1-1-2012

Effects Of Within-Modal Congruency, Cross-Modal Congruency, And Temporal Asynchrony On The Perception Of Percieved Audio-Visual Distance

Jonathan M.P. Wilbiks
Ryerson University

Follow this and additional works at: <http://digitalcommons.ryerson.ca/dissertations>

Part of the [Psychology Commons](#)

Recommended Citation

Wilbiks, Jonathan M.P., "Effects Of Within-Modal Congruency, Cross-Modal Congruency, And Temporal Asynchrony On The Perception Of Percieved Audio-Visual Distance" (2012). *Theses and dissertations*. Paper 1689.

This Thesis is brought to you for free and open access by Digital Commons @ Ryerson. It has been accepted for inclusion in Theses and dissertations by an authorized administrator of Digital Commons @ Ryerson. For more information, please contact bcameron@ryerson.ca.

EFFECTS OF WITHIN-MODAL CONGRUENCY, CROSS-MODAL
CONGRUENCY, AND TEMPORAL ASYNCHRONY ON THE PERCEPTION
OF PERCEIVED AUDIO-VISUAL DISTANCE

by

Jonathan Michael Paul Wilbiks
H.BSc University of Toronto, 2008
MA University of Sheffield, 2009

A thesis

presented to Ryerson University

in partial fulfillment of the

requirements for the degree of

Master of Arts

in the Program of

Psychology

Toronto, Ontario, Canada, 2012

© Jonathan Wilbiks, 2012

AUTHOR'S DECLARATION FOR ELECTRONIC SUBMISSION OF A THESIS

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

I understand that my thesis may be made electronically available to the public.

Congruency and temporal asynchrony in audio-visual perception

Effects of within-modal congruency, cross-modal congruency and temporal asynchrony on the perception of perceived audio-visual distance

Master of Arts 2012

Jonathan Michael Paul Wilbiks

Psychological Science

Ryerson University

ABSTRACT

When making decisions as to whether or not to bind auditory and visual information, temporal, spatial and congruency factors all contribute to the acceptance or rejection of multi-modal unity. While many of these factors have been studied in isolation, it is important to examine how they interact in a dynamic setting, in addition to evaluating ideas about the intrinsic relation between audition and the processing of time, and vision and the processing of space. Four experiments are presented, placing auditory and visual stimuli in a competitive binding scenario, to compare the effects of temporal and spatial factors both within and between modalities. Results support the dominance of auditory factors in temporal decision-making, and visual factors in spatial decision-making, with additional evidence for the presence of visual looming. With respect to audio-visual binding, the findings indicate precedence for temporal factors, with reliance on congruency factors only when the stimulus pairings are temporally ambiguous.

Acknowledgements

This thesis would not have been possible without the help of many individuals who contributed in many different ways.

I am very thankful to my supervisor, Dr. Ben Dyson, who has worked with me from the beginnings of thesis proposal through to the final stage of this write-up and always helped point me in the right direction when I started to stray off course.

I would also like to thank the members of the HEAR Lab, Raj Sandhu and Zara Chan, for their mostly informal feedback on my ideas in the planning and writing process, as well as for putting up with the incessant beeping of the experiments running in the lab.

Finally, I'd like to thank my family for always showing an interest in my work and asking me to explain it to them. I am most grateful to my wife, Hannah, who was quite literally the inspiration for this research, as she provided the initial question which I set out to answer through the thesis.

Table of Contents

Authors Declaration	ii
Abstract	iii
Acknowledgments.....	iv
Table of Contents	v
List of Tables	vii
List of Figures	viii
List of Appendices	ix
Introduction	
Multimodal Integration	1
Task Demands.....	2
Temporal Factors	4
Spatial Factors.....	6
Congruency Factors	9
Present Research	11
Experiment 1	
Method	14
Results	16
Discussion	18
Experiment 2	
Method	21
Results	21
Discussion	22

Experiment 3

Method	26
Results	28
Discussion	29

Interim Summary	30
-----------------------	----

Experiment 4

Method	32
Results	32
Discussion	34

General Discussion	36
--------------------------	----

Temporal Factors	37
Congruency Factors	39
Task Demands.....	40
Proposed Future Research.....	41
Concluding Remarks.....	43

References	45
------------------	----

Tables	52
--------------	----

Figures.....	63
--------------	----

Appendix.....	79
---------------	----

List of Tables

Table 1 – Probability of second responding (and SE) for 8 conditions of Experiment 1 (p. 51)

Table 2 – Summary of repeated measures ANOVA of Experiment 1 data (p. 52)

Table 3 – Probability of second responding (and SE) for 8 conditions of Experiment 2 (p. 53)

Table 4 – Summary of repeated measures ANOVA of Experiment 2 data (p. 54)

Table 5 – Reaction times (ms) (and SE) for each condition of Experiment 3 (p. 55)

Table 6 – Error rates (and SE) for each condition of Experiment 3 (p. 56)

Table 7 – Summary of repeated measures ANOVA of reaction time data of Experiment 3 (p. 57)

Table 8 – Summary of repeated measures ANOVA of error rate data of Experiment 3 (p. 58)

Table 9 – Probability of second responding (and SE) by condition for Experiment 4 (p. 59)

Table 10 – Summary of repeated measures ANOVA of Experiment 4 data (p. 60)

List of Figures

- Figure 1 – Experimental stimulus presentation timings for Experiments 1,2,4 (p. 62)
- Figure 2 – Comparison of time function for Experiments 1 and 2 (p. 63)
- Figure 3 – V1 x time interaction for Experiment 1 (p. 64)
- Figure 4 – V1 x V2 x time interaction for Experiment 1 (p. 65)
- Figure 5 – A1 x V interaction for Experiment 2 (p. 66)
- Figure 6 – A1 x time interaction for Experiment 2 (p. 67)
- Figure 7 – V x time interaction for Experiment 2 (p. 68)
- Figure 8 – Condition sets for Experiment 3 (p. 69)
- Figure 9 – Visual x auditory difficulty interaction for reaction times on Experiment 3 (p. 70)
- Figure 10 – Auditory difficulty x modality interaction for error rates on Experiment 3 (p. 71)
- Figure 11 – Congruency effect and AV difference for Experiment 3 (p. 72)
- Figure 12 – Condition x time interaction for Experiment 4 (p. 74)
- Figure 13 – S1 x time interaction for Experiment 4 (p. 75)
- Figure 14 – S3 x time interaction for Experiment 4 (p. 76)
- Figure 15 – Condition x S1 x S3 x time interaction for Experiment 4 (p. 77)

List of Appendices

Appendix A – Sample informed consent form used for experiments (p. 78)

Appendix B – Sample debriefing form used for experiments (p. 81)

Introduction

Multimodal Integration

In our lives, we are constantly exposed to stimuli from multiple sources and in many different modalities. These multimodal aspects of stimuli reach us through separate transducers (eyes, ears, nose, mouth, and skin) making for a plethora of individual sensory elements that enter our perceptual system. Making good decisions as to which pieces of information belong to the same source is an important contribution towards functioning properly in our world. The mechanisms involved in multimodal integration, specifically the binding of auditory and visual stimuli, have been shown however to both help and hinder our understanding of the external world. To take two basic examples from speech perception, Sumby and Pollack (1954) showed that the binding of lip movements of an individual to speech sounds facilitates the comprehension of speech in noisy environments. However, McGurk and MacDonald (1976) showed that illusory binding can lead to misperception of what is being spoken by a person, when these signals occur within a similar time window. When participants saw a mouth forming a /ba/ sound, but heard a /ga/ sound in close temporal proximity, they reported perceiving a /da/ sound. This error is a misattribution caused by incorrect binding of auditory and visual sensory inputs, occurring in part due to their close temporal coincidence.

Welch and Warren (1980) conceptualise the decision-making process associated with multimodal integration as a combination of cognitive and non-cognitive factors, through which a perceiver must make a decision as to whether the two (or more) sensory inputs they experience are different modality expressions of the same event, or two (or more) separate events.

Koelewijn, Bronkhorst and Theeuwes (2010) detail some of the factors that contribute to the perception of unity between multiple sensory inputs. These factors include the dominance of one modality over another determined by task demands (e.g. Stein & Stanford 2008), the spatial and

temporal coincidence of the composite signals (e.g., Calvert, Spence & Stein, 2004), and, whether the signals are congruent with one another (e.g., Molholm, Ritter, Javitt, & Foxe, 2004). This thesis will focus on the relationship between the temporal coincidence of auditory and visual stimuli and both within- and between-modality congruency via manipulations of perceived distance.

Task Demands

Welch and Warren (1980) proposed that when one is presented with an audiovisual stimulus and asked to make judgements about it, there is a disparity in the way judgements are made such that auditory or visual biases can arise as a function of task demands rather than any real constraint within audio-visual processing. Specifically, they state that when asked to make temporal judgements people rely primarily on auditory information (see Burr, Banks, & Morrone, 2009) but that they rely mostly on visual information when making spatial judgements (see Alais and Burr, 2004). This is attributed to a reliance on what has the highest fidelity in a given modality, which then gives the best opportunity of making correct assessments (Vroomen & Keetels, 2010). In other words, spatial acuity is best represented by the visual system but temporal acuity is best represented by the auditory system (Welch and Warren, 1980). Therefore, the potential explanation of auditory and visual influence may simply be a product of task demands and the relative weighting of one dimension over another. This asymmetry in the use of modality-specific information can result in failure to integrate multimodal signals, due to an underutilization of a modality based on the task on which it is being deployed. For example, when attempting to localize a the source of a signal in space, one may neglect auditory information in favour of higher-fidelity visual information, which could lead to faulty binding decisions. Roach et al. (2006) propose a multi-directional model for analysing audiovisual

stimuli, which they refer to as a maximum-likelihood estimation (MLE) system. They investigated interactions between auditory and visual stimuli using an attention task. The auditory stimuli were composed of a regularly occurring series of white noise bursts, while the visual stimuli were repeated flashes of an LED. Participants were asked to attend to one of the streams and indicate whether it had a higher rate of presentation than the other stream. While other studies have found distortions of perceived visual rate caused by concurrent auditory stimulation as a result of using a temporal task (e.g. *auditory driving*; Recanzone, 2003), Roach et al. (2006) found that the brain resolves discrepant information appropriately between modalities. Rather than the previously established biases for the two modalities, they found that the brain weighs information equally in determining which modality is more reliable in the given situation. If it seems that visual information is occurring on a regular time interval, it will be relied upon to make temporal judgments. Further evidence to this end comes from Alais and Burr (2004), who report that observers weigh the reliability of information from both auditory and visual streams to make positional judgements from audiovisual stimuli.

There is conflicting evidence as to whether the relationship between auditory and visual domains are strictly dominant in their given fields (auditory for temporal; visual for spatial), or whether there are some higher level mechanisms controlling which system is used for a given task. What is clear, however, is that temporal and spatial factors both play a significant role in the integration of audio-visual stimuli. The task that will be employed in the current research is one that seeks to use both temporal and spatial features to induce the unity of auditory and visual processing.

Temporal Factors

One of the clearest cues leading to the decision to bind sensory inputs is that they co-occur within the same, brief time frame. In terms of this temporal coincidence, it would seem that audio-visual binding should be optimal when the sensory components are presented simultaneously (Vroomen et al., 2004). However, research has revealed that binding can also occur on either side of truly simultaneous presentation. Researchers often describe a *temporal window of integration*: a range within which auditory and visual information can be bound. For example, van Wassenhove, Grant & Poeppel (2007) found that participants reported audio and visual elements as being fused (defined as an illusory McGurk combination of auditory and visual information) if they were presented within an approximate 200 ms window, from auditory information being presented -30 ms before visual information (hereafter, *audio lead*) to auditory information trailing visual information by +170 ms (hereafter, *audio lag*). A consistent asymmetry is apparent in previous data in that auditory and visual information are more likely to bind under conditions of auditory lag relative to auditory lead (van Wassenhove et al., 2007; Soto-Faracao and Alsius, 2009). Soto-Faracao & Alsius (2009) go even further to distinguish between the temporal window of integration (a region within which binding can occur) and the point of subjective simultaneity (PSS), which is the ideal point for multimodal binding. One reason for this temporal asymmetry may be due to a difference in speeds of transmission of auditory and visual information in the air and in the cortex, as shown by Fujisaki, Shimojo, Kashino, and Nishida (2004). They found that when visual and auditory information arrives simultaneously at their respective sensory transducers, cortical responses to audition are approximately 30 to 70 ms faster than responses to vision. This seems to be a compensation for the faster rate of transmission of visual information in air, allowing auditory information to catch

up in the brain after falling behind in the environment, and as such eventual cortical activation in visual and auditory areas is concurrent under conditions of peripheral auditory lag presentation. An important caveat of this work, as will be shown later, is that audio-visual compensation in terms of perceptual simultaneity may be limited to certain physical ranges. As indicated by the temporal window of integration, the influence of temporal proximity on audio-visual binding does have limits – research determining these limits have had varying results, but a general guideline for the judgment of synchronous presentation is that the auditory stimulus should be presented somewhere between 100 ms before to 200 ms after the visual stimulus (Dixon and Spitz, 1980; Lewald and Guski, 2003; Spence and Squire, 2003), although there is evidence for neurons in the superior colliculus that can integrate over a 600 ms time course (Stein & Stanford, 2008). These limits can also vary across participants, but tend to be relatively stable within an individual (Stone et al. 2003).

Similar influences on the temporal window of integration have been shown in terms of the apparent causal relationships between auditory and visual stimuli. As light travels faster in air than sound, the more common response would be that a visual event causes a subsequent auditory event. An example of this is the oft-repeated statement that thunder ‘comes from’ lightning. The two are, in fact, both generated by the same event, but because of the temporal order they are perceived in, causality is assigned to the former event. Rather than A causing V (or vice versa), there is a primary event, X, which causes both A and V. The temporal disparity of A and V arriving at the perceiver is a function of the relative speeds of sound and light in air, and this temporal difference is one of the factors contributing to the binding decision for these auditory and visual stimuli. Scholl and Nakayama (2000) discuss ‘causal capture events’, in which a causal interpretation of an ambiguous audiovisual display is only possible with less than

100 ms of difference between presentation of stimuli. If the auditory and visual stimuli are presented with an SOA greater than 100 ms, causality judgements decrease. Kohlrausch and van de Par (2000) confirm this finding, stating that causality is best perceived when a sound occurs 50 ms after the visual event (once again, providing evidence for the allowances made for audio lag). On the basis of this previous literature, it was possible to determine the likely range within which binding would occur in the current experiment. As such, the stimuli derived had a range of 400 ms overall, with a 200 ms region between two anchors (each of which has 100 ms duration) marking the start and end of trials.

Spatial Factors

While temporal factors play a large role in whether or not auditory and visual inputs are perceived as having a common cause, the contribution of spatial information also cannot be overlooked. The well-known ventriloquism effect (first described by Howard & Templeton, 1966) shows that auditory signals can be captured by visual stimuli that occur relatively closely in time and space, and as such the perceived position of an auditory signal can be altered by a particularly salient visual stimulus. This apparent visual dominance has been demonstrated in speech sounds, as well as more basic stimuli using visual spots and auditory pure tones (Berman & Welch, 1976; Radeau, 1985). Slutsky and Recanzone (2000) presented such stimuli at variable lags from the audio leading by 250 ms to the audio lagging by 250 ms and participants were asked to report whether the two stimuli began at the same time or not. These findings demonstrated that temporal disparity was less effective in disrupting the ventriloquism illusion if auditory information led visual than if it lagged behind, thereby underscoring that the temporal asymmetry of auditory lag impacts on decisions regarding spatial binding.

It is worth bearing in mind that the majority of audiovisual binding experiments are performed in standard laboratory settings, using computer screens with adjacent speakers (or headphones). This makes it relatively easy to study azimuth (left to right) variation but relatively difficult to study spatial effects in terms of distance on audiovisual binding, given that there is not enough space in most labs to generate distant stimuli. However, some recent experiments have taken on this challenging goal, with interesting results.

Kopinska & Harris (2004) presented sound-light pairings at variable lags, from audio leading by 200 ms to audio lagging by 200 ms. They presented these stimulus sets at distances ranging from 1 to 32 m, and asked participants to report when the sound-light pairs were occurring simultaneously. They report that sound-light pairings were judged as being simultaneous when the light and sound left the source at the same time. That is to say, the perceptual system compensates for distance such that long distances allow for longer lags between auditory and visual components. These findings build upon the earlier work of Engel & Dougherty (1971), who found that in order for observers to perceive simultaneity, sound needs to lag behind sight by 100 ms to account for 35 metres of perceived distance, once again highlighting the accommodation made for auditory lag conditions. Alais and Carlile (2005) found that compensation for distance follows a constant slope, allowing for 3.43 ms of delay for every metre of distance. They point out that this is very close to the speed of sound in air (343 m/s), indicating the sensitivity with which the brain adjusts for both long and short distances.

Sugita and Suzuki (2003) corroborated these findings, showing a point of subjective simultaneity (PSS) that was influenced by distance at a rate of approximately 3 ms per metre, up to a distance of around 10 m. However, they found that this seemingly linear trend of compensation is not constant beyond 10 metres. They concluded, therefore, that a different

compensatory mechanism is in operation beyond this point. A similar conclusion was reached by Lewald and Guski (2003) who found that a delay of up to 45 ms enables binding up to a distance of 20 metres, after which participants were no longer able to bind audio and visual information into a unitary percept.

In sum, the above research supports the idea of Spence and Squire (2003) that we have a sliding window for temporal integration based on the distance stimuli are away from us. So the previously reported window of audiovisual binding (Soto-Faracao & Alsius, 2009; van Wassenhove et al., 2007) appears to hold for relatively short distances, but shifts to accommodate greater auditory lags as distance from the signals increases. At the limit (over 20 m), there maybe nothing to reconcile auditory and visual signals generated by the same source, which arrive at the ear and the eye too far apart in time to be bound.

The studies presented above have tended to use static distance displays. That is to say, sources were presented at a set distance from the participant, but did not move towards, or away from, the participant. Within our natural environment, sources move in dynamic and potentially important ways such as the visual shape of a predator getting larger in size, or, an ambulance siren getting louder in intensity. Therefore, a final consideration in spatial variation is the movement (real or virtual) expressed in the sources and the degree to which dynamic change influences audio-visual binding. Cappe et al. (2009) looked at auditory and visual looming and receding by presenting an image of a disk which either got larger (looming) or smaller (receding), and sounds which got louder (looming) or quieter (receding). They found that multisensory looming showed the fastest reaction times when compared to multisensory receding or ambiguous pairings of stimuli (when auditory was receding while visual was looming, or the reverse). This adds to the previous literature on looming, showing that a

looming stimulus is more salient than a static or receding stimulus, and that when it is looming in both auditory and visual dimensions it increases this effect (Neuhoff, 2001; Maier et al., 2004). Therefore, a reasonable analog of dynamic distance appears to be the manipulation of size in the visual domain and the manipulation of loudness in the auditory domain. Moreover, these two properties represent prothetic (as opposed to metathetic; Smith & Sera, 1992) dimensions and index amodal characteristics of magnitude (Walsh, 2003). As such, there is a possibility of simulating distance within-modality both in terms of static (near or far) or dynamic (looming or receding) qualities by manipulating the values the auditory or visual stimuli (e.g., static near: large-large or loud-loud; dynamic looming: small-large or quiet-loud). At the same time, between-modality associations might be promoted as the ‘high magnitude’ (loud or large) and ‘low magnitude’ (quiet or small) stimuli may in fact be coded as the same amodal magnitude level in the brain. By considering variation in both size and loudness, it becomes possible to explore how both within- and cross-modal congruency impacts on AV binding.

Congruency Factors

In addition to the time course and spatial position within which auditory and visual stimuli occur, evidence has also been presented suggesting that the relationship between the auditory and visual stimuli can affect the likelihood of integration. The notion of *congruency* can be expressed at a perceptual (e.g., Walker, Bremner, Mason, Spring, Mattock, Slater & Johnson, 2009), semantic (e.g., Zhu, Zhang, Wu, Luo & Luo, 2010), or stimulus-response (e.g., Soetens, Maetens, Zeischka & Henderickx, 2010) levels (see Spence, 2011 for a review). An investigation of multiple possible combinations of auditory-visual congruency pairings by Marks (1987) found that there are relationships between pitch and visual lightness, pitch and visual brightness, loudness and brightness, and pitch and visual form. A more recent study by Evans

and Treisman (2010) tested for cross-modal mappings of auditory pitch onto visual position, size, spatial frequency, and contrast, and found that mapping auditory pitch onto visual position resulted in the shortest reaction times on an identification task. From this large corpus of evidence, it becomes clear that there are many possible congruent relationships between auditory and visual information, although the point at which this information is shared between senses varies.

In terms of perceptual congruency, auditory and visual information should conform to norms learned from real life experiences (Evans and Treisman, 2010). For example, small visual objects are more likely to generate high pitches whereas large visual objects are more likely to generate low pitches because, due to physical constraints related to sound production. Therefore, statistically speaking, we have encountered these ‘congruent’ pairings more often than the ‘incongruent’ ones, and have learned this relationship, although not all researchers agree on this basic point (see Bien et al., 2011; Ludwig et al., 2011). Parise & Spence (2009) confirmed that for audio-visual combinations of visual circle size and auditory pitch, congruent pairings (large circles associated with low pitch, small circles associated with high pitch) were more likely to be judged as synchronous at larger time lags, relative to incongruent pairings (large circles associated with high pitch, small circles associated with low pitch). Here, congruency between auditory and visual values is seen to expand the temporal window of integration, and the perception of synchrony.

Particularly relevant to the current research is a study reported by Gallace & Spence (2006). Here, two disks of a different size were sometimes accompanied by a sound simultaneous with the second disk. They found that responses were slowest when no sound was presented, and fastest when the sound was congruent with the second disk (with an incongruent sound falling in

the middle). The overall finding that the ‘synaesthetically’ congruent sound facilitated responding shows that participants were unable to avoid processing auditory information, even as they were asked to ignore it and attend to only the visual stimuli. Although there is evidence here to suggest that auditory information contributed to visual decision making, it is not currently clear whether the reverse obtains, namely, whether visual information will contribute to auditory decision making. Moreover, the Gallace & Spence (2006) study contained some aspects of a dynamic distance analog (i.e., big and small size of visual disk) but only studied the cross-modal congruency between the auditory tone and the second visual disk. The current study aims to compare elements of cross-modal and within-modal congruency in the same design.

Present Research

The present research elaborates on the procedure of Gallace & Spence (2006) in a number of ways to study the impact of both within- and cross-modal congruency (expressed as dynamic distance cues) and the role of temporal factors on audio-visual binding. The first two experiments were performed, using the same presentation paradigm but reversing the modality order. That is to say, Experiment 1 (VAV) examined a single roving auditory stimulus located at various time points from onset to offset with a primary visual anchor, through an ambiguous region, and from onset to offset with a secondary visual anchor. Experiment 2 (AVA) considered a single visual stimulus situated at the same time points between two auditory stimuli. However, in both cases, the roving stimulus had to be assigned as being caused by either the primary or secondary anchor. Experiment 3 considered the congruency relationships between auditory and visual stimuli in order to maximize the probability of binding. Experiment 4 combined the previous experiments, allowing direct comparisons between the paradigms of

Congruency and temporal asynchrony in audio-visual perception

Experiments 1 and 2 in a within-participants design, along with the increased cross-modal congruency between auditory and visual stimuli established in Experiment 3.

Given what we know about temporal presentations (Calvert, Spence & Stein, 2004) and causality (Scholl & Nakayama, 2000) we can predict that there will be a general bias in Experiment 1 (VAV) towards responding that the auditory stimulus came from the primary anchor, with a corresponding bias towards the secondary anchor in Experiment 2 (AVA). This pattern of responding aligns with previous research, and follows the expected ecological combination of pairing an auditory stimulus with a visual stimulus that preceded it. In terms of additional unity assumptions made at the level of between-modality congruency, we also would expect small visual stimuli to be more likely to be bound with quiet auditory stimuli, and large to go with loud (Walsh, 2003; Spence, 2011). It is hypothesized that these combinations will yield greater evidence of binding outside of the ‘normal’ range of temporal integration. More importantly, we also will be able to examine cases in which temporal coincidence and congruency relationships are put in competition with each other, and in doing so compare which will be more influential in the binding of the roving stimulus. Take, for example, during VAV context, a case of large-loud-small, where the roving stimulus is congruent with the primary anchor, but is presented simultaneously with the secondary anchor. Temporal coincidence dictates that it should be paired with the secondary anchor, but congruency relationships suggest the primary anchor. These experiments will shed light on how these two factors work together (or against one another) in the binding of auditory and visual stimuli.

Finally, we will be able to examine effects of simulated static and dynamic distance on audiovisual binding by examining within-modality congruency. In Experiment 1 (VAV) we have visual anchors being ‘far’ (small, small), ‘near’ (large, large), ‘looming’ (small, large) or

‘receding’ (large, small), and in Experiment 2 (AVA) auditory anchors can be similarly described in each of these ways (far = quiet, quiet; near = loud, loud; looming = quiet, loud; receding = loud, quiet). Therefore, we may expect ‘distant’ stimuli to show a greater auditory lag tolerance, as auditory stimuli may take longer to reach the perceiver than the visual stimulus coming from far away. Similarly, looming stimuli may be more likely to attract attention than receding stimuli (Neuhoff, 2001), leading to increased secondary responding. Finally, we will be able to consider the differential roles of audition and vision, given their traditional statuses in spatial effects (visual dominance) and temporal effects (auditory dominance). On the basis of the task demands literature, we would expect visual stimuli to show a stronger looming and/or receding effect than auditory stimuli, given that it is a spatial (distance) relationship (e.g. Cappe et al., 2009), and that auditory stimuli would show a stronger temporal effect, with more reliance on the time course of stimulus presentation (e.g. Soto-Faracao & Alsius, 2009).

Experiment 1

Method

Participants. Informed consent from 30 participants was obtained prior to the study, and they were compensated with partial credit in an undergraduate psychology course. The 27 participants that were used in the final sample had a mean age of 20.8 years ($sd = 5.4$), and included 26 females and 27 right handed individuals. The experimental procedure was approved by the Research Ethics Board of Ryerson University, and in keeping with ethical practice participants gave their informed consent based on preliminary information, and were debriefed as to the purpose of the experiment upon completion of the study. (Consent and Debrief forms are reproduced as Appendices A and B).

Stimuli and apparatus. 100 ms 1 kHz, with 5 ms linear on-set and off-set ramps were developed using SoundEdit 16 (MacroMedia). All sounds were played binaurally from free-field speakers (Harman Kardon) positioned either side of a computer monitor viewed approximately 57 cm away to encourage spatial coincidence between auditory and visual signals (Calvert et al., 2004). All sounds were calibrated using a Scosche SPL100 sound level meter to approximately 66 or 71 dB(C) to represent quiet and loud sounds, respectively. The visual stimulus consisted of a yellow asterisk, presented in the centre of a black screen in either 48 or 96 point Chicago font to represent small and large sizes, respectively. Stimulus presentation was controlled by PsyScope (Cohen, MacWhinney, Flatt & Provost, 1993) and responses were recorded using a PsyScope Button Box.

Design. An initial experimental block of 360 trials was developed involving the orthogonal combination of primary visual anchor size (V1; small, large), roving auditory stimulus volume (A; loud, quiet) and, secondary visual anchor size (V2; small, large). These

eight ($2 \times 2 \times 2$) sets of stimulus were further varied by changing the temporal presentation of the sound with respect to the visual stimuli. The two visual anchors were always presented with a 300 ms difference between onsets, and were played for 100 ms each. The sound could occur simultaneously with the onset of V1 (0 ms), and at 50 ms intervals until 100 ms after the offset of V2 (400 ms), for a total of 9 possible temporal presentations (see Figure 1). In order to maximize causal judgements, it was ensured that the roving stimulus was never more than 100 ms away from one of the anchors (as per Kohlrausch and van der Par, 2000). The first 3 of these time points (0, 50, 100 ms) were coincident with the presentation of V1, the last 3 (300, 350, 400 ms) were coincident with V2, and the middle 3 (150, 200, 250 ms) fell into an ambiguous region between the two visual stimuli. This made for a total of 72 conditions, which were subject to a variable delay of 100, 200, 300, 400, or 500 ms between trials, which was not considered in any further analysis. Participants completed three blocks of 360 trials, which were preceded by a single practice block consisting of 12 trials taken randomly from the experimental blocks. In both practice and experimental blocks, trial order was randomized.

Procedure

Each trial began with the presentation of a blank screen for 500 ms, which was then followed by the variable lag. Participants were then presented with V1 for 100 ms, and following a 300 ms interval V2 was presented for 100 ms. A sound (A) was presented at some time between V1 and V2 presentation at 50 ms intervals. Stimulus presentation was immediately followed by a response prompt saying “FIRST OR SECOND?” Participants were asked to respond by pressing the left-most button on a PsyScope button box if they thought the sound was caused by the first visual event (V1), and the right-most button if they thought the sound was

caused by the second visual event (V2). No feedback was provided as a result of the subjective nature of the task.

Results

Cell means for each participant were computed. Given that the probability of A binding to V2 should increase with time regardless of congruency relations, a positive correlation between V2 response and time was expected. Consequently, individual participants' scores were subjected to a correlation analysis between V2 response and time, to ensure that all participants were sensitive to the temporal factors related to the study. Three participants were found to have very low correlation with the overall mean scores ($r_s = -0.35, 0.33, 0.24$) relative to the rest of the sample, who had correlations of at least 0.63 (with 24 of 30 participants having correlation greater than 0.90). The exclusion of these three sets of data resulted in a final sample of 27 participants.

The overall means from each cell are presented in Table 1. The data was then subjected to a repeated measures ANOVA with within-subjects factors of V1 (small, large), A (quiet, loud), V2 (small, large), and time (0, 50, 100, 150, 200, 250, 300, 350, 400 ms) using the proportion of V2 responses as the dependent variable. The full results of the ANOVA are presented in Table 2. Post-hoc tests took the form of Tukey's HSD test ($p < .05$).

A main effect of time, $F(8,208) = 106.56, p < .001$, shows an S-shaped function displayed in the VAV portion of Figure 2. This confirms the general tendency for the auditory stimulus to bind with the primary visual anchor when presentations overlap and with the second visual anchor when presentations overlap. In the region of ambiguity (between 150-250 ms), the effect shows a steady increase in V2 responses, with significant increases in second responding between 150 and 200 ms ($p < .001$), 200 and 250 ms ($p < .001$), and 250 and 300 ms ($p < .001$).

In addition, it was found that within the region of V1 certainty (0-100 ms), there was no significant increase in V2 binding between 0 to 50 ms or between 50 and 100 ms, nor were there significant increases between 300 to 350 ms or 350 to 400 ms.

The interaction between V1 and time was significant, $F(8, 208) = 3.56, p = .001$.

Individual differences within the data showed that at 200 (small V1 = .457, SE = .032, large V1 = .360, SE = .030) and 250 ms (small V1 = .612, SE = .034, large V1 = .538, SE = .035), small V1 increasing the tendency for second responding. There was also a main effect of V2, $F(1,26) = 5.58, p = .026$, showing that when the secondary visual anchor was large there was greater second responding (large V2 = .45, SE = .02) relative to when the second visual stimulus was small (small V2 = .43, SE = .02).

These effects were subsumed by a significant V1 x V2 x Time interaction, $F(8, 208) = 3.02, p = .003$, and is shown in Figure 4. Tukey's HSD revealed a significant difference at 300 ms, in that when V1 was small, a large V2 led to significantly more second responding than when V2 was small. A non-significant trend in this direction (small-to-large) was also evident in time points 200, 250, 350 and 400 ms. This potential looming effect was robust late in the region of ambiguity, and was also represented by other differences in this interaction. When V1 was small and V2 large (i.e., looming), this led to significantly more V2 binding than when V1 is large and V2 small (i.e., receding) at 200 ms ($p < .001$) and 250 ms ($p < .001$; compare top and bottom sections of Figure 2 at 200 and 250 ms), and more V2 binding than when V1 and V2 were both large at 200 ms ($p < .001$) and 250 ms ($p < .001$).

There was no main effect of auditory stimulus, and though the A x Time interaction was significant, $F(8,208) = 3.56, p = .004$. However, there were no significant differences found between quiet and loud auditory stimuli at any time point. As such, it appears that the effects in

this VAV condition are being driven primarily by the stimulus relationships within the visual domain.

Discussion

The main effect of time supports the role of temporal factors in audiovisual binding, forming a function showing a gradual increase in binding to the second visual stimulus. Within the regions of certainty (when the audio was presented at the same time as one of the visual stimuli), the probability of association was largely insensitive to both auditory and visual quality. In considering the timepoint when the sound follows the visual by 100 ms, which is very close to the ideal point of subjective simultaneity (van Wassenhove et al., 2007), we see that the strength of the subjective temporal coincidence ensures primary responding, regardless of congruency factors. Within the region of uncertainty (when auditory stimulus occurred between 150 and 250 ms), there was a sharp increase in the likelihood of binding to the second visual stimulus, as the influence of visual congruency also comes into play.

A V1 x time interaction indicated that in the majority of the region of ambiguity, a small V1 leads to higher probability of binding to V2. This is particularly interesting in that based on causality (Scholl & Nakayama, 2000) and temporal window of integration (van Wassenhove et al., 2007) arguments, we would expect A to be bound with V1. In light of the literature referring to audiovisual binding over distance, there may have been an expectation for a wider binding window for a small (i.e. distant) V1. However, this is not the case in this data, and so it seems that the distance manipulation is not effective. This V1 x time interaction seems to be the complement of the main effect of V2 that was found, showing a general tendency for the auditory stimulus to be bound to the secondary visual anchor when it was large more so than when it is small. Therefore, both a lower salience V1 and a higher salience V2 may collectively

promote V2 binding. This must also be considered in light of the literature on looming (e.g. Neuhoff, 2001), which show that a stimulus that is approaching the participant has an alerting property. Although the visual stimulus size increase is not continuous, there is still potential for a looming effect to occur, given that the retinal image of the visual stimulus is increasing over a brief (200 ms) time course. Olsen and Stevens (2010) showed that stimulus continuity is not necessary for looming effects to be exhibited. Considering the interaction between the two anchors over time will reveal whether these effects are based on individual stimulus properties, or on perceived looming.

The final significant interaction was a V1 x V2 x Time interaction that showed that with a small V1, binding was the same from 0-150 ms, regardless of V2. However, these effects were in the region of ambiguity and throughout the time course of V2 presentation, a large V2 led to a greater probability of V2 binding. Perhaps there is a general bias towards binding to the secondary visual anchor, which is in contrast with the general findings in the literature (i.e. van Wassenhove et al., 2007; Soto-Faracao & Alsius, 2009). This demonstrates a bias in favour of temporal factors when they are unambiguously presented, as well as a preference for binding to the primary anchor when it is coincident over binding to the secondary anchor when it is coincident (Roseboom & Nishida, 2009). Following V1 presentation, preference for binding to the second visual stimulus increased and was modulated by small V1 or large V2. This unique effect of simulated looming (V1 small to V2 large) is not mirrored by a simulated receding (large, small) stimulus pair (Neuhoff, 2001) and so the effect cannot be attributed to stimulus change per se. In summary, it appears that stimulus characteristics only factor into the decision-making process when binding cannot be resolved through use of temporal factors.

Looking forward to Experiment 2, it will be interesting to consider whether there is a difference in the overall binding pattern between the VAV paradigm in Experiment 1 and the AVA paradigm in Experiment 2. While in Experiment 1 we saw no between-modality congruency effects driven by the magnitude of the roving auditory stimulus, so it seems that the two visual stimuli are more pertinent in creating the congruency effects that are observed. Given that visual stimuli are generally considered stronger in making spatial judgements (which, in this case, is simulated distance), it is possible that in Experiment 2 the roving visual stimulus will still drive some effect even though it will be the lone visual stimulus between two auditory anchors. That having been said, the task remains a temporal one, so the overall shape of the results should remain the same, with differences occurring when there are not sufficient temporal indicators to enable binding.

Experiment 2

Method

Informed consent from 29 participants was obtained prior to the experiment, and they were compensated with partial credit in an undergraduate psychology course. The 27 participants included in the final sample had a mean age of 19.7 years ($sd = 3.8$), and included 21 females and 23 right handed individuals. Stimulus parameters were identical to Experiment 1, apart from the order of presentation of stimuli (auditory – visual – auditory, rather than visual – auditory – visual).

Results

Participant data was screened according to the same method as the VAV experiment, and cell means were calculated following the same method as well. Two participants were excluded based on low correlation ($r_s = -.74, .33$), relative to an overall mean correlation of .95, leaving a final data set of 27 participants, from which the cell means were calculated.

The overall means from each cell are presented in Table 3. The data was then subjected to a repeated measures ANOVA with within-subjects factors of A1 (quiet, loud), V (small, large), A2 (quiet, loud), and time (0, 50, 100, 150, 200, 250, 300, 350, 400 ms) using the proportion of A2 responses as the dependent variable. The full results of the ANOVA are presented in Table 4.

There was a significant main effect of time ($F(8,208) = 139.08, p < .001$), following essentially the same S-shaped curve as found in VAV. It is of interest, however, that the AVA curve led to a greater number of secondary anchor binding responses than the VAV curve (see Figure 2), indicating that there is a greater tendency to bind V to A2 in AVA than binding A to V2 in VAV. This was confirmed by an additional ANOVA, using a between-participants

variable of experiment (VAV, AVA) and time (as in previous experiments). This ANOVA yielded a main effect of experiment, $F(1, 52) = 5.11, p = .028$, confirming the fact that the AVA paradigm leads to more binding to the secondary anchor ($M = .502, SD = .019$) than the VAV paradigm ($M = .442, SD = .019$). This was in line with predictions, given the tendency for audio to follow visual in nature, as well as previous findings (see Soto-Faracao & Alsius (2009), and Sugita & Suzuki (2003), as examples). The condition x time interaction was not significant, $F(8, 416) = 1.68, p = .10$, indicating that while there is a general difference in binding frequency between the two conditions, this did not modulate as a function of time although the effect is most clearly shown during the ambiguous region (150 – 250 ms) and during secondary anchor presentation (300 – 400 ms).

Within the AVA experiment, there were main effects of A1, $F(1, 26) = 33.78, p < .001$, and V, $F(1, 26) = 8.52, p = .007$. These two main effects are best understood by the interaction of A1 x V ($F = 7.35, p = .012$), which shows (Figure 5) that the probability of second responding was attenuated for the combination of loud A1 and large V. There were significant A1 x time, $F(8, 208) = 2.27, p = .024$, and, V x time, $F(8, 208) = 2.42, p = .016$ interactions, which are shown in Figures 6 and 7, respectively. Tukey's HSD showed that loud A1 reduced the likelihood of second responding relative to quiet A1 at 200 and 250 ms, and, small V reduced the likelihood of second responding relative to large V at 150 ms.

Discussion

As per Experiment 1, the main effect of time is shown through a function of gradual increase in binding to the second visual stimulus following the offset of the first auditory stimulus, and particularly for the region of ambiguity. A comparison between Experiments 1 (VAV) and 2 (AVA) showed a greater preference for second stimulus responding after the offset

of the primary anchor, and a steeper increase during the region of ambiguity in the case of AVA presentation (although these differences were not statistically significant). Within the time course of presentation of the primary anchor, there is no difference between the AVA and VAV paradigms. This is likely because there is little difference between the two paradigms in this situation. In fact, at the first (0 ms) time point, the two paradigms are identical, and so there is no reason to expect a difference. However, we do see differences between paradigms from 150-250 ms, in what has been termed the ‘ambiguous region’, with no differences when the roving stimulus occurs simultaneously with the secondary anchor. Within the ambiguous region, the responses for AVA are more likely to associate with the secondary anchor than the VAV responses. This is consistent with previous literature on causality in audiovisual binding (Scholl & Nakayama, 2000; Kohlrausch & van de Par, 2000), and the rule that auditory stimuli are more likely to follow visual stimuli than the opposite arrangement.

The A1 x V interaction shows a between-modality congruency effect, with a loud first auditory stimulus attracting a large visual stimulus more so than any other combination, evidenced by a reduction in second responding. However, this effect only works in one direction, without a corresponding effect for quiet first auditory and small visual stimuli. In this case, it is possible that this effect is due to a perceived distance effect, wherein the loud-large combination is perceived as being close to the participant, and so the lag between audio and visual is expected to be small. Conversely, the quiet-small combination is perceived as being distant from the participant, and so compensation occurs leading to the participant expecting the sound to lag further behind the visual stimulus in order to associate them (as in Kopinska & Harris, 2004; Sugita & Suzuki, 2003). Another explanation for this phenomenon could be along

this lines of an increased salience of high magnitude stimuli, and further research is necessary to parse out these possibilities.

There were also significant interactions of A1 x Time, following a similar pattern to the V1 x Time interaction from Experiment 1. The A1 x Time interaction once again indicates that when the primary anchor is of a high magnitude (in this case, loud), it is more likely to attract the binding of the roving stimulus. This similarity between the two experimental paradigms, even as different modalities are being responded to, supports Walsh's (2003) ATOM theory of magnitude, which holds that size and volume may be different modal representations of the same amodal magnitude level.

Overall in the first two experiments, there is a trend for stronger auditory effects in the temporal domain, and stronger effects of vision in the (simulated) spatial domain. We see more temporal precision in the AVA condition, as evidenced by the steeper curve during the ambiguous region, which indicates greater temporal discriminability using auditory anchors (Vroomen and Keetels, 2010). In contrast, there is a stronger within-modality congruency effect in the VAV condition, which could be defined as a spatial effect in terms of perceived distance. There is also evidence for visual stimuli influencing auditory judgements (in the AVA paradigm), but no evidence for auditory stimuli influencing visual judgements (in the VAV paradigm). This is in conflict with the findings of Gallace & Spence (2006), who used a very similar task (although with only a single lag between auditory and visual) and found that participants were unable to disregard auditory stimuli when they were meant to be attending to the visual task. It is likely that there may be some confound stemming from the relative discriminability of the auditory and visual values used. One alternate explanation for visual dominance is that the size difference between visual stimulus levels is more apparent than the

loudness difference between auditory stimulus levels. This disparity between the relative magnitudes of the two modalities led to an inequality in the difficulty of the auditory and visual discriminations. In order to help us resolve these issues, Experiment 3 was performed as a calibration exercise. With it, it was hoped that auditory and visual levels could be found, which maximized the congruency relationship between the high and low magnitude stimuli, as well as minimized the differences between auditory and visual discrimination difficulty. To do this, participants completed a selective attention task on bi-modal stimuli (see Patching & Quinlan, 2002) with the aim that the stimuli selected in Experiment 4 would be the largest with respect to cross-modal congruency and the smallest with respect to differences in processing differences between audition and vision.

Experiment 3

Method

Participants. Informed consent from 44 participants was obtained prior to the experiment, and they were compensated either with partial credit in an undergraduate psychology course or \$10 cash. 9 participants were eventually excluded on the basis of high error rates (above 30%, compared to a remaining group average of 13.5%). The remaining 36 participants had a mean age of 21.8 years ($SD = 6.3$), and included 26 females and 35 right handed individuals.

Stimuli and apparatus. All stimulus parameters were as per Experiments 1 and 2, apart from the manipulation of volume and size. All sounds were calibrated using a Scosche SPL100 sound level meter to approximately 71, 66, 61 or 55 dB(C) to represent four levels of volume (A1, A2, A3, A4). The visual stimulus consisted of a yellow asterisk, presented in the centre of a black screen in 96, 48, 24 or 12 point Chicago font to represent four levels of size (V1, V2, V3, V4).

Design. The four respective levels of volume and size were combined in such a way that each condition would consist of two possible volumes and sizes, with one of them always being the maximum volume (A1: 71 dB) and the maximum size (V1: 96pt font) as per Experiments 1 and 2. This yielded 9 possible condition sets (see Figure 8), and an initial experimental block of 96 trials was developed for each of these condition sets. These blocks involved orthogonal combinations of the two auditory stimuli (e.g., A1 and A2) and two visual stimuli (e.g. V1 and V3) within each condition set. These four combinations (e.g. A1V1, A1V3, A2V1, A2V3) consisted of two congruent (small/quiet, or, large/loud) and two incongruent (small/loud, or, large/quiet) pairings, and were further varied by the response modality – participants were asked

to respond to either the auditory or visual aspect of the bimodal stimuli. This made for a total of 8 stimulus-response combinations, and completing 12 repetitions of each combination yielded 96 total trials. Participants completed one block per condition, for a total of nine blocks of 96 trials. Their first experimental block was preceded by a 12 trial practice block in the same condition as their first experimental block. In both practice and experimental blocks, trial order was randomized.

The order in which participants completed conditions was counterbalanced by combining the 6 permutations of the levels in each modality. Given that levels V1 and A1 were fixed in all conditions, the levels of audio were varied within each level of vision, in a fixed order (A2, A3, A4 within V2; A2, A3, A4 within V3; A2, A3, A4 within V4). For the next participant, the levels of visual were varied within each level of audio, in the same fixed order (V2, V3, V4 within A2; V2, V3, V4 within A3; V2, V3, V4 within A4). Using the six permutations (234, 243, 324, 342, 423, 432), and by varying the order of manipulation (audio first or visual first) 12 combinations were possible. Three participants were run with each combination, and when participants were excluded due to high error rate, their data was replaced by another participant who received the same combination of conditions.

Procedure. Each trial began with the presentation of a blank screen for 500 ms. Participants were then simultaneously presented an auditory and visual stimulus for 100 ms. After a 250 ms delay, participants were given an audiovisual prompt consisting of a written and spoken ‘A’ (for audio responding) or ‘V’ (for visual responding). Participants were asked to respond by pressing the left-most button on a PsyScope button box if they thought the stimulus in question was small (visual) or quiet (auditory), and by pressing the right-most button if the stimulus in question was large (visual) or loud (auditory). Feedback was supplied in the form of

a green cross and corresponding sound for a correct answer, or a red cross and corresponding sound for an incorrect answer.

Results

Data were trimmed by removing any responses falling outside of two standard deviations from the mean within each cell. Additionally, only reaction time data from correct responses were used in analysis. Individual trials were categorized as congruent (small/quiet, or, large/loud) and incongruent (small/loud, or, large/quiet) cases and performance was compared across the 9 conditions, yielding a 3 (size difference: hard, intermediate, easy) x 3 (volume difference: hard, intermediate, easy) x 2 (response modality: audio, visual) x 2 (congruence: congruent, incongruent) ANOVA.

Full results of the ANOVA on reaction times are shown in Table 5 below. The RT ANOVA showed a main effect of congruency, $F(1,35) = 21.85, p < .001$, with congruent audiovisual pairings ($M = 686$ ms, $SE = 26$ ms) being responded to significantly faster than incongruent pairings ($M = 713$ ms, $SE = 24$ ms). The presence of the congruency effect confirms inefficient separation of auditory and visual information at the time of responding, thus audio-visual binding. The interaction between visual and auditory difficulty was significant, $F(4,140) = 3.00, p = .021$ (see Figure 9) although no significant pairwise comparisons were shown.

Full results of the error rate ANOVA are displayed in Table 6. The ANOVA again produced the expected main effect of congruency, $F(1,35) = 24.08, p < .001$, with congruent trials ($M = .118$, $SE = .012$) producing significantly less errors than incongruent trials ($M = .151$, $SE = .014$). The error rate data also showed main effects of auditory difficulty, $F(2,70) = 12.61, p < .001$, and response modality, $F(1,35) = 40.93, p < .001$, subsumed by a significant two-way

interaction between auditory difficulty and response modality, $F(2,70) = 23.09$, $p < .001$ (see Figure 10). Tukey's HSD test showed auditory difficulty did not modulate as a function of visual response but did modulate as a function of auditory response. Error rates decreased as auditory difficulty decreased and were generally higher than visual responses.

Discussion

In this calibration experiment, we found a main effect of between-modality congruency, but without interactions for visual and auditory difficulty. Participants appeared globally sensitive to the difficulty of the auditory discrimination, but only when it was task-relevant. As such, there is no inferential evidence to suggest which of the stimulus pairings is the 'best match' to produce a maximum congruency effect in subsequent experiments. In Figure 11, it is apparent that the 'easy' auditory discrimination produced the largest congruency effect, regardless of which of the three visual discriminations it is paired with.

There are several possible reasons for the lack of significant findings with regard to which auditory and visual levels to use. Given that I chose just four respective levels in each modality, it is possible that none of the combinations were congruent enough with one another to yield significant findings. To that end, it may have been useful to use a continuous set of stimulus levels (e.g. 56, 57, 58...71 dB) rather than the four equidistant levels that were employed. However, if this was done it would greatly increase the number of trials required, and as such it would likely require a narrowing of the range of stimuli. Using the four discrete levels allowed a balance between resolution and range for the stimulus levels. Another option could be to use individually calibrated stimulus pairings, following Sandhu & Dyson (2012). They used pairings that each individual participant showed were maximally crossmodally correspondent in an initial block, rather than trying to find one set that was consistent across participants. Given

the level of individual differences shown in the sample in the experiment, this may have been a better route to take. That being said, there are still useful conclusions to be drawn from the experiment as it was conducted.

While there is no statistically significant evidence for which pairing to use for Experiment 4, there is still value in using the numerical differences found in Experiment 3 to select the pairing which produces the greatest congruency difference (that is, lower reaction times and error rates for congruent than incongruent pairings), while having a minimal difference between auditory and visual responding (as per Vroomen and Keetels, 2010). To this end, it was decided that the best pairing was an easy auditory discrimination ('loud' = 71 dB, 'quiet' = 56 dB) with a moderate difficulty visual judgment ('large' = 96 point font, 'small' = 24 point font). In terms of RTs, this pairing produced a large congruency effect (36 ms) and minimal audio-visual differences (-2 ms). This is superior to the congruency effect (26 ms) found in the stimuli used in Experiments 1 and 2, and more importantly is a much smaller AV difference than the 23 ms difference found in Experiments 1 and 2. In terms of error rates, while the chosen pairing did have a small audio-visual difference (6.7%; compared to 16.4% in Experiments 1 and 2), it did not have a particularly high congruency effect (2.0%; compared to 1.0%). Overall, it was still decided that this was the best pairing to proceed with.

Interim Summary

The first pair of complementary experiments yielded results demonstrating a general tendency for audiovisual binding to occur in accordance with principles of temporal coincidence, but also showed that this binding is modulated by congruency relations between stimuli, particularly in the region of temporal ambiguity in which the to-be-bound stimulus is not physically coincident with either competing stimulus. In terms of VAV performance,

Congruency and temporal asynchrony in audio-visual perception

congruency was defined according to the relationships within visual stimuli alone. In terms of AVA performance, congruency was defined between auditory and visual stimulus modalities. One concern regarding the difference in the data is due to the salience of visual and auditory values. In order to maximize and equate the congruency / incongruency relationship between auditory and visual stimuli in Experiment 4, the findings of Experiment 3 were used to choose a pair of volumes and sizes that were found to have a high level of congruency, as well as a low level of difference between auditory and visual responding. To confirm the global differences between Experiments 1 and 2, the comparison between VAV and AVA in Experiment 4 became a within-participants factor. To make the final experiment manageable, the time points used in Experiment 4 were reduced from 9 (at 50 ms intervals) to 5 (at 100 ms intervals). In doing so, care was taken to retain the points of focus, using the onset of the primary anchor (0 ms), the offset of the primary anchor (100 ms), the ambiguous region (200 ms), the onset of the secondary anchor (300 ms) and the offset of the secondary anchor (400 ms).

Experiment 4

Method

Informed consent from 30 participants was obtained prior to the experiment, and they were compensated with partial credit in an undergraduate psychology course. The 27 participants making up the final sample had a mean age of 20.2 years ($sd = 2.0$) and included 24 females and 27 right handed individuals. All aspects of Experiment 4 were identical to Experiment 1 and 2, except for the following. Sounds were calibrated using a Scosche SPL100 sound level meter to approximately 56 or 71 dB(C) for quiet and loud sounds, while visual stimuli were presented at either 24 or 96 point Chicago font to represent small and large sounds, respectively. Two initial experimental blocks of 240 trials were developed involving expressions of the orthogonal combination of primary anchor variation (V1; small, large *or* A1; quiet, loud), roving stimulus variation (A; quiet, loud *or* V; small, large), secondary anchor variation (V2; small, large *or* A2; quiet, loud), roving stimulus time (0, 100, 200, 300, 400 ms), and, variable lag (see Experiments 1 and 2), not considered in subsequent analyses. Participants completed two blocks of 240 trials for the VAV condition and two blocks of 240 trials for the AVA condition, each of which were preceded by a single practice block consisting of 12 trials taken randomly from the experimental blocks. In both practice and experimental blocks, trial order was randomized. The order in which participants completed the blocks (VAV first or AVA first) was also counterbalanced between participants.

Results

Means were calculated for each condition, and are displayed in Table 10. The data were then subjected to a 2 (condition; VAV, AVA) x 2 (first stimulus [S1]; low magnitude [small or quiet], high magnitude [large or loud]) x 2 (second stimulus [S2]; low, high) x 2 (third stimulus

[S3]; low, high) x 5 (time; 0, 100, 200, 300, 400 ms) ANOVA, the full results of which are displayed in Table 11. There was, as expected, a main effect of condition, $F(1,26) = 11.68$, $p = .002$, with the AVA paradigm yielding more second responding than the VAV paradigm. This main effect mirrored the previous comparisons between Experiment 1 and 2, as did the significant interaction between condition and time ($F(4,104) = 15.07$, $p < .001$; see Figure 12). Tukey's HSD ($p < .05$) revealed that the difference between the conditions was now significant at 200, 300, and 400 ms, revealing more second anchor responding in AVA. A significant S2 by time interaction ($p = .008$) was present, but comparisons revealed no significant differences in terms of magnitude at any time point.

Main effects of S1 and S3 are best considered in terms of their significant interactions with time. The interaction between S1 and time, $F(4,104) = 19.31$, $p < .001$, is shown in Figure 13 shows that a low magnitude S1 (that is, a small visual or a quiet auditory stimulus) leads to more secondary responding than a high magnitude S1, and that this difference is significant at 100, 200, and 300 ms, which is identical to the results shown in Experiments 1 and 2. The S3 by time interaction, $F(4,104) = 3.16$, $p = .017$, follows a similar trend, but in the opposite direction. In this case, Figure 14 shows that a high magnitude (large visual or loud auditory stimulus) S3 leads to more second responding, with the difference being significant at 200 and 300 ms, consistent with the V2 x time interaction found in Experiment 1. This effect was not initially apparent in Experiment 2 and it is likely due to the lower discriminability of auditory stimuli in Experiment 2 and their subsequent improvement in Experiment 4. Through increasing the discriminability by conducting Experiment 3, the effect was found to exist in both VAV and AVA paradigms. Generally then, it seems that the likelihood of binding is increased by low magnitude primary anchors and high magnitude secondary anchors around the ambiguous

temporal region. In this regard, the data from Experiment 4 replicate the V1 x V2 x time interaction found in Experiment 1, wherein low magnitude primary anchors and high magnitude secondary anchors increased the likelihood of second responding.

A final four-way interaction between condition x S1 x S3 x time ($F(4,104) = 4.30, p = .003$; Figure 15) gives some insight into the constraints associated with within-modality congruency. We see that at 200 ms the VAV and AVA conditions represents the S1 magnitude effect previously discussed; low magnitude stimuli (small, quiet) increase the likelihood of second responding relative to high magnitude stimuli (large, loud; Tukey's HSD, $p < .05$). However, at 300 ms only the VAV condition shows that low magnitude S1 increases second responding only when in combination with high magnitude S3 ($p < .05$). This replicates exactly the visual looming effect observed in Experiment 1 (small V1, large V2) also observed at 300 ms.

Discussion

The main findings from Experiment 4 clearly replicate effects found in Experiments 1 and 2, respectively. The difference between the VAV and AVA paradigms are equivalent within the conditions of Experiment 4 are equivalent to the comparison between Experiments 1 and 2. By including paradigm as a within-participants factor rather than our post-hoc comparison with a between-participants factor the interaction between paradigm and time became significant, showing that there was a difference in the way participants responded based on temporal factors on the two tasks. This is in line with expectations from many sources (see van Wassenhove et al., 2007 for an example) indicating that to perceive audio-visual stimuli as simultaneous, the ideal temporal condition is a slight auditory lag. In the VAV paradigm, this would be expressed

as a preference to bind the A to the first V, while in the AVA paradigm this would be expressed as a preference to bind the V to the second A.

The S1 x time interaction in Experiment 4 mimics the V1 x time from Experiment 1 and the A1 x Time found in Experiment 2, with a general trend for a low magnitude first stimuli to be more attractive to secondary binding than high magnitude first stimuli. This effect was independent of modality, occurring for both VAV and AVA at 200 ms. With regard to congruency, we see a strengthening of within-modality congruency (between S1 and S3) relationship, with a low level primary anchor and a high level secondary anchor leading to more secondary binding, and which is significant at the 200 ms point in both conditions. While this effect occurs for both visual and auditory pairings, there is a further effect of looming that exists at 300 ms only in the visual pairing. That is to say, the simulation of visual looming via low-magnitude primary anchor and high-magnitude secondary anchor departs significantly from other conditions (such as receding, or constant simulated distance). This is in agreement with Cappe et al.'s (2009) finding that visual looming is more pertinent to the observer than visual receding, and, auditory looming and receding. Between-modality congruency relationships remained weak overall, and perhaps have been weakened further by minimizing the discrimination differences between auditory and visual stimuli, which were present in the stimuli chosen for Experiments 1 and 2. In Experiment 1, congruency as defined by stimulated distance was strong due to the use of visual anchors. In Experiment 2, weak effects of (spatial) congruency provided by the auditory anchors were helped by vision (Vroomen et al., 2004). In Experiment 4, clearer auditory discriminations might have meant that the visual 'crutch' was no longer needed, thus eliminating significant between-modality effects.

General Discussion

Through the experiments reported herein, the processes of audiovisual binding, and the interplay of temporal and stimulus factors that affect how decisions are made in a competitive setting have been probed. Experiment 1 examined the selective binding of an auditory stimulus, situated between two visual stimuli, with the participants being asked to choose which of those visual stimuli the auditory stimulus came from. In addition to temporal variation, the stimuli could vary on their level of magnitude, with two volumes for auditory stimuli, and two sizes for visual stimuli. The results showed that temporal factors played a large role, with temporally coincident stimuli being paired most often, and that there was also an effect of stimulus magnitude that modulated this relationship. In Experiment 1 we failed to find a between-modality congruency relationship, revealing a general tendency to bind to high magnitude stimuli (Walsh, 2003). There was also an additional looming effect, with higher secondary binding when the primary visual anchor was small and the secondary anchor was large. This seems to be an alerting effect based on simulated motion, when it appears that the visual anchors are in fact one entity, which is moving towards the participant. Experiment 2 followed the same paradigm, but with inverted modalities. By placing one visual stimulus between two auditory stimuli, both temporal and congruency effects were modulated. An interesting finding was that, while the temporal curves of the two paradigms were similar in shape, the curve for Experiment 2 showed a preference for binding to the second stimulus relative to Experiment 1, essentially strengthening the effects of binding at temporally coincident points. Temporal separation between auditory and visual rovers were held constant across Experiments 1 and 2 and so differences could not be attributed to reliability differences as a function of range. Additionally for Experiment 2, within-modality effects were abolished and between-modality effects were

shown, specifically shown as strong associations between loud primary anchors and large visual roves. In contrast to the use of identical temporal variation between VAV and AVA, one concern was that the visual magnitude values may have been more discriminable than the auditory magnitude values. This may have led to the between-modality effects in Experiment 2. This necessitated Experiment 3, which sought to equalize these difficulty levels and maximize the congruency relationships between auditory and visual stimuli. By presenting nine different combinations of auditory and visual stimuli, the pairing which best satisfied the joint criteria of largest congruency effects and smallest modality differences was determined, and these stimuli were used in Experiment 4. In this final experiment, participants were exposed to simplified paradigms based on Experiments 1 and 2 using a within-participants design. The findings largely replicated the findings of the first two experiments: whereas AVA showed stronger temporal divisions, VAV showed stronger congruency divisions. It was also the case that while Experiment 1 only showed within-modality congruency effects, and Experiment 2 only showed between-modality congruency, Experiment 4 showed only within-modality magnitude effects across both paradigms, with a final looming effect observed only in VAV. From this experimental series, we can speak to many of the issues regarding audio-visual integration. In turn, the effects of temporal factors, the effects of within-modality and between-modality congruency (as expressed via modulation of distance), and task demands will be considered. Future research possibilities will be presented, and theoretical conclusions will be drawn based on the literature reviewed and the research conducted.

Temporal Factors

Although the experimental series has considered the interplay of temporal and congruency factors, everything we have found in the results must be viewed through the lens of

an overarching temporal effect. All of the temporal manipulations in the main experiment showed a similar pattern of data, with regions of relatively consistent binding when the roving stimulus was presented coincidentally with either the primary or secondary anchor, and a region of ambiguity between the two anchors, during which the probability of secondary responding steadily increased (see Figure 2 for an example of this pattern). Having established this effect, it became possible to consider differences between the patterns in the various paradigms and stimuli used throughout the experiments. Specifically, there was a tendency for the AVA paradigm to show more secondary binding than the VAV paradigm (beyond the primary region of certainty), which is easily attributable to a causality relationship stemming from real life experience. As shown by previous research (see Scholl & Nakayama, 2000, and, Roseboom, Nishida, & Arnold, 2009), sensory systems are more likely to perceive auditory and visual stimuli as having a common source when the visual leads the auditory, due to the differential speeds of transmission of light and sound in air. We also see that the rate of increase during the region of ambiguity in the AVA paradigm is steeper than in the VAV paradigm, with a trend found between Experiments 1 and 2, and a significant interaction when comparing between paradigms within Experiment 4. This indicates that the temporal resolution when there are two auditory anchors and a visual roving stimulus is finer tuned than when there are two visual anchors and an auditory roving stimulus. Traditionally, audition has been put forward as the dominant modality when making temporal judgements (Burr, Banks, & Morrone, 2009), and vision has been held to be dominant in spatial judgements (Alais & Burr, 2004). Our findings support these ideas as we will see, given that temporal resolution is greater with auditory anchors (AVA), whereas congruency (i.e. spatial) effects are stronger in with visual anchors (VAV).

It is clear from Figures 2 (Experiments 1 and 2) and 12 (Experiment 4) that temporal cues are the dominant contribution in analysing audio-visual causality, with congruency factors coming into play when temporal information cannot provide a clear resolution. That is, effects of congruency are essentially absent when there is temporal overlap with one of the anchors (0, 50, 100 ms for the first anchor, and, 300, 350, 400 ms for the second anchor) but come into play during the region of ambiguity. Therefore, congruency works in the service of temporal factors in resolving audio-visual causality. Aspect of both within-modality and between-modality congruency will now be discussed.

Congruency Factors

In the experiments that were conducted, within-modal and cross-modal congruency (size and volume) were employed to serve as simulated distance cues (Cappe et al., 2009). Due to the relationship between retinal image and perceived size, and aural intensity and perceived volume, it was very easy to employ these cues to simulate distance, and in doing so create the illusion of common source due to two stimuli both being ‘near’ or ‘far’ relative to the participant. If, for example, a large star was paired with a loud sound, we would have expected cross-modal binding indicating that the simulated stimulus was close. At the same time, a small primary visual anchor followed by a large secondary visual anchor would be perceived as moving towards the participant. These simulated distance effects, along with lower level congruency effects, were manifest in various ways throughout the experiments.

An interesting contrast between the VAV and AVA paradigms is that while having two auditory anchors seemed to increase the temporal resolution of the causality assignment, having two visual anchors tended to increase the level of congruency effects affecting participants’ decisions. Taking into account the results of the calibration experiment, which showed that

visual discrimination was initially easier than auditory discrimination, one might attribute this stronger effect to the asymmetry in modalities. However, given that these findings still existed in Experiment 4 (when the asymmetry had been minimised), that account is not supported. Making a competitive decision between two auditory anchors relies more on temporal factors, which fits with canonical modality dominances. The preference for congruency information when using two visual factors also follows with this account, especially when we define the stimulus variation as simulated distance (i.e. space; Vroomen & Keetels, 2010).

In addition to the effects of congruency, we also find an alerting effect shown only by visual, looming anchor pairs. That is to say, when the anchors go from low magnitude to high magnitude, only in the visual anchor condition, secondary responding increases. This is supported by previous research (see Neuhoﬀ, 2001), and serves as a type of ecological survival mechanism. When one perceives a looming body, it is important to determine quickly whether it poses an immediate threat or not. Conversely, when a body is receding, it can be assumed that it is not a threat. Static stimuli should be analysed for threat, but the time course for making that assessment is not as critical, and should vary directly with proximity.

Task Demands

Considering the effects across experiments can also speak to a potential response bias argument for the results. In Experiment 4, participants completed both VAV and AVA conditions, in a counterbalanced order. The difference between the two paradigms, and the implied causality relationship between audition and vision (Guski & Troje, 2003) could lead to a need for participants to respond differently in the two paradigms, simply because they are different. However, in Experiments 1 and 2, participants completed only one of the two

paradigms. In this situation, it is much more likely that the observed effects would come from differences in the stimuli themselves (rather than processing configuration across paradigms).

An additional interesting finding related to the task is the overall mean response across the experiments. The timing of the roving stimulus that was employed was symmetrically distributed with regard to the two anchors, which should imply a baseline of 0.5 – equal probability of primary and secondary responding. However, when the mean proportion of secondary responding across Experiments 1, 2, and 4 was computed, it was found to be less than 0.5 ($M = .470$, $SE = .009$). A one-sample t -test to compare this value to the test value of 0.5 was significant, $t(80) = -3.038$, $p = .003$, which shows that indeed our baseline was significantly different from the expected mean of 0.5. Some light is shed on this seemingly anomalous result by Roseboom et al.'s (2009) finding that when multimodal binding decisions are made, they are hard to break. In our paradigm, this implies that there should be an advantage (or, a bias towards binding to) the primary anchor, and this is borne out in the data. As such, it will be important in future research to keep this tendency in mind, and ensure that any comparisons to baseline use the true baseline rather than 0.5.

Proposed Future Research

There are many ways to build upon the findings of this research, and some of them will be explored in this section. The most obvious would be to eschew the use of congruency as an analogue for distance and use real distance. This approach would increase the ecological validity of the findings, because the visual and auditory stimuli would truly be proximal and distal to the participant, rather than being simulated as such on one screen and pair of speakers. This would require an apparatus with displays and speaker arrays at various distances from the participant, but a successful replication would demonstrate that the effects shown are based on distance (and

allow for the use of simulated distance in future studies, as necessary). A failed replication, however, would point to a semantic congruency effect driving the data, and not a simulated effect of distance (perceptual congruency; see Spence (2011) for a discussion of different levels of congruency). In that case, it would be useful to eliminate the effects of distance by modifying the stimuli. The perceived motion effect (which is the only effect of distance found) is due to the fact that the primary and secondary anchors are of differing magnitudes but of the same token. That is, for visual looming, a small star is followed by a large star in quick succession, leading to implied visual motion. By changing the type of shape (e.g. square and circle rather than star and star), this implied motion would be eliminated, and any congruency would assumedly be cross-modal between the magnitude of the roving auditory stimulus and one of the anchors.

Furthermore, it would be interesting to remove the idea of magnitude altogether, and present auditory and visual stimuli that are all the same volume (auditory) and size (visual), enabling the exploration of other congruency relationships. One option would be to use an amorphous shape (similar to a bouba) and a sharp shape (kiki; Ramachandran & Hubbard, 2001; Hossain, unpublished), and pair them with a sinusoidal sound wave (congruent with bouba) and a square or sawtooth wave (congruent with kiki). These shapes were established as having some level of synaesthetic congruency with astounding consistency, even cross-culturally. By employing the same overall paradigm, but using contour congruency rather than magnitude congruency, we would be able to show whether certain types of congruency are more or less effective in making a causality judgement than others. In the current research, magnitude was used to simulate distance, and as such it is unclear whether the non-temporal effects observed were limited to spatial congruency. Performing the above proposed experiments would determine whether the

effects we are observing are being driven by distance effects, magnitude congruency, or whether any type of congruency can reveal the same effects.

Concluding Remarks

The research takes the concepts of temporal and congruency factors that have been employed by previous researchers, and is the first research of its type to put these factors in a competitive binding scenario with one another. In doing so, asymmetries have been revealed between the responding shown on the two paradigms used, and these asymmetries align themselves with classically held domains of modality dominance. Future research ideas have been suggested above, and should be conducted to further elucidate the findings of the current study. For the time being, however, suffice it to say that we have found a dynamic relationship between temporal and spatial factors, and this relationship is dependent on the amount of auditory and visual information that is available to the observer.

In determining unity between auditory and visual information, our findings indicate that we first use temporal factors, and rely on congruency factors only when the temporal information is not clear enough to provide us with a conclusion. Generally speaking, auditory information shows better temporal resolution, while visual information is better-used for spatial (and, in this case, congruency) information. Our experimental series presented the two modalities in identical paradigms, therefore equating the reliabilities of the auditory and visual information. As such, the observed asymmetries are more likely to be due to intrinsic properties of modality with regard to their utility in the given tasks. So, it seems, that temporal information is the primary means by which we make audiovisual binding decisions, and that we prefer to get that temporal information from an auditory source. We then use visually gleaned spatial information to

Congruency and temporal asynchrony in audio-visual perception

support the temporal information when it does not provide conclusive evidence for a binding decision to be made.

References

- Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current Biology*, 14 (3), 257-262.
- Alais, D., & Carlile, S. (2005). Synchronizing to real events: subjective audiovisual alignment scales with perceived auditory depth and speed of sound. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (6), 2244-2247.
- Arnold, D. H., Johnston, A., & Nishida, S. (2005). Timing sight and sound. *Vision Research*, 45, 1275-1284.
- Bermant, R. I., & Welch, R. B. (1976). Effect of degree of separation of visual-auditory stimulus and eye position upon spatial interaction of vision and audition. *Perceptual and Motor Skills*, 43, 487-493.
- Bien, N., ten Oever, S., Goebel, R., & Sack, A. T. (2012). The sound of size: crossmodal binding in pitch-size synesthesia: a combined TMS, EEG and psychophysics study. *NeuroImage*, 59 (1), 663-672.
- Burr, D., Banks, M. S., & Morrone, M. C. (2009). Auditory dominance over vision in the perception of interval duration. *Experimental Brain Research*, 198 (1), 49-57.
- Calvert, G. A., Spence, C., & Stein, B. E. (2004). *The handbook of multisensory processing*. Cambridge, MA: MIT Press.
- Cappe, C., Thut, G., Romei, V., & Murray, M. M. (2009). Selective integration of auditory-visual looming cues by humans. *Neuropsychologia*, 47 (4), 1045-1052.
- Cohen, J., MacWhinney, B., Flatt, M. & Provost, J. (1993). PsyScope: An interactive graphic system for designing and controlling experiments in the psychology laboratory using

- Macintosh computers. *Behavior Research Methods, Instruments, and Computers*, 25, 257-271.
- Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception*, 9, 719-721.
- Engel, G. R., & Dougherty, W. G. (1971). Visual-auditory distance constancy. *Nature*, 234, 308.
- Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, 10 (1), 1-12.
- Fujisaki, W., Shimojo, S., Kashino, S. & Nishida, S. (2004). Recalibration of audio-visual simultaneity. *Nature Neuroscience*, 7, 773-778.
- Fujisaki, W., Koene, A., Arnold, D., Johnston, A., & Nishida, S. (2006). Visual search for a target changing in synchrony with an auditory signal. *Proceedings of the Royal Society B*, 273, 865-874.
- Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & Psychophysics*, 68 (7), 1191-1203.
- Guski, R., & Troje N. F. (2003). Audiovisual phenomenal causality. *Perception & Psychophysics*, 65, 789-800.
- Heron, J., Whitaker, D., McGraw, P. V., & Horoshenkov, K. V. (2007). Adaptation minimizes distance-related audiovisual delays. *Journal of Vision*, 7 (13):5, 1-8.
- Hossain, S. (unpublished). Shapes and sounds: an exploration of audiovisual crossmodality. *Unpublished manuscript*.
- Howard, I. P., & Templeton, W. B. (1966). *Human Spatial Orientation*. Oxford, England: John Wiley & Sons.

- Kanaya, S., & Yokosawa, K. (2011). Perceptual congruency of audio-visual speech affects ventriloquism with bilateral visual stimuli. *Psychonomic Bulletin and Review*, 18(1), 123-128.
- Koelewijn, T. Bronkhorst, A. & Theeuwes, J. (2010). Attention and the multiple stages of multisensory integration: A review of audio-visual studies. *Acta Psychologica*, 134, 372-384.
- Kohlrausch, A., & van de Par, S. (2000). Experimente zur Wahrnehmbarkeit von Asynchronie in audio-visuellen Stimuli [Experiments on the perception of asynchrony with audio-visual stimuli]. In *Fortschritte der Akustik* (DAGA 2000, pp. 316-317). Oldenburg: DEGA Gesellschaft.
- Kopinska, A. & Harris, L. R. (2004). Simultaneity constancy. *Perception*, 33, 1049-1060.
- Lewald, J., & Guski, R. (2003). Cross-modal perceptual integration of spatially and temporally disparate auditory and visual stimuli. *Cognitive Brain Research*, 16 (3), 468-478.
- Lewald, J., & Guski, R. (2004). Auditory-visual temporal integration as a function of distance: no compensation for sound-transmission time in human perception. *Neuroscience Letters*, 357, 119-122.
- Ludwig, V. U., Adachi, I., & Matsuzawa, T. (2011). Visuoauditory mappings between high luminance and high pitch are shared by chimpanzees (*Pan troglodytes*) and humans. *Proceedings of the National Academy of Sciences* 108 (51), 20661-20665.
- Maier, J. X., Neuhoff, J. G., Logothetis, N. K., & Ghazanfar, A. A. (2004). Multisensory integration of looming signals by Rhesus monkeys. *Neuron*, 43, 177-181.
- Marks, L. E. (1987). On cross-modal similarity: Auditory-visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, 13, 384-394.

- McGurk, H., & MacDonald J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748.
- Melara, R. D., & O'Brien, T. P. (1987). Interaction between synesthetically corresponding dimensions. *Journal of Experimental Psychology: General*, 116 (4), 323-336.
- Michotte, A. (1946). *La perception de la causalité*. Louvain: Publications Universitaires.
[English translation: *The perception of causality*. London: Methuen, 1963.]
- Molholm, S., Ritter, W., Javitt, D. C. & Foxe, J. J. (2004). Multisensory visual-auditory object recognition in humans: A high-density electrical mapping study. *Cerebral Cortex*, 14, 452-465.
- Neuhoff, J. G. (2001). An adaptive bias in the perception of looming auditory motion. *Ecological Psychology*, 13 (2), 87-110.
- Olsen, K. N. & Stevens, C. J. (2010). Perceptual overestimation of rising intensity: Is stimulus continuity necessary? *Perception*, 39, 695-704.
- Parise, C. V. & Spence, C. (2009). 'When birds of a feather flock together': Synesthetic correspondences modulate audio-visual integration in non-synesthetes. *PLoS One*, 4, e5664.
- Patching, G. R., & Quinlan, P. T. (2002). Garner and congruence effects in the speeded classification of bimodal signals. *Journal of Experimental Psychology: Human Perception and Performance*, 28 (4), 755-775.
- Radeau, M. (1985). Signal intensity, task context, and auditory-visual interactions. *Perception*, 14(5), 571-577.
- Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia – a window into perception, thought and language. *Journal of Consciousness Studies*, 8 (12), 3-34.
- Recanzone, G. H. (2003). Auditory influences on visual temporal rate perception. *Journal of Neurophysiology*, 89, 1078-1093.

- Roach, N. W., Heron, J., & McGraw, P. V. (2006). Resolving multisensory conflict: a strategy for balancing the costs and benefits of audio-visual integration. *Proceedings of the Royal Society B*, 273, 2159-2168.
- Roseboom, W., Nishida, S. & Arnold, D. H. (2009). The sliding window of audio-visual simultaneity. *Journal of Vision*, 9, 1-8.
- Sandhu, R., & Dyson, B. J. (2012). Re-evaluating visual and auditory dominance through modality switching costs and congruency analyses. *Acta Psychologica*, 140 (2), 111-118.
- Scholl, B. J., & Nakayama, K. (2000, November). *Contextual effects of the perception of causality*. Poster presented at the annual meeting of the Psychonomic Society, New Orleans. [Abstract published in *Abstracts of the Psychonomic Society*, 5, 91].
- Slutsky, D. A., & Recanzone, G. H. (2001). Temporal and spatial dependency of the ventriloquism effect. *NeuroReport*, 12 (1), 7-10.
- Smith, L. B., & Sera, M. D. (1992). A developmental analysis of the polar structure of dimensions. *Cognitive Psychology*, 24 (1), 99-142.
- Soetens, E. Maetens, K., Zeischka, P. & Henderickx, D. (2010). Congruency reversals in an accessory signal Simon task with auditory and visual stimuli. *Acta Psychologica*, 134, 391-397.
- Soto-Faracao, S. and Alsius, A. (2009). Deconstructing the McGurk-MacDonald illusion. *Journal of Experimental Psychology: Human Perception and Performance*, 35, 580-587.
- Spence, C. (2011). Crossmodal correspondences: a tutorial review. *Attention, Perception, and Psychophysics*, 73 (4), 971-995.
- Spence, C. & Squire, S. (2003). Multisensory integration: maintaining the perception of synchrony. *Current Biology*, 13, R519-R521.

- Stein, B. E., & Stanford T. R. (2008). Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience*, 9, 255-266.
- Stone, J. V., Hunkin, N. M., Porrill, J., Wood, R., Keeler, V., Beanland, M., Port, M., & Porter, N. R. (2001). When is now? Perception of simultaneity. *Proceedings of the Royal Society B*, 268, 31-38.
- Sugita, Y., & Suzuki, Y. (2003). Audiovisual perception: implicit estimation of sound-arrival time. *Nature*, 421, 911.
- Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America*, 26, 212–215.
- Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in Cognitive Science*, 14 (9), 400-410.
- Teder-Sälejärvi, W. A., McDonald, J. J., Di Russo, F., Hillyard, S. A. (2002). An analysis of audio-visual crossmodal integration by means of event-related potential (ERP) recordings. *Cognitive Brain Research*, 14, 106-114.
- van Wassenhove, V., Grant, K. W. & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neurophysiologica*, 45, 598-607.
- Vroomen, J., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Recalibration of temporal order perception by exposure to audiovisual asynchrony. *Cognitive Brain Research*, 22 (1), 32-35.
- Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: a tutorial review. *Attention, Perception, & Psychophysics*, 72 (4), 871-884.

- Walker, P., Bremner, J. G., Mason, U., Spring, J., Matlock, K., Slater, A. & Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science*, 21, 21-25.
- Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space, and quantity. *Trends in Cognitive Sciences*, 7 (11), 483-488.
- Welch, R. B., & Warren, D. H. (1980). Immediate perceptual response to intersensory discrepancy. *Psychological Bulletin*, 88, 638-667.
- Zhu, X. R., Zhang, H. J., Wu, T. T., Luo, W. B. & Luo, Y. J. (2010). Emotional conflict occurs at an early stage: Evidence from the emotional face-word Stroop task. *Neuroscience Letters*, 478, 1-4.
- Zmigrod, S., & Hommel, B. (2011). The relationship between feature binding and consciousness: evidence from asynchronous multi-modal stimuli. *Consciousness and Cognition*, available online first: doi: 10.1016/j.concog.2011.01.011.

Table 1 **Probability of second responding (and SE) for 8 conditions of Experiment 1**

<i>Time from V1 (ms)</i>	<i>0</i>	<i>50</i>	<i>100</i>	<i>150</i>	<i>200</i>	<i>250</i>	<i>300</i>	<i>350</i>	<i>400</i>
<i>V1 / A / V2</i>									
<i>Small/Quiet/Small</i>	<i>.17 (.03)</i>	<i>.18 (.03)</i>	<i>.22 (.04)</i>	<i>.29 (.03)</i>	<i>.47 (.04)</i>	<i>.59 (.04)</i>	<i>.65 (.04)</i>	<i>.75 (.04)</i>	<i>.77 (.04)</i>
<i>Small/Quiet/Large</i>	<i>.19 (.04)</i>	<i>.20 (.04)</i>	<i>.19 (.03)</i>	<i>.26 (.03)</i>	<i>.50 (.04)</i>	<i>.64 (.04)</i>	<i>.76 (.04)</i>	<i>.80 (.04)</i>	<i>.82 (.04)</i>
<i>Small/Loud/Small</i>	<i>.14 (.03)</i>	<i>.18 (.03)</i>	<i>.22 (.04)</i>	<i>.22 (.02)</i>	<i>.37 (.03)</i>	<i>.56 (.04)</i>	<i>.69 (.04)</i>	<i>.73 (.04)</i>	<i>.78 (.04)</i>
<i>Small/Loud/Large</i>	<i>.14 (.03)</i>	<i>.19 (.04)</i>	<i>.25 (.04)</i>	<i>.28 (.03)</i>	<i>.49 (.04)</i>	<i>.66 (.04)</i>	<i>.78 (.04)</i>	<i>.74 (.04)</i>	<i>.81 (.04)</i>
<i>Large/Quiet/Small</i>	<i>.14 (.03)</i>	<i>.12 (.04)</i>	<i>.18 (.04)</i>	<i>.23 (.04)</i>	<i>.37 (.04)</i>	<i>.55 (.04)</i>	<i>.69 (.05)</i>	<i>.72 (.04)</i>	<i>.81 (.04)</i>
<i>Large/Quiet/Large</i>	<i>.17 (.04)</i>	<i>.18 (.04)</i>	<i>.20 (.04)</i>	<i>.24 (.03)</i>	<i>.38 (.03)</i>	<i>.56 (.04)</i>	<i>.65 (.04)</i>	<i>.70 (.04)</i>	<i>.78 (.04)</i>
<i>Large/Loud/Small</i>	<i>.12 (.04)</i>	<i>.11 (.03)</i>	<i>.16 (.04)</i>	<i>.20 (.04)</i>	<i>.35 (.04)</i>	<i>.51 (.05)</i>	<i>.70 (.04)</i>	<i>.78 (.05)</i>	<i>.79 (.04)</i>
<i>Large/Loud/Large</i>	<i>.13 (.03)</i>	<i>.18 (.04)</i>	<i>.18 (.04)</i>	<i>.21 (.04)</i>	<i>.34 (.04)</i>	<i>.54 (.04)</i>	<i>.73 (.04)</i>	<i>.70 (.05)</i>	<i>.78 (.04)</i>

Table 2 Summary of repeated measures ANOVA of Experiment 1 data

<i>Metric</i>	<i>df</i>	<i>F</i>	<i>MSE</i>	<i>p</i>
1 st Visual (V1)	1,26	3.28	.258	=.082
Audio (A)	1,26	1.87	.028	=.184
2nd Visual (V2)	1,26	5.58	.045	=.026
Time (T)	8,208	106.56	.142	<.001
V1 x A	1,26	0.11	.012	=.747
V1 x V2	1,26	1.11	.131	=.302
V1 x T	8,208	3.56	.012	=.001
A x V2	1,26	2.65	.007	=.116
A x T	8,208	2.93	.009	=.004
V2 x T	8,208	1.94	.012	=.056
V1 x A x V2	1,26	1.36	.011	=.254
V1 x A x T	8,208	1.43	.010	=.187
V1 x V2 x T	8,208	3.02	.012	=.003
A x V2 x T	8,208	1.16	.010	=.324
V1 x A x V2 x T	8,208	1.40	.010	=.199

Note: Statistical significance in bold

Table 3 Probability of second responding (and SE) for 8 conditions of Experiment 2

<i>Time from A1 (ms)</i>	<i>0</i>	<i>50</i>	<i>100</i>	<i>150</i>	<i>200</i>	<i>250</i>	<i>300</i>	<i>350</i>	<i>400</i>
<i>A1 / V / A2</i>									
<i>Quiet/Small/Quiet</i>	<i>.17 (.03)</i>	<i>.21 (.04)</i>	<i>.22 (.03)</i>	<i>.38 (.04)</i>	<i>.53 (.04)</i>	<i>.73 (.04)</i>	<i>.82 (.04)</i>	<i>.82 (.04)</i>	<i>.86 (.03)</i>
<i>Quiet/Small/Loud</i>	<i>.16 (.03)</i>	<i>.18 (.03)</i>	<i>.23 (.04)</i>	<i>.38 (.04)</i>	<i>.60 (.04)</i>	<i>.76 (.03)</i>	<i>.78 (.03)</i>	<i>.82 (.04)</i>	<i>.84 (.03)</i>
<i>Quiet/Large/Quiet</i>	<i>.16 (.03)</i>	<i>.17 (.04)</i>	<i>.22 (.03)</i>	<i>.30 (.03)</i>	<i>.53 (.04)</i>	<i>.72 (.04)</i>	<i>.80 (.04)</i>	<i>.86 (.03)</i>	<i>.85 (.03)</i>
<i>Quiet/Large/Loud</i>	<i>.18 (.04)</i>	<i>.13 (.03)</i>	<i>.22 (.04)</i>	<i>.32 (.04)</i>	<i>.55 (.04)</i>	<i>.70 (.04)</i>	<i>.80 (.04)</i>	<i>.82 (.04)</i>	<i>.83 (.04)</i>
<i>Loud/Small/Quiet</i>	<i>.17 (.03)</i>	<i>.20 (.04)</i>	<i>.21 (.04)</i>	<i>.33 (.04)</i>	<i>.51 (.04)</i>	<i>.69 (.03)</i>	<i>.75 (.04)</i>	<i>.83 (.03)</i>	<i>.84 (.03)</i>
<i>Loud/Small/Loud</i>	<i>.19 (.04)</i>	<i>.17 (.04)</i>	<i>.21 (.04)</i>	<i>.38 (.04)</i>	<i>.55 (.03)</i>	<i>.67 (.04)</i>	<i>.80 (.04)</i>	<i>.82 (.03)</i>	<i>.85 (.03)</i>
<i>Loud/Large/Quiet</i>	<i>.12 (.03)</i>	<i>.15 (.03)</i>	<i>.19 (.04)</i>	<i>.26 (.03)</i>	<i>.48 (.04)</i>	<i>.67 (.04)</i>	<i>.74 (.04)</i>	<i>.80 (.04)</i>	<i>.83 (.03)</i>
<i>Loud/Large/Loud</i>	<i>.13 (.03)</i>	<i>.13 (.03)</i>	<i>.15 (.03)</i>	<i>.27 (.03)</i>	<i>.44 (.04)</i>	<i>.64 (.04)</i>	<i>.76 (.04)</i>	<i>.80 (.03)</i>	<i>.82 (.03)</i>

Table 4 Summary of repeated measures ANOVA of Experiment 2 data

<i>Metric</i>	<i>df</i>	<i>F</i>	<i>MSE</i>	<i>p</i>
1st Audio (A1)	1,26	33.78	.014	<.001
Visual (V)	1,26	8.52	.054	=.007
2 nd Audio (A2)	1,26	0.02	.018	=.886
Time (T)	8,208	139.08	.130	<.001
A1 x V	1,26	7.35	.012	=.012
A1 x A2	1,26	0.13	.015	=.722
A1 x T	8,208	2.27	.008	=.024
V x A2	1,26	2.28	.010	=.143
V x T	8,208	2.42	.011	=.016
A2 x T	8,208	1.21	.011	=.296
A1 x V x A2	1,26	0.47	.012	=.501
A1 x V x T	8,208	0.73	.009	=.663
A1 x A2 x T	8,208	1.32	.009	=.237
V x A2 x T	8,208	0.74	.010	=.659
A1 x V x A2 x T	8,208	0.60	.010	=.775

Note: Statistical significance in bold

Table 5 **Reaction times (ms) (and SE) for each condition of Experiment 3**

<i>Discrimination Level</i>	<i>Auditory Responding</i>		<i>Visual Responding</i>	
	<i>Congruent</i>	<i>Incongruent</i>	<i>Congruent</i>	<i>Incongruent</i>
<i>Visual / Auditory</i>				
<i>Hard / Hard</i>	713 (51)	755 (56)	705 (32)	717 (33)
<i>Hard / Moderate</i>	687 (41)	709 (39)	686 (34)	710 (29)
<i>Hard / Easy</i>	671 (30)	711 (30)	672 (26)	703 (26)
<i>Moderate / Hard</i>	672 (33)	698 (32)	642 (26)	658 (25)
<i>Moderate / Moderate</i>	730 (46)	767 (48)	703 (39)	719 (30)
<i>Moderate / Easy</i>	683 (34)	703 (29)	668 (36)	720 (37)
<i>Easy / Hard</i>	684 (32)	715 (36)	661 (27)	674 (26)
<i>Easy / Moderate</i>	699 (31)	703 (31)	685 (31)	707 (28)
<i>Easy / Easy</i>	731 (49)	749 (40)	662 (40)	719 (35)

Table 6 Error rates (and SE) for each condition of Experiment 3

<i>Discrimination Level</i>	<i>Auditory Responding</i>		<i>Visual Responding</i>	
	<i>Congruent</i>	<i>Incongruent</i>	<i>Congruent</i>	<i>Incongruent</i>
<i>Visual / Auditory</i>				
<i>Hard / Hard</i>	.251 (.034)	.238 (.035)	.065 (.012)	.097 (.014)
<i>Hard / Moderate</i>	.144 (.029)	.196 (.029)	.054 (.011)	.094 (.014)
<i>Hard / Easy</i>	.144 (.025)	.170 (.026)	.058 (.014)	.104 (.016)
<i>Moderate / Hard</i>	.229 (.023)	.266 (.035)	.059 (.014)	.075 (.011)
<i>Moderate / Moderate</i>	.155 (.024)	.209 (.030)	.044 (.010)	.087 (.013)
<i>Moderate / Easy</i>	.132 (.024)	.147 (.021)	.059 (.012)	.084 (.013)
<i>Easy / Hard</i>	.257 (.026)	.278 (.034)	.043 (.009)	.073 (.012)
<i>Easy / Moderate</i>	.166 (.026)	.199 (.032)	.052 (.011)	.098 (.012)
<i>Easy / Easy</i>	.156 (.023)	.194 (.034)	.061 (.016)	.110 (.016)

Table 7 Summary of repeated measures ANOVA of RT data of Experiment 3

<i>Metric</i>	<i>df</i>	<i>F</i>	<i>MSE</i>	<i>p</i>
Visual (V)	2,70	0.06	82818.5	=.947
Audio (A)	2,70	0.28	124143.7	=.762
Modality (M)	1,35	1.30	104554.7	=.263
Congruency (C)	1,35	21.85	10636.8	<.001
V x A	4,140	3.00	31202.5	=.021
V x M	2,70	0.60	19524.1	=.552
V x C	2,70	0.10	5214.5	=.902
A x M	2,70	0.52	15048.0	=.595
A x C	2,70	1.32	5668.4	=.275
M x C	1,35	0.00	5236.7	=.958
V x A x M	4,140	1.78	8995.0	=.136
V x A x C	4,140	0.13	5993.9	=.971
V x M x C	2,70	1.09	3945.0	=.341
A x M x C	2,70	2.03	10866.3	=.138
V x A x M x C	4,140	0.74	4001.5	=.565

Note: Statistical significance in bold

Table 8 Summary of repeated measures ANOVA of error rate data of Experiment 3

<i>Metric</i>	<i>df</i>	<i>F</i>	<i>MSE</i>	<i>p</i>
Visual (V)	2,70	0.56	.026	=.576
Audio (A)	2,70	12.61	.018	<.001
Modality (M)	1,35	40.93	.120	<.001
Congruency (C)	1,35	24.08	.015	<.001
V x A	4,140	0.63	.009	=.642
V x M	2,70	1.09	.014	=.342
V x C	2,70	0.19	.005	=.831
A x M	2,70	23.09	.014	<.001
A x C	2,70	3.02	.005	=.055
M x C	1,35	0.34	.012	=.566
V x A x M	4,140	0.92	.007	=.456
V x A x C	4,140	0.92	.004	=.457
V x M x C	2,70	0.71	.006	=.495
A x M x C	2,70	0.42	.006	=.658
V x A x M x C	4,140	0.83	.005	=.508

Note: Statistical significance in bold

Table 9 **Probability of second responding (and SE) by condition in Experiment 4**

<i>Time from Stim 1 (ms)</i>	<i>0</i>	<i>100</i>	<i>200</i>	<i>300</i>	<i>400</i>
VAV					
<i>Low/Low/Low</i>	.19 (.03)	.24 (.04)	.41 (.05)	.57 (.06)	.62 (.05)
<i>Low/Low/High</i>	.19 (.04)	.19 (.03)	.48 (.05)	.66 (.05)	.72 (.05)
<i>Low/High/Low</i>	.15 (.03)	.22 (.03)	.42 (.05)	.57 (.05)	.65 (.05)
<i>Low/High/High</i>	.14 (.03)	.23 (.04)	.47 (.05)	.68 (.06)	.72 (.05)
<i>High/Low/Low</i>	.14 (.02)	.16 (.03)	.27 (.04)	.53 (.05)	.67 (.06)
<i>High/Low/High</i>	.16 (.03)	.17 (.03)	.35 (.05)	.52 (.05)	.64 (.05)
<i>High/High/Low</i>	.13 (.03)	.16 (.03)	.27 (.04)	.51 (.05)	.66 (.06)
<i>High/High/High</i>	.14 (.03)	.16 (.03)	.33 (.04)	.54 (.05)	.66 (.06)
AVA					
<i>Low/Low/Low</i>	.16 (.03)	.21 (.03)	.47 (.04)	.69 (.05)	.75 (.05)
<i>Low/Low/High</i>	.15 (.02)	.21 (.03)	.54 (.05)	.71 (.05)	.76 (.06)
<i>Low/High/Low</i>	.08 (.02)	.19 (.03)	.53 (.05)	.74 (.05)	.74 (.06)
<i>Low/High/High</i>	.12 (.02)	.16 (.02)	.56 (.05)	.73 (.05)	.78 (.06)
<i>High/Low/Low</i>	.13 (.02)	.16 (.03)	.43 (.04)	.66 (.05)	.76 (.05)
<i>High/Low/High</i>	.11 (.02)	.19 (.04)	.38 (.04)	.70 (.05)	.77 (.05)
<i>High/High/Low</i>	.10 (.03)	.13 (.02)	.39 (.04)	.69 (.05)	.75 (.05)
<i>High/High/High</i>	.10 (.02)	.16 (.03)	.35 (.04)	.72 (.05)	.79 (.06)

Table 10 **Summary of repeated measures ANOVA of Experiment 4**

<i>Metric</i>	<i>df</i>	<i>F</i>	<i>MSE</i>	<i>p</i>
Condition (C)	1,26	11.68	.151	=.002
1st Stimulus (S1)	1,26	46.21	.044	<.001
2 nd Stimulus (S2)	1,26	0.39	.033	=.539
3rd Stimulus (S3)	1,26	11.09	.028	=.003
Time (T)	4,104	194.84	.185	<.001
C x S1	1,26	2.56	.052	=.121
C x S2	1,26	0.09	.023	=.770
C x S3	1,26	2.67	.024	=.114
C x T	4,104	15.07	.048	<.001
S1 x S2	1,26	0.51	.011	=.482
S1 x S3	1,26	0.96	.049	=.336
S1 x T	4,104	19.31	.019	<.001
S2 x S3	1,26	0.11	.021	=.744
S2 x T	4,104	3.67	.015	=.008
S3 x T	4,104	3.16	.014	=.017
C x S1 x S2	1,26	0.23	.018	=.639
C x S1 x S3	1,26	0.17	.054	=.685
C x S1 x T	4,104	1.39	.016	=.244

Congruency and temporal asynchrony in audio-visual perception

C x S2 x S3	1,26	0.03	.013	=.862
C x S2 x T	4,104	1.35	.011	=.258
C x S3 x T	4,104	1.48	.016	=.215
S1 x S2 x S3	1,26	0.16	.017	=.695
S1 x S2 x T	4,104	2.26	.013	=.068
S1 x S3 x T	4,104	3.02	.012	=.021
S2 x S3 x T	4,104	0.29	.014	=.883
C x S1 x S2 x S3	1,26	0.28	.011	=.605
C x S1 x S2 x T	4,104	0.57	.017	=.687
C x S1 x S3 x T	4,104	4.40	.018	=.003
C x S2 x S3 x T	4,104	0.91	.014	=.459
S1 x S2 x S3 x T	4,104	0.42	.013	=.794
C x S1 x S2 x S3 x T	4,104	0.51	.013	=.728

Note: Statistical significance in bold

Figure 1. Experimental stimulus presentations for Experiments 1, 2, and 4. First stimulus (S1) and third stimulus (S3) are always presented from 0-100 ms and 300-400 ms, respectively. A 100 ms long second stimulus (S2) is presented with its onset at one of nine potential times (top scale) for Experiments 1-2, and five potential times (bottom scale) in Experiment 4.

Figure 2. Comparison of responding over time between Experiment 1 (VAV) and Experiment 2 (AVA). Error bars show standard error.

Figure 3. Graph showing significant interaction between V1 and time in Experiment 1. Error bars show standard error, and significant differences (Tukey's HSD; $p < .05$) are indicated by asterisks.

Figure 4. Graph showing significant three-way interaction between V1 x V2 x time in Experiment 1. Error bars show standard error, and significant differences (Tukey's HSD; $p < .05$) are indicated by asterisks.

Figure 5. Graph showing significant interaction between A1 and V in Experiment 2. Error bars show standard error, and significant differences (Tukey's HSD; $p < .05$) are indicated by asterisks.

Figure 6. Graph showing significant interaction between A1 and time in Experiment 2. Error bars show standard error, and significant differences (Tukey's HSD; $p < .05$) are indicated by asterisks.

Figure 7. Graph showing significant interaction between V and time in Experiment 2. Error bars show standard error, and significant differences (Tukey's HSD; $p < .05$) are indicated by asterisks.

Figure 8. Condition sets used in Experiment 3. All sets used V1 and A1 (96 point font, 71 dB), and the other three levels of audio and visual stimuli were paired in each possible combination, yielding nine total combinations, as shown below.

Standard (V1 – A1) 96 point - 71 dB	V2 48 point	V3 24 point	V4 12 point
A2 66 dB	Hard Visual Hard Auditory	Moderate Visual Hard Auditory	Easy Visual Hard Auditory
A3 61 dB	Hard Visual Moderate Auditory	Moderate Visual Moderate Auditory	Easy Visual Moderate Auditory
A4 56 dB	Hard Visual Easy Auditory	Moderate Visual Easy Auditory	Easy Visual Easy Auditory

Figure 9. Graph showing significant interaction for reaction times between auditory discrimination difficulty and visual discrimination difficulty for Experiment 3. Error bars show standard error.

Figure 10. Graph showing significant interaction between auditory difficulty and modality for error rates in Experiment 3. Error bars show standard error.

Figure 11. Graphs showing congruency effects (blue graphs) and audio-visual differences (red graphs) for reaction times (first pair) and error rates (second pair) for 9 possible combinations of audio-visual stimuli sets in Experiment 3. Error bars show standard error. Negative values indicate higher reaction time or error rate for visual response. Red arrows show the condition set used in Experiments 1 and 2, and green arrows show the condition set chosen for use in Experiment 4.

Congruency and temporal asynchrony in audio-visual perception

Figure 12. Graph showing significant interaction between experimental condition and time in Experiment 4. Error bars show standard error, and significant differences (Tukey's HSD; $p < .05$) are indicated by asterisks.

Figure 13. Graph showing significant interaction between S1 and time in Experiment 4. Error bars show standard error, and significant differences (Tukey's HSD; $p < .05$) are indicated by asterisks.

Figure 14. Graph showing significant interaction between S3 and time in Experiment 4. Error bars show standard error, and significant differences (Tukey's HSD; $p < .05$) are indicated by asterisks.

Figure 15. Graph showing significant four-way interaction between condition x S1 x S3 x time in Experiment 4. Error bars show standard error.

Appendix A – Participant Consent Agreement

Ryerson University

Consent Agreement

AUDIO-VISUAL BINDING IN ADULTS: EFFECTS OF SIZE AND VOLUME OF OBJECTS

You are being asked to participate in a research study. Before you give your consent to be a volunteer, it is important that you read the following information and ask as many questions as necessary to be sure you understand what you will be asked to do.

Investigators: Jonathan Wilbiks, Ben Dyson

Purpose of the Study: This study is part of an ongoing research program where we hope to more fully understand the way in which the brain processes auditory and visual information, how information from different senses interact, and how those processes and interactions change as a function of age and expertise. We are hoping to test 24 individuals in this study, and wish to use only those individuals who self-report as having normal (or corrected-to-normal) hearing and vision.

Description of the Study: The study will take place in the HEAR Lab, located in the Psychology Research and Training Centre at 105 Bond Street, unless otherwise stated. Experiments will take approximately 1 hour so *please ask your experimenter now if you are unclear as to the time commitments of the current study*. Prior to the study, you will have the study explained to you and the opportunity to take part in a practice block so you are familiar with the procedure. You will be given the chance to ask any questions you may have regarding the study, prior to reviewing the consent agreement. During the study, age, gender, handedness and degree of musical expertise (by educational level and years of playing) will be requested. After the study, you will be fully debriefed as to the purpose of the study, and given a further opportunity to ask questions.

In both practice and experimental blocks, at each trial, you will see an image displayed on the screen, followed by a sound, and then see a second image. You will then be asked whether the tone seemed to come from the first or second visual stimulus. The sound may be loud or quiet, and the images may be small or large. Please try to respond as quickly and as accurately as possible. You will not receive feedback as to whether your response was right or wrong.

What is Experimental in this Study: Previous research has examined how auditory and visual information integrate with one another across time. The study is experimental in the respect that we are investigating auditory and visual integration using a unique design. The study is also experimental in the respect that we manipulate variables like size and volume across trials, present these trials in a random order to participants.

Risks or Discomforts: There are no known long-term risks associated with behavioral testing of the manner proposed. One short-term risk is fatigue. Effects of fatigue will be offset by providing participants with the opportunity to take breaks in-between blocks of trials. *If you feel uncomfortable at any time during the experiment, you may discontinue participation, either temporarily or permanently.*

Benefits of the Study: The potential benefits of the study for science and society are a greater understanding of how the cognitive processing of stimuli occurs. The studies may also offer avenues into how to tailor more aesthetically pleasing experiences as a result of understanding how the senses interact. However, there are no immediate benefits that you can reasonably expect from the study.

Confidentiality: Confidentiality will be maintained in all aspects of data dissemination. Only individuals involved in the research team will have access to a central password-protected electronic file that matches participant ID numbers with identifying information, and to original research records which will be stored in a locked file cabinet. All data will be stored for a minimum of 5 years after collection. Participants have the option of receiving a summary of their performance after participation, and should make this known to the researchers at the time of testing. Participants also have the option of removing their data from the study after participation, and should contact an individual listed on the debriefing form to do so. Please note that after publication of a data set (usually no sooner than 3 months after participation) it is not possible to remove data.

Incentives to Participate: Three different incentive schemes are offered. Payment can be offered for experimental participation at the rate of \$10 per hour. If either the participant or experimenter chooses to discontinue the study, participants will be paid \$5 pro-rata. For certain individuals, course credit is available as an incentive, awarded either on the basis of participation or a walk-through in which the participant can take part in the study but not submit their data. Please indicate which incentive you require:

☐

MONEY

☐

COURSE CREDIT
(PARTICIPATION)

☐

COURSE CREDIT
(WALKTHROUGH)

Voluntary Nature of Participation: Participation in this study is voluntary. Your choice of whether or not to participate will not influence your future relations with Ryerson University. If you decide to participate, you are free to withdraw your consent and to stop your participation at any time without penalty or loss of benefits to which you are allowed. At any particular point in the study, you may refuse to answer any particular question or stop participation altogether.

Questions about the Study: If you have any questions about the research now, please ask. If you have questions later about the research, you may contact.

Jonathan Wilbiks
jwilbiks@psych.ryerson.ca
001 416-979-5000 x2186

Congruency and temporal asynchrony in audio-visual perception

Dr. Ben Dyson
ben.dyson@psych.ryerson.ca
001 416-979-5000 x2063

If you have questions regarding your rights as a human subject and participant in this study, you may contact the Ryerson University Research Ethics Board for information.

Research Ethics Board
c/o Office of the Vice President, Research and Innovation
Ryerson University, 350 Victoria Street
Toronto, ON, M5B 2K3, Canada
001 416-979-5042

Agreement: Your signature below indicates that you have read the information in this agreement and have had a chance to ask any questions you have about the study. Your signature also indicates that you agree to be in the study and have been told that you can change your mind and withdraw your consent to participate at any time. You have been given a copy of this agreement. You have been told that by signing this consent agreement you are not giving up any of your legal rights.

Informed consent for study participation

Name of Participant (please print)

Signature of Participant

Date

Signature of Investigator

Date

Appendix B – Debriefing Form

Ryerson University

Debriefing Form

Audio-Visual Binding in Adults: Effects of Size and Volume of Objects

Dear Participant:

Thank you very much for your participation in our study. Your time and commitment to psychological research at Ryerson University is very much appreciated.

The study you took part in will contribute to ongoing auditory and visual research conducted in the H.E.A.R Lab. Our lab is dedicated to designing and implementing research studies that will help us better understand how the brain represents what we hear and see, and how this information is integrated.

The particular study you took part in was designed to assess how individuals learn to integrate information from the different senses. The timing of the sound varied from occurring simultaneously with the first visual stimulus to simultaneously with the second visual stimulus, and at various increments in between. We are interested in the way in which your perception of the source of the sound is influenced by the time of presentation, as well as things like the sizes of visual stimuli and the volume of the auditory stimulus. We expect that, although generally you will choose the visual that the sound occurs closest to, there will also be an effect of the size and volume, with louder sounds more likely to come from larger visuals, for example.

If you have any questions regarding your participation in this study, or would like to receive information about the results once they are available, feel free to contact Dr. Ben Dyson. We would be happy to provide you with your own data as well as the overall findings of our study.

Finally, if you are interested in taking part and learning more about visual and auditory perception research in the H.E.A.R Lab, feel free to contact Dr. Dyson.

Jonathan Wilbiks
MA Student
Ryerson University
jwilbiks@psych.ryerson.ca

Dr. Ben Dyson
Professor of Psychology
Ryerson University
ben.dyson@psych.ryerson.ca