P. (2010). Dynamics of Active Sensing and perceptual selection. Current Opinion in Neurobiology, 20(2), 172–176. https://doi.org/10.1016/j.conb.2010.02.010

Schutz, M., & Kubovy, M. (2009). Causality and cross-modal integration. Journal of Experimental Psychology: Human Perception and Performance, 35(6), 1791–1810. https://doi.org/10.1037/a0016455

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. Nature, 385(6614), 308. https://doi.org/10.1038/385308a0

Senkowski, D., Schneider, T. R., Foxe, J. J., & Engel, A. K. (2008). Crossmodal binding through neural coherence: implications for multisensory processing. Trends in Neurosciences. https://doi.org/10.1016/j.tins.2008.05.002

Shams, L., & Beierholm, U. R. (2010). Causal inference in perception. Trends in Cognitive Sciences, 14(9), 425–432. https://doi.org/10.1016/j.tics.2010.07.001

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. Trends in Cognitive Sciences, 12(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Singer, W., & Gray, C. (1995). Visual feature integration and the temporal correlation hypothesis. Annual Review of Neuroscience, 18, 555–586. https://doi.org/10.1146/annurev.ne.18.030195.003011

Stein, B. E., Burr, D., Constantinidis, C., Laurienti, P. J., Alex Meredith, M., Perrault, T. J., … Lewkowicz, D. J. (2010). Semantic confusion regarding the development of multisensory integration: a practical solution. European Journal of Neuroscience, 31(10), 1713–1720. https://doi.org/10.1111/j.1460-9568.2010.07206.x

Stein, B. E., & Stanford, T. R. (2008). Multisensory integration: Current issues from the perspective of the single neuron. Nature Reviews Neuroscience, 9(4), 255–266. https://doi.org/10.1038/nrn2331

Stein, B. E., Stanford, T. R., & Rowland, B. A. (2009). The neural basis of multisensory integration in the midbrain: Its organization and maturation. Hearing Research, 258(1–2), 4–15. https://doi.org/10.1016/J.HEARES.2009.03.012

Sumby, W. H., & Pollack, I. (1954). Visual Contribution to Speech Intelligibility in Noise. The Journal of the Acoustical Society of America, 26(2), 212–215. https://doi.org/10.1121/1.1907309

Thut, G., Schyns, P. G., & Gross, J. (2011). Entrainment of perceptually relevant brain oscillations by non-invasive rhythmic stimulation of the human brain. Frontiers in Psychology. https://doi.org/10.3389/fpsyg.2011.00170

Ullman, S. (2007). Object recognition and segmentation by a fragment-based hierarchy. Trends in Cognitive Sciences, 11(2), 58–64. https://doi.org/10.1016/J.TICS.2006.11.009

Ullman, S., & Sali, E. (2000). Object Classification Using a Fragment-Based Representation. In S. W. Lee, H. H. Bulthoff, & P. T (Eds.), Biologically Motivated Computer Vision (1811th ed., pp. 73–87). Berlin, Heidelberg: Springe. https://doi.org/10.1007/3-540-45482-9_8

Ullman, S., Vidal-Naquet, M., & Sali, E. (2002). Visual features of intermediate complexity and their use in classification. Nature Neuroscience, 5(7), 682–687. https://doi.org/10.1038/nn870

Vatakis, A., & Spence, C. (2007). Crossmodal binding: evaluating the "unity assumption" using audiovisual speech stimuli. Perception & Psychophysics, 69(5), 744–756. https://doi.org/10.3758/BF03193776

Venezia, J. H., Thurman, S. M., Matchin, W., George, S. E., & Hickok, G. (2016). Timing in audiovisual speech perception: A mini review and new psychophysical data. Attention, Perception, & Psychophysics, 78(2), 583–601. https://doi.org/10.3758/s13414-015-1026-y

Vinck, M., Oostenveld, R., van Wingerden, M., Battaglia, F., & Pennartz, C. M. A. (2011). An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. NeuroImage, 55(4), 1548–1565. https://doi.org/10.1016/J.NEUROIMAGE.2011.01.055

Voss, A., Rothermund, K., & Voss, J. (2004). Interpreting the parameters of the diffusion model: an empirical validation. Memory & Cognition, 32(7), 1206–1220. https://doi.org/10.3758/BF03196893

Wagenmakers, E.-J. (2009). Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. European Journal of Cognitive Psychology, 21(5), 641–671. https://doi.org/10.1080/09541440802205067

Wallace, M. T., Roberson, G. E., Hairston, W. D., Stein, B. E., Vaughan, J. W., & Schirillo, J. A. (2004). Unifying multisensory signals across time and space. Experimental Brain Research, 158(2), 252–258. https://doi.org/10.1007/s00221-004-1899-9

Watanabe, K., & Shimojo, S. (2001). When Sound Affects Vision: Effects of Auditory Grouping on Visual Motion Perception. Psychological Science, 12(2), 109–116. https://doi.org/10.1111/1467-9280.00319

Womelsdorf, T., & Fries, P. (2007). The role of neuronal synchronization in selective attention. Current Opinion in Neurobiology. https://doi.org/10.1016/j.conb.2007.02.002

Womelsdorf, T., Schoffelen, J.-M., Oostenveld, R., Singer, W., Desimone, R., Engel, A. K., & Fries, P. (2007). Modulation of Neuronal Interactions Through Neuronal Synchronization. Science, 316(5831), 1609–1612. https://doi.org/10.1126/science.1139597

Zion-Golumbic, E., & Schroeder, C. E. (2012). Attention modulates “speech-tracking” at a cocktail party. Trends in Cognitive Sciences. https://doi.org/10.1016/j.tics.2012.05.004

Zion Golumbic, E. M., Poeppel, D., & Schroeder, C. E. (2012). Temporal context in speech processing and attentional stream selection: A behavioral and neural perspective. Brain and Language, 122(3), 151–161. https://doi.org/10.1016/j.bandl.2011.12.010

Zoefel, B., Archer-Boyd, A., & Davis, M. H. (2018). Phase Entrainment of Brain Oscillations Causally Modulates Neural Responses to Intelligible Speech. Current Biology, 28(3), 401–408.e5. https://doi.org/10.1016/J.CUB.2017.11.071