Do Temporal Modulations in Timbre Impact Visual Perception

Emil Bergh

Submitted in partial fulfillment of the requirements for the
Master of Music in Music Technology
in the Department of Music and Performing Arts Professions
Steinhardt School
New York University

Primary Thesis Advisor: Prof. M. Farbood
Secondary Thesis Advisor: Prof. P. Ripolles
Thesis 2 - Section 01, MM Music Technology
Steinhardt, New York University
Prof. A. Roginska
Apr 26, 2023

ABSTRACT

Humans frequently resort to the use of semantics used by other sensory modalities, for example, vision and touch, when attempting to describe our perception of timbre. We have only just begun to understand whether these cross-modal descriptions of timbre reflect multisensory processing and to what extent linguistics or prior conceptual knowledge may influence these descriptions. If our perception of timbre does involve processing via more than one sensory dimension, could we use timbral characteristics to influence our perception of the correlating modalities? Furthermore, how would these cross-modal interactions occur in a temporal context - modulating from one state to another?

To investigate how temporal modulations in timbre impact visual perception, we conducted an experiment consisting of two parts. In both parts, the timbre of the auditory stimulus was modulated (changed over time), from bright to dark or dark to bright, over a 2-second period. In both parts of the experiment, the visual stimulus's brightness changed from bright to dark or dark to bright. The function of the audio stimuli was to either enhance or detract from response accuracy or reaction time to the visual stimuli. In part 1 of the experiment, auditory and visual stimuli were played in congruent and incongruent pairings to measure the impact of timbre on modulations in visual brightness. In part 2 of the experiment, congruent visual and auditory pairings were played with varying modulation intervals to measure the influence of changes in timbre on visual events in time.

In part 1, visual stimuli changing from bright to dark showed significant differences in accuracy between congruent and incongruent pairings; however, none of the other stimulus pairs showed significant results. While there were no significant differences in RT between congruent and incongruent pairings within each directional condition, there were significant differences observed in RT between the three brightness conditions, with stimuli modulating from dark to bright proving easier to identify.

In part 2, the auditory stimulus ended before, after, or at the same time as the visual stimulus. We observed significant results in both accuracy and RT between the three temporal conditions, indicating cross-modal interaction between visual and auditory senses when reacting to visual events. There were no significant differences observed in changes to directional brightness.

ACKNOWLEDGEMENTS

Kelly Mellecker without whom none of this would have been possible.

Morewaread Farbood, Pablo Ripolles, and Alex Sigman for their invaluable guidance and advice over the past two years as I discovered a whole new world of research.

Paul Geluso and Agnieszka Roginska for their inspiration and insight into the various exciting avenues within music technology.

Monique Pimontel, Abigail Greenleaf, and Faxin Zhou for equipping me with the tools I needed when I didn’t even know where to look for the toolbox.

Peter Traver, Darren Blanckensee, Jiawen Mao, Richa Namballa, Max Hein, and Tyler Kennedy for the conversations that inspired ideas, and the shared efforts that got us through the unknowns. Moreover, everyone else I was fortunate to cross paths with, who showed passion and excitement to learn and explore.

Last but certainly not least, Zachary Wallmark for setting the direction of this thesis with his research.

TABLE OF CONTENTS

LIST OF FIGURES

1. INTRODUCTION

Timbre can be defined as

"Any property other than pitch, duration, and loudness that allows two sounds to be distinguished." (McAdams, 2019, p. 1)

As humans, we frequently resort to the use of semantics used by other sensory modalities, e.g., vision and touch, when attempting to relate our perception of timbre. We have only just begun to understand whether these cross-modal descriptions of timbre reflect multisensory processing and to what extent linguistics or prior conceptual knowledge may influence these descriptions.

As we look to a future where our lifestyles are increasingly exposed to cross-platform, multimodal, immersive experiences - whether switching between handheld and desktop computing, augmented reality wearables, or fully immersive extended reality experiences - we are constantly exposed to a wide range of cross-modal stimuli. Understanding how these stimuli interact with one another is crucial to creating an effective ecosystem of digital applications and tools - where many humans spend most of their time interacting within both professional and personal contexts.

This thesis will investigate timbre's role in the cross-modal perception of visuals. We will explore this topic based on the hypothesis that temporal modulations in timbre will affect subjects' perception of visual brightness when subjected to incongruent stimulus pairings.

The research will consist of two experiments:

The first experiment will seek to establish if the brightness of a visual stimulus and task-irrelevant auditory prime are modulated in both congruent and incongruent directions (dark <> bright), what the effect of the prime-stimulus pairings would be on the participant's cross-modal perception.

The second experiment will seek to establish if the brightness of a visual stimulus and task-irrelevant auditory prime are modulated congruently but at synchronous and asynchronous rates, what the effect of the prime-stimulus pairings would be on the participant's cross-modal perception.

Our aim with these experiments will be:

- To establish how we might use one set of stimuli (e.g. a modulated sound) to influence the other (modulated brightness).

- To better understand the impact of temporality when examining the relationship between cross-modal sensory perception in the description of timbre.
- Establish whether changes (modulations) to incongruent audio-visual stimulus pairings affect participants' perception of said stimuli over a short period
- Establish whether asynchronous changes to congruent audio-visual stimulus pairings are identifiable when modulated over a short period.

Understanding the effect of temporal variables in cross-modal perception could have implications for our understanding of immersive auditory experiences and how we may use sound stimuli to manipulate timing and narrative.

While ample research has been done on the impact of pitch and amplitude as it relates to the perception of height, size, and brightness, we are just beginning to unpack the crossmodal perception of timbre. Where previous research has examined the impact of incongruent auditory stimuli on static visual prime stimuli, our study will add the independent variable of time to better understand these stimuli in an interactive context.

2. LITERATURE REVIEW

2.1 Introduction

Creating and conducting an experiment to measure the effect of audio-visual prime/stimulus pairings on cross-modal perception requires an understanding of three categories:

1. **Semantics** - Humans are generally considered to lack a sensory vocabulary to describe timbre. While tonal properties relating to pitch, duration, and loudness are described with linear scales ((low/high, slow/fast, soft/loud), we use attributes associated with other senses or modalities. In this section of our review, we will aim to establish a foundation of common descriptors as they relate specifically to timbre.
2. **Cross-modal Perception** - Our neurons share multiple modality pathways. Therefore cross-modal combinations of sensory stimuli are bound to either enhance or detract from the perception of one stimulus over the other when perceived simultaneously. This section reviews current research on various cross-modal pairings, for example, audio-tactile and audio-visual. Our focus will be on cross-modal interactions of timbre and visual perception - specifically, current gaps in this research topic relate to observations of cross-modal interactions when including temporality as an independent variable. Within the scope of audio-visual pairings, we will briefly examine research into cross-modal perceptions of pitch, duration, and loudness.
3. **Testing Methods** - Paramount to the practical interpretation of cross-modal auditory-visual interactions are the methods we use to capture and interpret the generated data. One of the most common effects used to measure choice behavior in mathematics and neuroscience is the Speed-Accuracy Trade-off (SAT). It provides an accurate view of how decision-speed and decision-accuracy interact. This section will examine the role the SAT has played in similar research topics and how this methodology might be applied to our specific area of interest.

2.2 Semantics

To start our investigation, we need to establish a foundation of commonly used terms that will ensure a shared understanding of the concepts discussed by both the researcher and the participant involved in the study.

Across the various sources covered in this review, one of the main recurring themes is that timbre appears to fall outside the standard tonal properties we use to describe a sound. Siedenburg et al. (2019) highlight some of the first psychophysical research into timbre's perceptual status - defining it as a complex auditory parameter. Siedenburg references early research by Hermann von Helmholtz, a German physicist noted for shifting his research "from exterior to interior aspects of the perceptual process" (Green & Butler, 2002, p. 246). Helmholtz notably used Fourier's theorem, concluding that it "closely described physical and physiological reality." Helmholtz states: "The quality of the musical portion of a compound tone depends solely on the number and relative strength of its partial simple tones, and in no respect on their difference of phase" (Helmholtz, 1877, p. 126) - alluding to the abstract nature of our perception of timbre.

McAdams, S. (2019, p. 23) expands on the idea of complex auditory parameters by describing timbre as a characteristic that displays both "spectral and temporal properties of the audio signal." Saitis and Weinzierl (2019, p.119) explore the idea that timbre is an intuitive concept that describes attributes not accounted for by properties like frequency or intensity - but rather "Conceptualized and communicated primarily through sensory attributes from different modalities, onomatopoeic attributes, and nonsensory/abstract attributes." Wallmark and Kendall (2018, p. 2) go on to reinforce this idea by stating that "Timbre exists at the confluence of the physical and the perceptual" and that it is this "multidimensional nature" that makes it so challenging to provide a sense of consistency that goes beyond conceptual descriptions.

Wallmark and Kendall (2018, p. 25) go on to say that "timbre descriptors cannot be attributed strictly to the physical, perceptual, or verbal frames; rather, they constitute a "hybrid" that arises cognitively." This brings us to another point in our investigation of semantic processing of timbre - that it may happen "more conceptually, prior to lexical activation" (Wallmark & Kendall, 2018, p. 25), and also that the cross-modal nature of these perceptual processes might overlap with a form of latent or weak synesthesia prevalent amongst the general population.

When considering the above, it is essential to note that we should interpret the words used to describe timbre in a broader sense than their dictionary definitions in that they "signify by tapping into a complex network of interdependent connotations and implications accrued over lifetimes of associative learning (both of an individual and a language community), not by referring to any established definition" (Wallmark & Kendall, 2018, p. 25).

When describing auditory properties such as pitch, loudness, and tempo, we tend to use one-dimensional descriptors with linear scales, for example, low/high, slow/fast, and soft/loud (Wallmark & Kendall, 2018).

Wallmark and Kendall (2018) and Saitis and Weinzierl (2019) narrowed down a set of three standard semantic dimensions for timbre as follows:

- **Luminance:** Brightness, sharpness
- **Texture**: Roughness, harshness
- **Mass:** Fullness, richness.

These descriptors share a relative prevalence across languages and cultures; however, more research is required.

Using these words leads to another question: Do verbal descriptions, such as bright vs. dull, reflect a perceptual or affective evaluation of sound qualities? Warrenburg (2020) extensively analyzed the Previously Used Musical Stimuli (PUMS) database. They found that "the results suggest that the literature relies on nine emotional terms, focuses more on perceived emotion than on induced emotion, and contains mostly short musical stimuli." (Warrenburg, 2020, p. 240). We can surmise that these descriptions may lean more towards perceptual sound qualities.

Following this thought is the idea that "previous knowledge and concept of the sound source plays a role in perception" (McAdams, 2019, p. 23). Saitis & Weinzierl (2019, p. 144) note that "People systematically make many crossmodal mappings between sensory experiences presented in different modalities or within the same modality," pointing to the fact that one potential way to circumvent lexical processing when asking participants to describe stimuli is to make use of forms rather than words - i.e., using shapes or objects as identifiers as opposed to text. Furthermore, participants may associate timbral descriptions with the instrument producing the sounds. Warrenburg (2020) notes in their paper that if prior knowledge of the sound source may influence the experiment's outcome, we should use unfamiliar sound stimuli (e.g., synthesized sounds).

Canette et al. (2021) hit on a few interesting points relating to the semantic activation caused by sound stimuli - Importantly, they found that textural sounds with blended timbres that evolved over time promoted “greater semantic network activation than do rhythmic structures” (p. 155) - indicating that the time during which participants are exposed to a specific stimuli plays a vital role in triggering a given response.

Finally, in their article “Semantic crosstalk in timbre perception,” Wallmark (2019) highlights a few key factors that enforce our study’s hypothesis. Firstly, cross-modal terms used to describe timbre may reflect similarities in the characteristics of sound and vision or touch. This again indicates a “weak synesthetic congruency between interconnected sensory domains” (p. 1) and enforces the idea that processing of one perceptual stimulus may be mildly impaired in the presence of another mismatched or task-irrelevant stimulus.

2.3 Cross-Modal Perception

Cross-modal perception can take on many dimensions. In this review, we will focus mainly on audio-visual pairings; however, we will briefly look at the audio-tactile pairing in the context of how it may relate to our study.

First, we should unpack how we, as humans process information. Marks (2004) notes that the “information processing capacity of humans is limited” (p. 85), and in their experiment goes on to pose the question of how well we can process information from one modality while ignoring information from another. Stein et al. (1996) gets right at the heart of the neurological matter and argues that the level of central nervous system activity is associated with stimulus activity. The idea is that many neurons do not fit neatly within modality-specific categories but can be influenced by multiple sensory modalities, resulting in cross-modal perception. “If these multisensory neurons participate in such fundamental functions as perceived intensity, the presence of a nonvisual (i.e., auditory) cue may have a significant effect on the perceived intensity of a visual cue.” (Stein et al., 1996, p. 497). One key concept worth noting here, by Marks (2004) is that the failures of our selective attention can be regarded as “dimensional interactions.”

A notable outcome of any perceptual stimulus in the context of audio-visual media is the concept of immersion. While the measure of immersion is beyond the scope of this research, we must contextualize the relevance of cross-modal perception as it relates to immersive experiences. Argawal et al. (2020) define immersion as: “A state of deep mental involvement in which the subject may experience disassociation from the awareness of the physical world due to a shift in their attentional state" (p. 6). They go on to identify two significant perspectives on immersion:

- An individual's psychological state, and
- The objective properties of a system and its technology

These two descriptions can also be referred to as "immersive tendency" (at the individual level) and "immersive potential" (at the system level). They describe immersion as a "cognitive construct" in which multiple cross-modal percepts are activated simultaneously to either enhance or detract from an experience (Agrawal et al., 2020, p. 9). This points toward our hypothesis that timbral modulations can influence our perception of temporal visual stimuli.

Wallmark et al. (2021) pose the following questions: "Would the ''dark'' deeds of a villain be perceived as darker if accompanied by a ''dark'' timbral palette, and perhaps less so if set to ''bright'' sounds? Would a protagonist's ''rough'' day be facilitatively underscored by ''rough'' sounding instruments?" (p. 14)

This idea is echoed in Salselas et al. (2021) when they pose the question: "In what ways can sound be used as a subliminal element that shifts or induces focus?" (p. 737)

Salselas et al. (2021) reflect Agrawal et al.'s (2020) concept of "cognitive construct" in their description of the "audiovisual contract" between sound and visual stimuli, offering that the same visuals, using different sounds, can create multiple contexts. This idea adds to the argument for our hypothesis of modulating the timbre of an auditory stimulus either congruently or incongruently with visual stimuli.

When considering how sound may interact cross-modally with visual stimuli, it is also important to unpack the role of the sound sources being used for our experiments. As mentioned above, Salselas et al. (2021) described an "audiovisual contract" between sound and image, which allows the sound designer to steer narratives with these combined stimuli. "Therefore, sound design, whether it is music, audio effects, or foley, has the ability to manipulate and intensify the visuals and has always had a fundamental role in storytelling in the context of linear audiovisual media narratives." (Salselas, I., Penha, R., & Bernardes, G., 2021, p. 739)

In the context of our research, creating sound stimuli that are unfamiliar to our test subjects is essential because it has been noted that users tend to relate or map their perception of a sound's characteristics to its source, i.e. if a user knows what a trumpet sounds like, they will map those attributes to their description of the sound (Siedenburg et al., 2019; McAdams, 2019; Warrenburg, 2020)

Iwamiya (2013) identifies two subcategories within their explanation of perceived congruency, namely:

- Formal congruency: The matching of auditory and visual temporal structures.
- Semantic congruency: The similarity between auditory and visual affective impressions.

Formal congruency is particularly interesting because it focuses on synchronizing sound and visual events. This synchronization increases perceived congruence between modalities - and is the variable we will try to isolate and manipulate in our experiments.

Marks (2004) explains three key concepts for the interpretation and measuring of cross-modal perception:

- Garner Interference: "when variation in the irrelevant dimension interferes with the processing of the relevant dimension." (p. 89)
- Congruence effect: "Responses are more accurate and quicker when a stimulus has two or more perceptual attributes that are congruent." (p. 87)

Further elaborating on the concept of "dimensional interactions" (mentioned above), they go on to explain that these can be observed in tasks of speeded classification - where test subjects are asked to classify multiple stimuli as fast as possible to measure the subject's response time (RT). We will explore more about testing methods in the following section.

2.4 Testing Methods - The Speed-Accuracy Trade-off (SAT)

This section will briefly describe the most popular testing methods and examine examples of previous experiments that are relevant to our research topic. The most frequently referenced effect in the research on cross-modal perception examined for this review is the "Speed-Accuracy Trade-off" (SAT). SAT is defined as follows in the "Encyclopedia of Clinical Neuropsychology": "The complex relationship between an individual's willingness to respond slowly and make relatively fewer errors compared to their willingness to respond quickly and make relatively more errors is described as the speed-accuracy tradeoff." (Kreutzer et al., 2011, p. 2344)

Heitz (2014) provides an in-depth overview of the effect, tracing it as far back as the mid-1800s when Herman von Helmholtz was performing experiments on nerve conductivity.

“Helmholtz’s logic was perhaps just as important as his discovery: one can use the time of an overt movement as a dependent measure, and by altering the antecedent conditions, estimate the duration of intermediary components.” (Heitz, 2014, p. 1)

Following this, the first time a relationship between choice accuracy and decision time was identified happened in 1911 when V.A.C. Henmon performed a simple discrimination experiment. “His data revealed an orderly relation, suggesting they were not independent” (Heitz, 2014, p. 2) - this relationship came to be known as the “speed-accuracy relation.”

Towards the late 1950s, mathematical decision models were being introduced to SAT to show that “two-choice decisions could be modeled as a stochastic process.” (Heitz, 2014, p. 2) These developments led to various ways of SAT data analysis, including the Speed Accuracy Trade-off Function, Conditional Accuracy Function, and the Quantile Accuracy Function. The most commonly referenced method for measuring cross-modal perception appears to be SATF. “The SATF plots mean RT and accuracy rate for each SAT condition separately. It reflects the efficacy of the experimental manipulation and quantifies how accuracy trades off with RT, on average.” (Heitz, 2014, p. 8). Overall, SATs offer an in-depth perspective of strategic alterations to the decision process; however, SAT experiments can be costly as they require a large number of subjects. The benefit of this cost is high accuracy in measuring neural mechanisms of the decision-making process.

What follows are three key examples of previous experiments that were performed using SAT and their relevance to our research:

In their experiment combining simple visual movement and unidirectional pitch shift, Arita et al. (2005) found that “the combination of a rising image and an ascending pitch scale, and that of a falling image and a descending pitch scale did indeed create higher perceived congruence than the alternative combinations. This experiment confirmed that the vertical correspondence of direction between visual movement and pitch shift effectively created perceived congruence.” (Arita et al., 2005) This experiment is particularly interesting because the investigators used changes over time in their audio and visual stimuli to measure users’ perceived congruence. Similarly, we will aim to modulate timbre both in frequency and timing variables to measure their impact on congruence when combined with visual stimuli.

Guest et al. (2002) investigated the effect of modulated frequency content in auditory stimuli on tactile perception. Again the temporal variable in this study is of crucial interest. The study found that “attenuating high frequencies led to a bias towards an increased perception of tactile smoothness” (p. 161). Furthermore, “These experiments demonstrate the dramatic effect that auditory frequency manipulations can have on the perceived tactile roughness and moistness of surfaces, and are consistent with the proposal that different auditory perceptual dimensions may have varying salience for different surfaces.” (p. 161)

Wallmark et al.’s (2021) research article “Does Timbre Modulate Visual Perception? Exploring Crossmodal Interactions” is of fundamental interest to our research topic. This article delves into the idea that we commonly use non-auditory terms to describe our perception of sounds and timbres. It goes on to pose the question of whether multisensory processing is taking place when listening to different timbres. It then sets up an experiment to examine how test subjects' perceptions of images changed with different audio stimuli. Our thesis research will be in close proximity to this topic, as we aim to investigate the subject’s perception of visual stimuli by manipulating brightness in both the auditory and visual modalities in combination with the temporal variable.

2.5 Conclusion

In the sections above, we have explored the semantics involved when talking about audio and visual stimuli in the context of cross-modal perception. We established that when describing auditory properties such as pitch, loudness, and tempo, we tend to use one-dimensional descriptors with linear scales, for example, low/high, slow/fast, and soft/loud. Conversely, timbre is a cognitive hybrid of physical, perceptual, or verbal descriptors. Wallmark et al. (2018) and Saitis et al. (2019; 2020) narrowed down a set of three standard semantic dimensions for timbre as follows:

- **Luminance:** Brightness, sharpness
- **Texture**: Roughness, harshness
- **Mass:** Fullness, richness

Next, we broadly explored cross-modal perception as a research topic, discussing the processes that take place when subjects perceive stimuli across more than one modality simultaneously and pinpointed a few key concepts relevant to our research, including “dimensional interaction,” “formal congruency, “semantic congruency,” “Garner Interference, and the “Congruence effect.”

Lastly, we looked at the Speed Accuracy Trade-off and its relevance as a testing method for our research.

Sound is inherently transient - and when applied in a narrative context, where changes in stimulus progress the participant's experience from one state to another, identifying how sounds can enhance or manipulate other sensory stimuli is of crucial importance.

This thesis aims to further our understanding of how cross-modal sensory stimuli impact each other in interactive contexts. We will specifically examine how modulations in timbre can affect brightness perception when both modalities are modulated over short periods.

In this study, we are performing two experiments based on the following two hypotheses:

1. Modulated changes in timbre, from bright to dark or vice versa, will influence participants' perception of visual changes in brightness over a short period of time.
2. Asynchronous modulations in the brightness of timbre and visual stimuli will influence participants' ability to observe the changes in one modality versus the other over a short period of time.

If our experiments confirm these hypotheses, they will open up new areas of exploration into how we could use changes in timbre to influence or manipulate time-sensitive actions, e.g., cues, prompts, and tasks in interactive and immersive experiences.

3. METHOD

3.1 Experiment Design

The focus of this study was to measure the influence of auditory timbre on visual perception. We performed an experiment consisting of 2 parts. In both parts 1 and 2, the timbre of the auditory stimulus was modulated (changed over time), from either bright to dark, or dark to bright, over a 2-second period. In both parts of the experiment, the brightness of the visual stimulus changed from a base color to either bright or base color to dark, in congruent and incongruent pairings with the auditory stimuli. The function of the audio stimuli was for the timbral effects to either enhance or detract from responses to the visual stimuli.

Both parts 1 and 2 made use of the speeded response paradigm to measure participants' reaction times in determining changes in the visual stimuli. The aim of the study was to identify whether auditory stimuli, with a timbre described as characteristically “dark” or “bright”', have an effect on visual perception response speed and accuracy when the stimuli are modulated or changed over a short time period. If the descriptors used to describe timbre were solely associated with the linguistic domain, then the auditory stimulus should not have any impact on the test subjects’ responses to the paired visual stimulus.

Alternatively, if there was interference (in incongruent pairings) or facilitation (in congruent pairings) in the accuracy or response time due to the auditory inputs, this would indicate crossmodal activity in subjects’ responses. To further the hypothesis of congruency facilitation between audio and visual stimuli pairings, in part 2 we investigated whether congruent pairings actively sped up response time, and impaired or altered temporal perception.

Additionally, part 1 of the experiment incorporated deceptive “forced choice” questions, where either the auditory or visual stimuli did not modulate, and participants were asked to respond whether the visual stimuli changed brightness.

Part 1 aimed to measure if cross-modal congruency had an effect on visual perception. The audio and visual stimuli were either played in the same direction (congruently - sound and visual both change from bright to dark) or in opposite directions (incongruently - sound = bright to dark, visual = dark to bright).

Part 2 aimed to measure if cross-modal congruency between visual and sound had an effect on temporal perception of visual changes. The audio and visual stimuli were played in the same direction (congruently) but at different speeds or modulated at different rates, e.g., the visual faded to dark slower than the audio stimulus modulated timbre. The two parts were run consecutively for each participant during the same session.

3.2 Participants

In total 62 participants were recruited for the experiment. Criteria for participants to be included in the study were: 18+ years of age, normal hearing ability, and normal or corrected-to-normal vision without color blindness. Criteria for exclusion included abnormal vision abilities and abnormal hearing abilities, as this study's focus was on cross-modal perception between color and sound. Participants were informed of these requirements in the recruitment messaging.

Participants were sourced through NYU Steinhardt recruitment channels - primarily asking fellow students to participate in the experiments. Participants were invited to attend the experiments in person which were conducted on a per-individual basis. Participants were presented with a randomized sequence of stimuli and asked to respond with three possible answers to each stimulus. At the end of the experiment, participants were asked to complete a short survey based on the Goldsmith Musical Sophistication Index (Müllensiefen et al., 2014), intended to establish their musical experience. The researcher was present with the participant throughout the session to introduce and guide them through the experiment and also assist with any questions they may have had during the tasks. Female to male ratio was 28 to 34, and ages ranged from 20 to 41.

3.3 Materials

3.3.1 Stimuli

Visual

Three visual stimuli were created using Figma graphic design software. The color for the baseline grey was created by applying 50% opacity to black (hex value 0x000000), resulting in a hex value of 0x808080. The bright stimulus was set at 60% opacity, resulting in grey 0x666666. The dark stimulus was set at 40% opacity, resulting in grey 0x999999. These stimuli were based on those used by Wallmark et al (2021). In their research, a separate control experiment was conducted to validate that participants were able to distinguish between the color differences of the 3 stimuli used. To add the variable of temporality, a 2-second animated fade was created for each of the 3 stimuli:

Stimuli	**Hex Value**
Baseline to Bright	0x808080 to 0x666666
Baseline to Dark	0x808080 to 0x999999
Control	0x808080 to 0x808080

Table 1 - List of Stimuli

Where a 640 x 640-pixel square was used for the Wallmark study, it was decided to utilize the entire monitor screen to display the animation, thereby minimizing the possibility of any neighboring color elements on the computer screen affecting the participants' perception of the changes in color.

Each of the animations was earmarked with a start and end color block (containing a fixation cross) of orange 0xF1D387 to indicate the start and end of the animation.

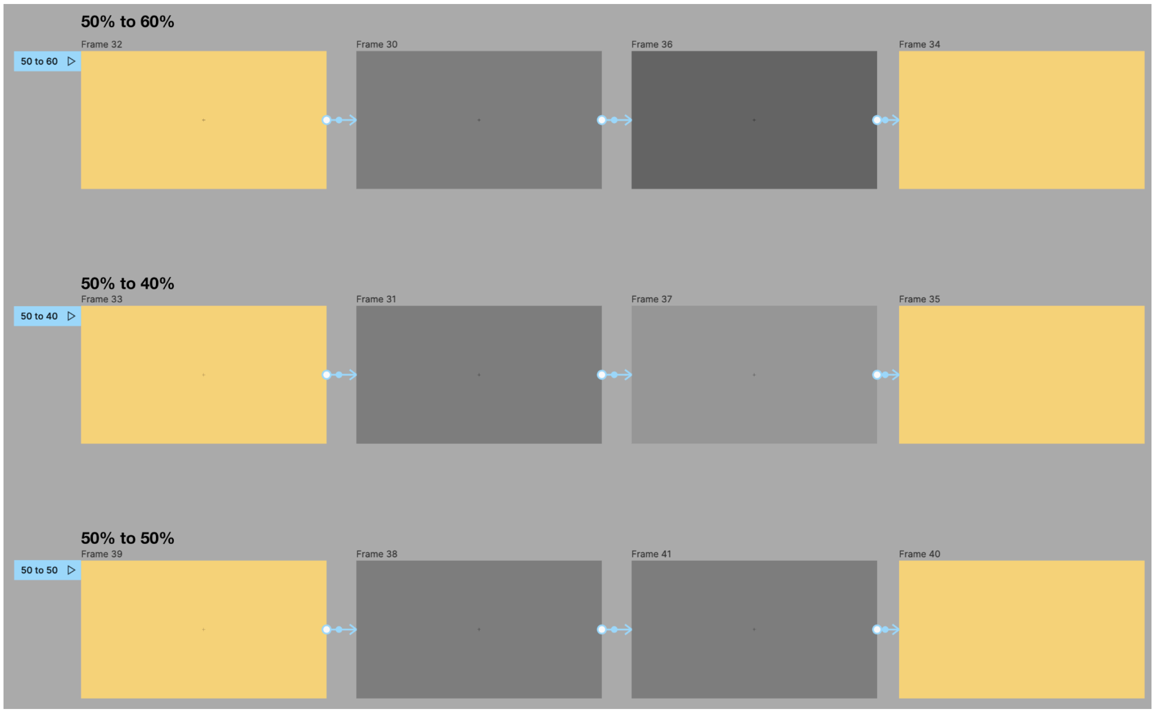

Figure 1 - Visual stimuli color variations

Audio

The audio consisted of 6 synthesized stimuli. Two of the stimuli were selected from a previous study (Wallmark, 2019) which sought to establish perceived timbral brightness in audio stimuli. Wallmark states in that study “93 natural-instrument and synthesized signals were rated by participants on a 7-point bipolar semantic differential scale (very dark–very bright)." (2019) The additional 4 sounds were created using the Operator instrument in Ableton Live 11. Each of the sounds was modeled to mirror similar timbral characteristics as those used in the aforementioned study.

For the current study only synthesized sounds were used in order to avoid any prior semantic associations with identifiable instruments. The 6 stimuli were processed with a 12dB slope low pass filter. Each of the samples was analyzed for their perceived loudness and spectral centroid, sampled at 48 kHz, and normalized to -22dB LUFS. For the 6 bright-to-dark stimuli, the filter was swept from 22kHz to 1kHz. For the 6 dark-to-bright stimuli, the filter was swept from 1kHz to 22kHz. The control stimuli were processed at a constant frequency of 4.47kHz by the filter - the logarithmic mid-point between 1kHz & 22kHz.

Synthesized Sound Names
Anger Management
Basic Glide
Big Pulse Waves*
Diagraph Persona
Ele Weble
Icy Synth Lead*

*Table 2 - *Synthesized signals (Wallmark, 2019)*

Temporality and Modulation

Human vision is optimized for detecting small and sudden changes in light and motion (Shrednoff, 2012). At the same time, small variations in timing can have pronounced effects on our perception of visual modulations. Too short, and subjects might not pick up any changes, too long, and the modulation becomes too subtle over time to gauge reaction time (Head, 2016). A time period of 2 seconds was used as the duration of modulation across the stimuli.

This duration was established by conducting an online survey of 15 participants. The visual stimuli were played without the accompanying audio stimuli, and participants were informed of the direction of change in brightness before making a selection. Each participant was asked to select an interval at which brightness changes were most noticeable - from 3 options: 500ms, 1000ms, and 2000ms. The results showed that 53% of respondents showed a preference for the 2000ms time duration, vs 36% for 1000ms and 11% for 500ms.

3.4 Procedure

3.4.1 Part 1: The temporal effects of timbral modulation on visual perception

Procedure

The study took place in a quiet room with darkened lighting to minimize outside sound and visual interference. The experiment was conducted using a Mac computer, and designed using the Psychopy application (Peirce et al., 2019). A 27-inch monitor was used to display the visual stimuli in full-screen mode, and the audio was played back over a pair of Sony MDR-7506 headphones. Participants had the opportunity to set a comfortable listening volume prior to commencing the experiment. A custom game controller was used to allow participants to input their answers, eliminating any potential reaction time interference or incorrect key presses caused by using the computer keyboard.

As with Wallmark (2021), participants were asked to respond as fast and accurately as possible to each task once the stimuli pairing had completed. Each stimuli pairing was 2 seconds in duration and ended on the orange 0xF1D387 screen. Participants were given 6 practice trials to familiarize themselves with the test procedure. Any responses recorded before the animation was completed were discarded. Response options were to identify whether the visual stimuli faded from a) “Baseline to bright”, b) “Baseline to dark”, or c) “No change”.

The 3 visual stimuli were each presented with 3 variations of the 6 auditory stimuli in a randomized order, resulting in a total of 108 trials.

3.4.2 Part 2: The effects of cross-modal congruency on temporal visual perception

Procedure

Part 2 of the experiment took place immediately following the first. The same equipment, location, and response requirements were used as in Part 1. Participants were asked to identify whether the visual stimuli stopped changing in brightness before, after, or at the same time as the auditory stimuli. The visual stimuli were 2 seconds in length, with the auditory stimulus ending slightly before or slightly after the visual. An end color 1 stop darker/brighter was used to indicate the end of the darker or brighter modulation respectively. Only the Icy Synth sample was used as an audio stimulus, with 2 timbral modulations - brighter and darker, and 3 timing variations.

Participants were asked to respond as fast and accurately as possible at the end of the stimulus pair modulation from one of the following options:
"Did the visual color fade end: a) Before the auditory stimulus, b) After the auditory stimulus, c) Same time."

The 2 visual stimuli were each presented with 2 timbre variations in 3 timing variations. For the first 48 trials, the timing intervals were held at 0.5s shorter or longer than the 2s visual stimulus. For the following 48 trials, the interval was decreased to 0.3s shorter or longer than the visual. The order of presentation was randomized, and participants completed 96 trials in total.

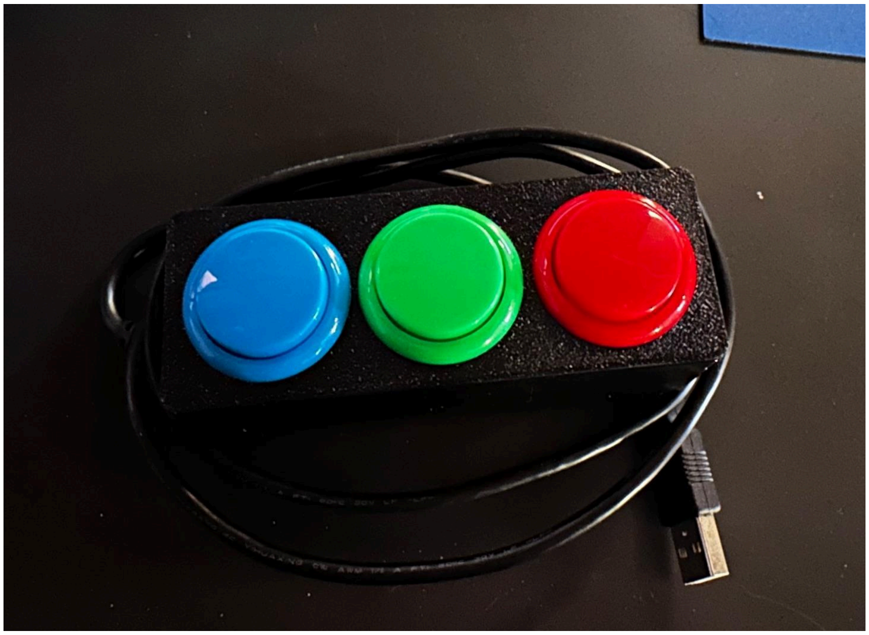

Figure 2 - Game controller used for response input

4. ANALYSIS

The results were analyzed by capturing responses for accuracy and response time for each participant in a database created within the PsychoPy application. Analysis of the data was done by running a mixed-factor ANOVA using the JASP application.

Outlier thresholds were applied to the data by determining IQR and filtering the data within JASP, removing 141 responses across both parts 1 and 2 of the experiment (Whelan, 2008). A correlation analysis for musical sophistication based on each participant's GMSI score was included in the model.

4.1 Part 1 - Timbre Modulation Analysis

4.1.1 Accuracy

Accuracy was analyzed by creating a Repeated Measures ANOVA. 2 Factors with 3 levels each were created. The first factor was *Visuals,* with levels *Visual Neutral* (Vn), *Visual Down* (Vd), and *Visual Up* (Vu). The second factor was *Audio,* with levels of *Audio Neutral* (An), *Audio Down* (Ad), and *Audio Up* (Au). The Repeated Measures ANOVA measured participants' response accuracy to the direction of the visual stimulus relative to each audio stimulus played alongside it.

The *Visual Up* (Vu) stimulus showed the highest accuracy across all audio primes, with the congruent Vu/Au pair displaying the highest mean accuracy, 0.990. There was no significant difference in accuracy between the congruent and incongruent (Vu/Au) and Vu/Ad) audio pairings (p = 1.000). There was also no significant difference in accuracy between the congruent and incongruent neutral (Vu/Au and Vu/An) audio pairings (p = 1.000).

The *Visual Down* (Vd) stimulus showed the second highest accuracy, with the congruent Vd/Ad pair displaying the highest mean accuracy, 0.972. There was a significant difference in accuracy between the congruent and incongruent (Vd/Ad and Vd/Au) audio pairings (p = .010). There was also a significant difference in accuracy between congruent and incongruent neutral audio (Vd/Ad and Vd/An) pairings (p = .038). The difference in accuracy scores between Vu/Au and Vd/Ad congruent pairs was insignificant (p = 1.000).

The *Visual Neutral* (Vn) stimulus recorded the lowest overall accuracy, with the incongruent Vn/Au pair displaying the lowest mean accuracy, 0.851. There was no significant difference in accuracy between the congruent neutral pairing (Vn/An) and darker incongruent (Vn/Ad) pairings (p=1000). There was a significant difference between congruent neutral pairing (Vn/An) and incongruent brighter pairing (Vn/Au) (p = .001). There was also a significant difference between both incongruent pairings (Vn/Au and Vn/Ad) (p = 0.001).

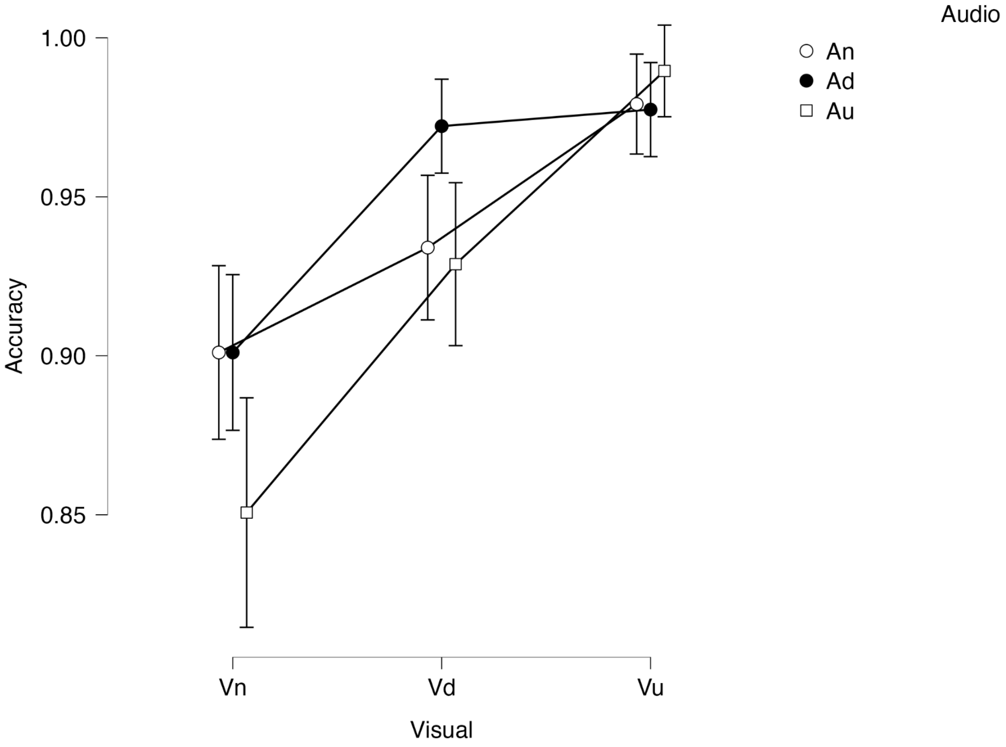

Figure 3 - Mean Accuracy for each visual stimulus across 3 audio primes

4.1.2 Reaction Time

Reaction Time was analyzed by creating a Repeated Measures ANOVA. 2 Factors with 3 levels each were created. The first factor was *Visuals,* with levels *Visual Neutral* (Vn), *Visual Down* (Vd), and *Visual Up* (Vu). The second factor was *Audio,* with levels of *Audio Neutral* (An), *Audio Down* (Ad), and *Audio Up* (Au). The Repeated Measures ANOVA measured participants' response reaction time to the direction of the visual stimulus relative to each audio stimulus played alongside it.

The Vu stimulus had the fastest reaction time across all audio primes, with the congruent Vu/Au pair displaying an RT mean of 0.410s. There was no significant difference in RT between congruent and incongruent (Vu/Au and Vu/Ad) audio pairings (p = 1.000). There was also no significant difference in RT between congruent and incongruent neutral (Vu/Au and Vu/An) audio pairings (p = .937). RT mean for incongruent (Vu/Ad) and incongruent neutral (Vu/An) displayed a slower but insignificant RT mean than the congruent pairing (p = 1.000).

The Vd stimulus showed the second fastest RT, with congruent pairing (Vd/Ad) displaying an RT mean of 0.420s. There was no significant difference in RT between congruent and incongruent (Vd/Ad and Vd/Au) audio pairings (p = 1.000), and no significant difference in RT between congruent and incongruent neutral (Vd/Ad and Vd/An) audio pairings (p = .381). The RT mean for incongruent (Vd/Au) and incongruent neutral (Vd/An) displayed a slower but insignificant RT mean than the congruent pairing.

The Vn stimulus showed the slowest RT, with the congruent pair (Vn/An) displaying an RT mean of 0.511s. The incongruent pairing Vn/Ad showed an insignificant difference in RT mean of 0.515s (p = 1.000). The incongruent pairing Vn/Au showed a slower RT mean of 0.527s, however, this difference was also insignificant (p = 1.000).

There were significant results for each instance where the Vn stimulus pairings were compared with the audio prime of the Vu and Vd pairings, e.g. Vn/An vs Vd/An and Vu/An, Vn/Au vs Vd/Au and Vu/Au, and Vn/Ad vs Vd/Ad and Vu/Ad. These comparisons were not the focus of this study, however, this result will be discussed in section 5 below.

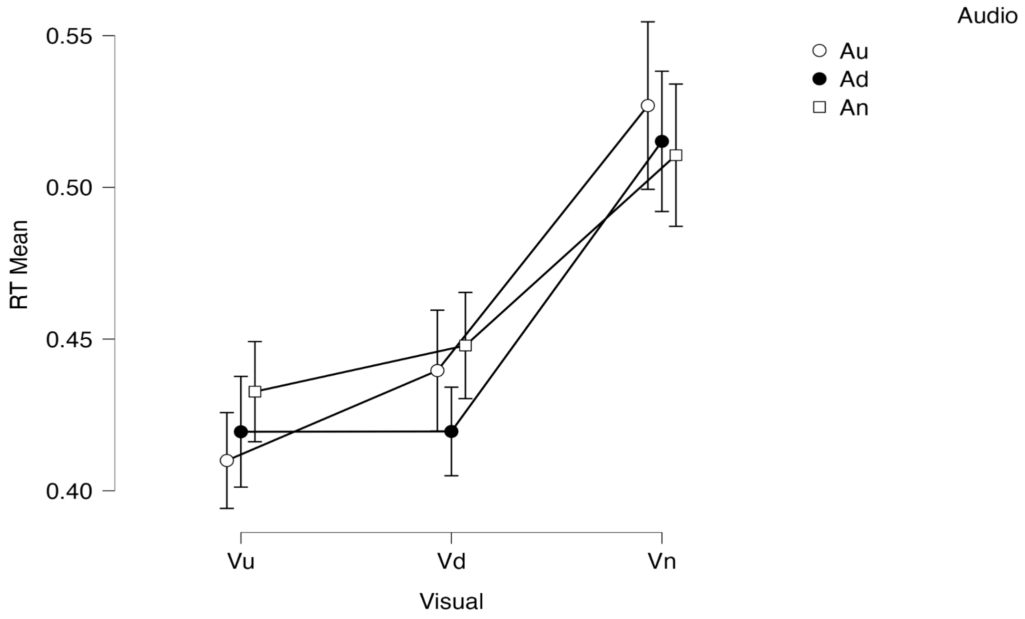

Figure 4 - RT Mean for each visual stimulus across 3 audio primes

4.2 Part 2 - Temporal Modulation Analysis

All audio/video pairings were congruent in perceived brightness (Vu/Au and Vd/Ad). For each pairing, 3 variations in audio prime were created to play longer, shorter, or end at the same time as the visual stimulus. Accuracy and Reaction Time was measured by asking participants to answer whether the visual ended before, after, or at the same time as the audio prime. Halfway through part 2 of the experiment, the interval between the visual and auditory stimuli was decreased from 0.5s to 0.3s. The purpose of decreasing the interval was to increase the difficulty as participants progressed through the experiment. The participants were not informed of this change.

4.2.1 Accuracy - 0.5s Interval

Accuracy was analyzed by creating a Repeated Measures ANOVA. 2 Factors were created: *Visual* and *Time*. The *Visual* factor contained levels *Visual Up* (Vu), and *Visual Down* (Vd). The Time factor contained levels *Before* (sound longer), *Same*, and *After* (sound shorter). The Repeated Measures ANOVA measured participants' response accuracy to the direction of the visual stimulus relative to each audio stimulus played alongside it.

Mean accuracy for both Vu and Vd displayed the same patterns across all time conditions, with no significant p values:

Visual Direction	Before (Sound longer)	After (Sound shorter)	Same
Vu	0.685	0.924	0.769
Vd	0.678	0.912	0.738
P value	1.000	1.000	0.772

Table 3 - Mean Accuracy for visual stimuli across 3 temporal variables at 0.5s interval.

Visual and sound interactions with regards to accuracy:

The Vu Same condition (Visual and audio ending at the same time) displayed a slightly higher, but insignificant, accuracy than the Vu Before condition (p = .129). The Vd Same condition displayed a higher accuracy than the Vd Before condition, however, had an insignificant difference (p = 0.478).

The Vu After condition displayed significantly higher accuracy than condition Vu Same (p = <.001). The Vd After condition displayed significantly higher accuracy than the Vd Same condition (p = <.001).

The Vu Before condition displayed significantly lower accuracy than the Vu After (p = <.001). The Vd Before condition displayed a significantly lower accuracy than the Vd After condition (p = <.001).

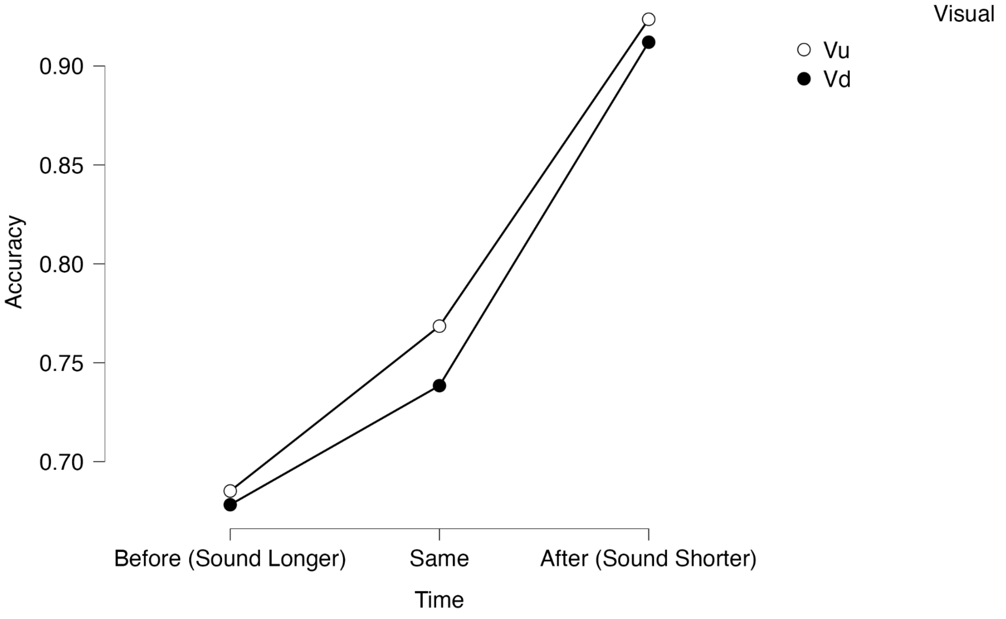

Figure 5 - Mean Accuracy for 2 visual stimuli modulating brighter (Vu) and darker (Vd), with congruent audio prime ending at 0.5s before and after.

4.2.2 Reaction Time - 0.5s Interval

Reaction Time was analyzed by creating a Repeated Measures ANOVA. 2 Factors were created: *Visual* and *Time*. The *Visual* factor contained levels *Visual Up* (Vu), and *Visual Down* (Vd). The Time factor contained levels *Before* (sound longer), *Same*, and *After* (sound shorter). The Repeated Measures ANOVA measured participants' response reaction time to the direction of the visual stimulus relative to each audio stimulus played alongside it.

RT Mean for both Vu and Vd displayed the same patterns across all time conditions, with no significant p values:

Visual Direction	Before (Sound longer)	After (Sound shorter)	Same
Vu	0.732s	0.553s	0.557s
Vd	0.761s	0.569s	0.542s
P value	1.000	1.000	1.000

Table 4 - RT Mean for visual stimuli across 3 temporal variables 0.5s interval.

Visual and sound interactions with regards to timing:

The Vu Same condition displayed a significantly lower RT compared to Vu Before (p = <.001). The Vd Same condition displayed a significantly lower RT compared to Vd Before (p = <.001).

The Vu Same condition displayed an insignificantly lower RT compared to Vu After (p = 1.000). The Vd Same condition displayed an insignificantly lower RT compared to Vd After (p = 1.000).

The Vu Before condition displayed a significantly higher RT compared to Vu After (p = <.001). The Vd Before condition displayed a significantly higher RT compared to Vd After (p = <.001).

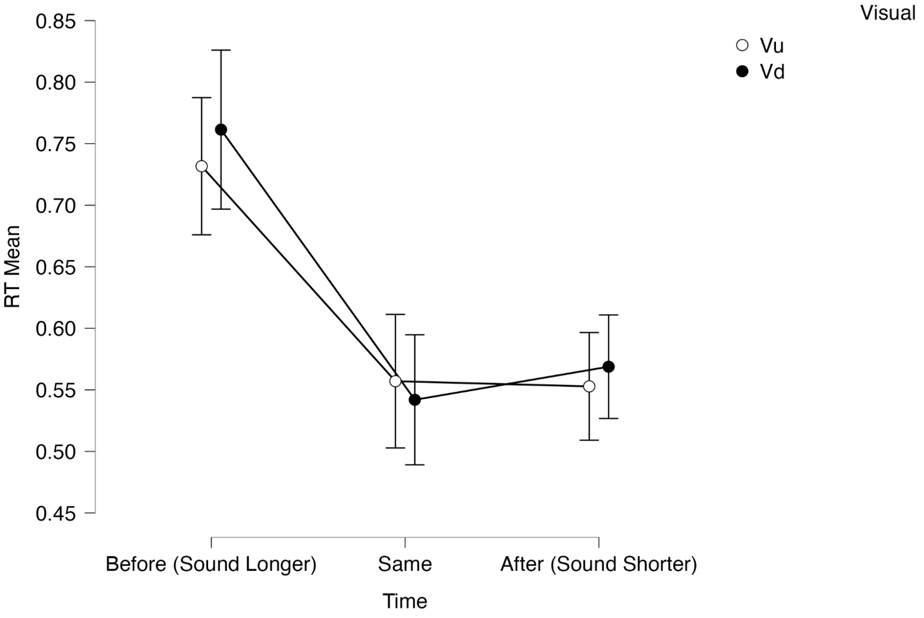

Figure 6 - RT Mean for 2 visual stimuli modulating brighter (Vu) and darker (Vd), with congruent audio prime ending at 0.5s before and after.

4.2.3 Accuracy - 0.3s Interval

The same Repeated Measures ANOVA was used as discussed in 4.2.1 above. Mean accuracy for both Vu and Vd displayed the same patterns across all time conditions, with no significant p values:

Visual Direction	Before (Sound longer)	After (Sound shorter)	Same
Vu	0.516	0.931	0.782
Vd	0.481	0.894	0.742
P value	0.514	0.514	0.514

Table 5 - Mean Accuracy for visual stimuli across 3 temporal variables at 0.3s interval.

Visual and sound interactions with regards to accuracy:

The Vu Same condition displayed significantly higher accuracy than Vu Before (p = <.001). The Vd Same condition displayed significantly higher accuracy than Vd Before (p = <.001).

The Vu Same condition displayed a significantly lower accuracy than Vu After (p = .020). The Vd Same condition also displayed a significantly lower accuracy than Vd After (p = .020).

The Vu Before condition displayed a significantly lower accuracy than Vu After (p = <.001). The Vd Before condition displayed a significantly lower accuracy than Vd After (p = <.001).

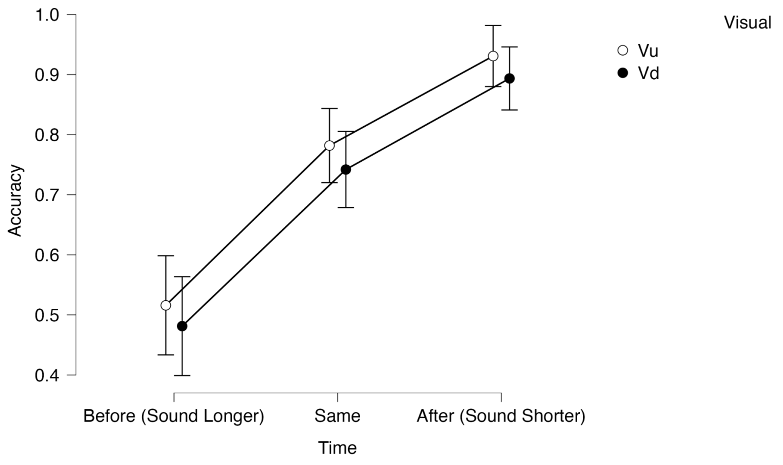

Figure 7 - Mean Accuracy for 2 visual stimuli modulating brighter (Vu) and darker (Vd), with congruent audio prime ending at 0.3s before and after.

4.2.4 Reaction Time - 0.3s Interval

The same Repeated Measures ANOVA was used as discussed in 4.2.2 above. RT Mean for both Vu and Vd displayed similar patterns across all 3 temporal conditions, however, a significant difference was displayed for the Vu and Vd *Before* conditions:

Visual Direction	Before (Sound longer)	After (Sound shorter)	Same
Vu	0.490s	0.412s	0.414s
Vd	0.616s	0.372s	0.424s
P value	0.002	1.000	1.000

Table 6 - RT Mean for visual stimuli across 3 temporal variables 0.3s interval.

Visual and sound interactions with regards to timing:

The Vu Same condition displayed a lower, but insignificant difference in RT compared to Vu Before (p = .349). The Vd Same condition displayed a significantly lower RT compared to Vd Before (p = <.001).

The Vu Same condition displayed an insignificant difference in RT compared to Vu After (p = 1.000). The Vd Same condition displayed a higher RT compared to Vd After, however, the difference was also insignificant (p = .992).

The Vu Before condition displayed a higher, but insignificant difference in RT compared to Vu After (p =.335). The Vd Before condition displayed a significantly higher RT compared to Vd After (p = <.001).

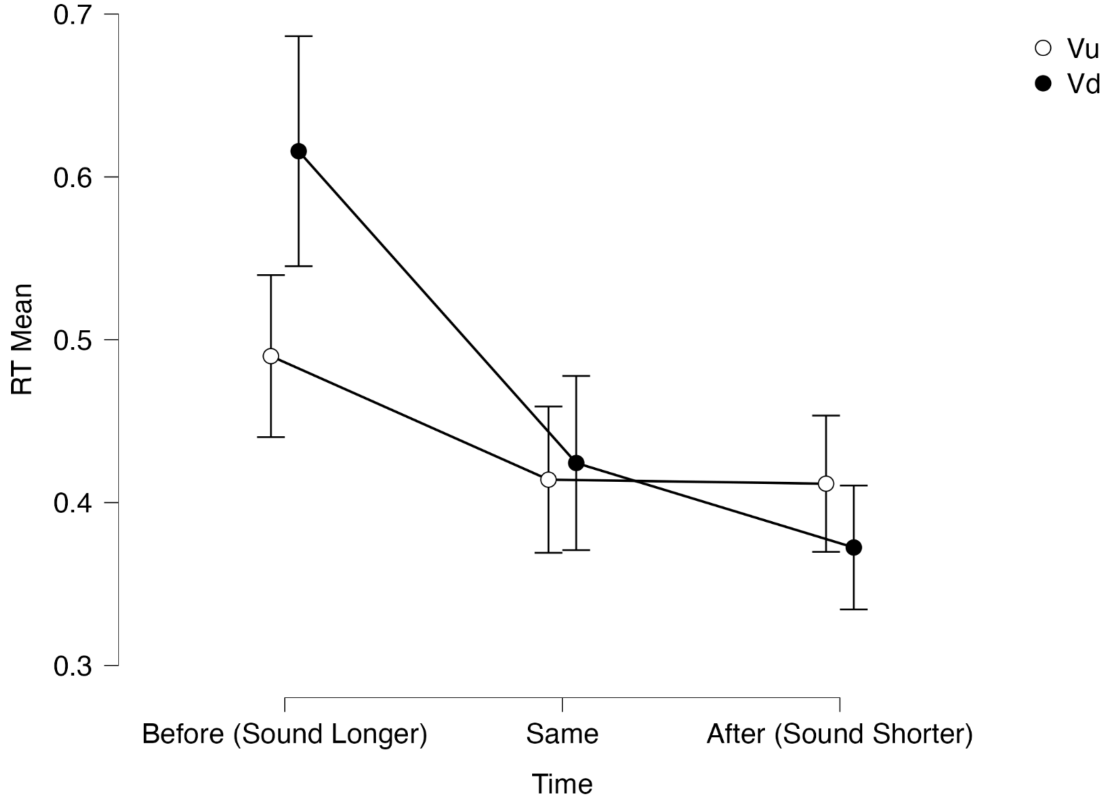

Figure 8 - RT Mean for 2 visual stimuli modulating brighter (Vu) and darker (Vd), with congruent audio prime ending at 0.3s before and after.

4.3 Musical Sophistication

A classical regression test was performed to measure the correlation between participants' scores gathered from the Goldsmith Musical Sophistication Index against both mean accuracy and mean reaction time data. Correlations were made using Pearson's r coefficient.

4.3.1 Part 1 - Timbre Modulation

There was no significant correlation between mean accuracy and musical sophistication, r = -.003, p = .980.

There was a significant correlation between mean RT and musical sophistication, p = .014. Pearson's r = -.311, showing a moderate negative correlation, in this case, a lower reaction time correlating with higher musical sophistication.

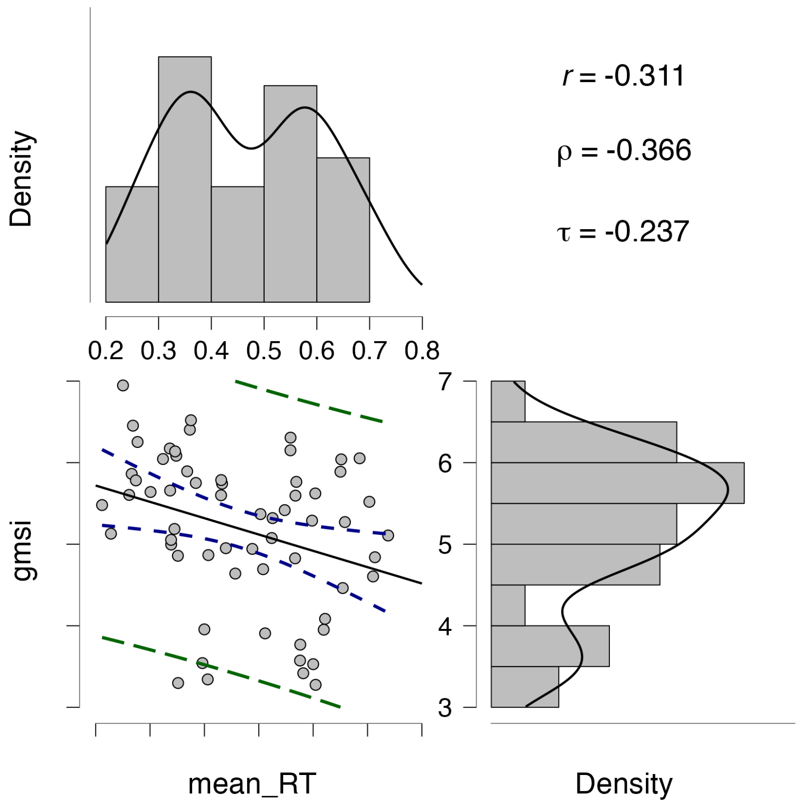

Figure 9 - Correlation between GMSI and Reaction Time shows reaction time decreasing as sophistication goes increases.

4.3.2 Part 2.1 - Temporal Modulation - 0.5s

There was a significant correlation between mean accuracy and musical sophistication, p = <.001, and Pearson's r coefficient r = .473, indicating a moderate positive correlation - with response accuracy increasing as musical sophistication increased.

There was also a significant correlation between mean RT and musical sophistication, p = .012, and the r = -.315, indicating a moderate negative correlation - with reaction time decreasing as musical sophistication increased.

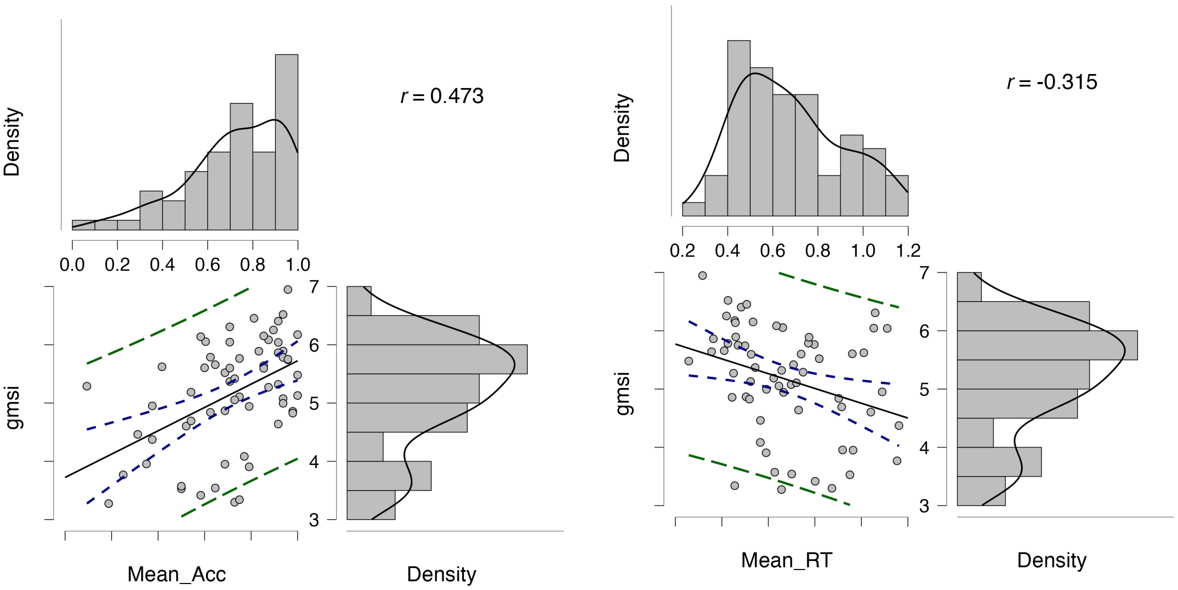

Figure 10 - Correlation between GMSI and mean accuracy and mean RT

4.3.3 Part 2.2 - Temporal Modulation - 0.3s

There was no significant correlation between either the mean accuracy or mean RT with musical sophistication once the time interval was decreased to 0.3 seconds.

4.4 Speed Accuracy Trade-off

A series of classic regression tests was performed to measure the correlation between mean accuracy and mean reaction time across both parts 1 and 2 of the experiment. Correlations were made using Pearson's r coefficient.

4.4.1 Part 1 - Timbre Modulation

Two sets of correlations were done for Part 1. Firstly, we ran a correlation test between the mean accuracy and mean RT for each user across all conditions. This returned an insignificant correlation. Secondly, we ran correlation tests for each of the 9 conditions individually, e.g. mean accuracy and mean RT for just the *Visual Up/Audio Down* condition, etc. All 9 conditions returned insignificant results.

4.4.2 Part 2.1 - Temporal Modulation - 0.5s

Two sets of correlations were done for Part 2.1. Firstly, we ran a correlation test between the mean accuracy and mean RT for each user across all conditions. There was a significant correlation between accuracy and RT, p = <.001, however, the Pearson coefficient was strongly inverted, r = -.571, with high accuracy scores corresponding to low reaction times, and vice versa.

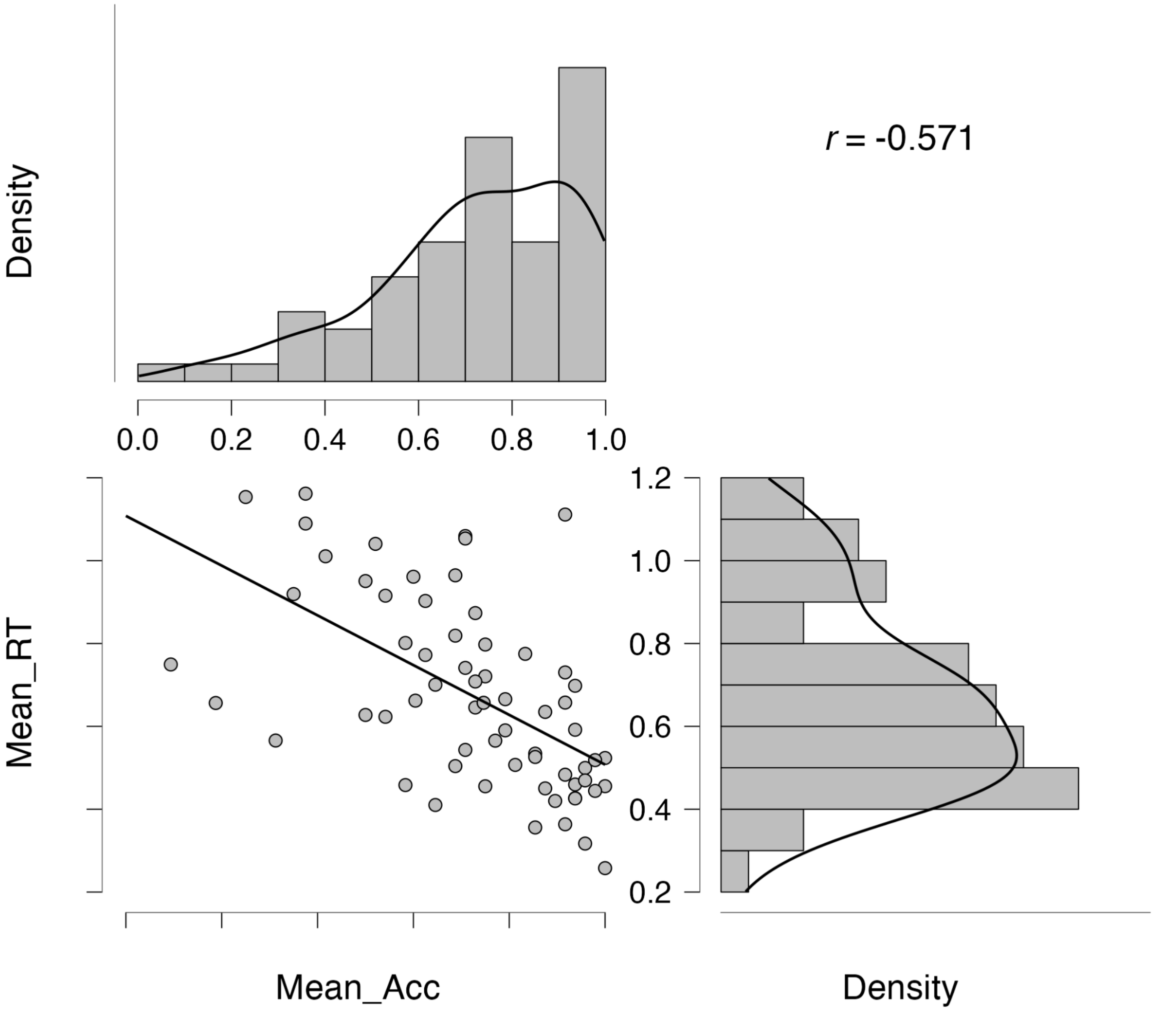

Figure 11 - Correlation between mean accuracy and mean RT, showing inverse coefficient.

Second, we ran correlation tests for each of the 6 conditions individually. All 6 conditions returned significant results, with the Pearson coefficient reflecting the same inversion observed in the first correlation test. The values are shown in the table below:

Correlation	Pearson r	p-value
SU mean accuracy - mean RT	-.329	.009
SD mean accuracy - mean RT	-.398	.001
AU mean accuracy - mean RT	-.268	.047
AD mean accuracy - mean RT	-.293	.028
BU mean accuracy - mean RT	-.621	<.001
BD mean accuracy - mean RT	-.586	<.001

Table x - Significant correlation between all conditions mean accuracy and RT, with inverse Pearson coefficient.

4.4.3 Part 2.2 - Temporal Modulation - 0.3s

The same 2 sets of correlation tests were performed as mentioned in 4.4.2 above. The first correlation test between the mean accuracy and mean RT for each user across all conditions displayed a very significant correlation between accuracy and RT, p = <.001, with the Pearson coefficient strongly inverted, r = -.606, again showing high accuracy scores corresponding to low reaction times, and vice versa.

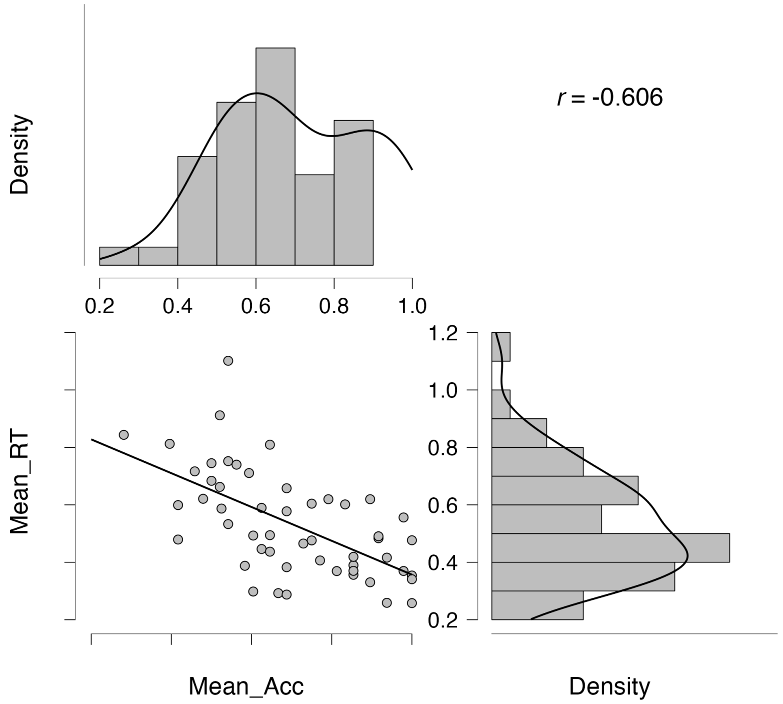

Figure 12 - Correlation between mean accuracy and mean RT, showing inverse coefficient.

After the first test, we again ran correlations for each of the 6 individual conditions. This time only 3 returned significant correlations:

Correlation	Pearson r	p-value
AU mean accuracy - mean RT	-.571	<.001
BU mean accuracy - mean RT	-.499	.002
BD mean accuracy - mean RT	-.500	.001

Table x - Significant correlation between 3 conditions mean accuracy and RT, with inverse Pearson coefficient.

5. DISCUSSION

This research aimed to further our understanding of how cross-modal sensory stimuli impact each other in interactive contexts. We specifically examined how modulations in timbre may affect visual perception of brightness when both modalities are modulated over short time periods, observing reactions in both the sensory (reaction time) and decisional (accuracy) domains as stimuli were altered from one state to another. Our research was divided into 2 parts.

5.1 Part 1

In part 1 our focus was on determining whether the timbre of an audio stimulus, when modulating changes in timbre, from bright to dark or vice versa, would influence participants' perception of visual changes in brightness, over a short period of time. The results of our experiment showed that there are trends in the data of visual perception being influenced by timbre in a manner that could indicate cross-modal interactions. Of the 3 visual stimuli paired across 3 auditory stimuli, there were 9 conditions total, 6 congruent vs incongruent comparisons.

When measuring accuracy, 3 of the 6 comparisons returned significant results: *Visual Down /Audio Down* (Vd/Ad) vs *Visual Down/Audio Up*, Vd/Ad vs *Visual Down/Audio Neutral* (Vd/An), and Vn/An vs Vn/Au. For both Vu and Vn, when paired with the incongruent Ad stimulus, there was no significant influence on accuracy when measured against their congruent pairings. For both Vd and Vn, when paired with Au, there was significant influence compared to their congruent pairings. Lastly, the Vd/Ad incongruent pair showed a significant decrease in accuracy compared to the congruent Ad/Ad pairing. From this, we can gather that the Au stimulus, modulating from a dark to bright timbre, had a bigger impact on participants' perception of visual brightness in incongruent pairs, than auditory stimuli modulating darker or remaining neutral. Furthermore, both of the Vd incongruent pairings showed significant decreases in accuracy compared with their congruent pair, indicating that visuals modulating darker were more susceptible to cross-modal interference from modulating timbre than the Vu or Vn stimuli. When considering no significant results from the Vu conditions, we can infer that participants were more confident identifying brighter visual changes than those remaining neutral or modulating darker.

When measuring reaction time none of the congruent/incongruent comparisons returned significant results, however, both the Vu and Vd incongruent pairs displayed slower reaction times than their congruent pairings, alluding to the assumption that timbre had some influence on participants' perception of visual brightness, however, more research will need to be done in order to establish concrete parameters around this effect.

Similar to Wallmark (2021) we observed that the Vu stimulus, modulating brighter, generated higher accuracy and faster RTs regardless of cross-modal congruency than darker or neutral visual stimuli. As theorized by Wallmark, these differences could result from varying levels of luminance intensity of the visual stimuli. Further research is required to identify specifically how subjects respond to varying luminous intensities in relation to timbral brightness.

There was a significant difference in RT between the Vn auditory pairings and the Vu and Vd stimulus pairings. We theorize here that it was significantly more difficult for participants to identify the neutral visual stimulus, however, this difficulty can not be attributed to the accompanying auditory stimulus, as the congruent Vn/An pairing mean RT was not significantly different from the incongruent Vn pairings.

We measured the correlation between both mean accuracy and mean RT with each participant's musical sophistication index, gathered from the GMSI questionnaire that was completed after each experiment. Part 1 of the experiment showed no significant correlation between accuracy and musical sophistication. Considering that participants were asked to focus on answering each trial based on visual observations, this result was to be expected, as the visual sensory modality does not play a role in musical training. Part 1 did however show a significant correlation when observing reaction times, with faster RT correlating with higher musical sophistication. We can therefore infer that those participants with higher musical training possess better hand-eye coordination than those with lower musical sophistication.

In performing regression tests to measure the correlation between speed and accuracy for part 1 of the experiment, we found no significant results. We primarily attribute this to very little variance in both the accuracy and RT data sets. As discussed earlier in this section, we did observe some cross-modal influence, however, participants were overwhelmingly able to correctly identify the direction of brightness modulation for each of the visual stimuli.

5.2 Part 2.1

In part 2 we built on the hypothesis of part 1, by investigating whether asynchronous modulations in the brightness of timbre, when accompanying modulating visual stimuli, would influence participants' ability to accurately observe those changes over a short period of time. Part 2 of the experiment was divided into 2 sub-sections. The first with an interval of 0.5 seconds difference in duration between the visual and auditory stimuli, the second with an interval of 0.3 seconds difference in duration. In both sub-sections, visual-auditory pairings were congruent, in each case modulating dark to bright or bright to dark.

In part 2.1 (0.5s) there was no significant difference in response accuracy across all 3 temporal conditions (Visual ending before, Visual ending after, and same time), regardless of the change in direction of either the visual brightness or timbral brightness. It can be noted that in each of the 3 temporal conditions, the Visual up/Audio up (brighter) stimulus pairing recorded marginally higher accuracy than the Visual down/Audio down pairing, for each temporal condition. As noted in the discussion in part 1 of the experiment, these differences could result from varying levels of luminance intensity of the visual stimuli.

When measuring accuracy for each temporal condition, we did however observe significant differences in accuracy responses, regardless of the change in brightness. Participants recorded the highest accuracy in trials for the Visual after condition, with significant differences in accuracy when compared to both the Same and Before conditions. The difference in accuracy between the Same and Before conditions were insignificant. Considering the insignificant difference between brighter and darker stimulus pairs for each temporal condition, this indicates that participants found instances where the auditory stimulus ended after or at the same time as the visual stimulus much harder to identify than when ending before the visual stimulus.

It is therefore clear that there is significant cross-modal interaction taking place between visual and auditory senses when participants were asked to respond to temporal visual events, however, because Part 1 of our experiment did not provide significant evidence that timbre influenced the visual perception of brightness, we can not ascribe the observed interactions to the timbral characteristics of the auditory stimuli. Further research is needed to establish whether other auditory characteristics would have similar effects on participants’ response accuracy.

When observing reaction time, there were no significant differences in RT between the brightness directions across all 3 temporal conditions, however, there were significant results when comparing the reaction times of the 3 temporal conditions to each other. Participants recorded the fastest reaction times for the Visual after and Visual same condition conditions across both up and down brightness directions, with insignificant differences in RT correlation between those two temporal conditions. Participants recorded the slowest reaction time for the Visual before condition across both brightness directions, with significant differences in RT when compared to both the aforementioned temporal conditions. Interestingly, where the significant difference in accuracy lay between the After condition (high accuracy) and both the Same and Before conditions (significantly lower); for RT the significant difference lay between the Before condition (higher RT) and the Same and After conditions (significantly lower RT). It is clear that participants found the Before condition most difficult to identify, through both lower accuracy and slower response times. With the caveat of Part 1's insignificant results, when considering the results of Part 2 in the context of timbral modulation, we can infer that where changes in auditory brightness modulated slower than their visual counterparts, there was significant cross-modal interference, making it more difficult for participants to identify temporal visual events. Conversely, where auditory brightness stimuli modulated faster than their visual counterparts, ending before the visual stimuli reached the end of their transition, participants found it much easier to identify temporal visual events.

The Same condition showed contradictory patterns across sensory and decisional dimensions, recording lower accuracy, but faster reaction times in relation to the other 2 temporal conditions. The lower accuracy measure might be ascribed to a misleading effect of participants anticipating the Before condition, where sound continued after the visual stimulus, however, this does not account for the faster reaction times.

There was a significant correlation between participants' musical sophistication, and both accuracy and RT, with higher accuracy and lower (faster) RT corresponding to higher scores on the GMSI. In complex tasks where participants had to relate small time intervals to visual changes on the screen and react as fast as possible through tactile inputs, we can infer that skills acquired through instrument playing and site reading would be of benefit.

While we did observe significant correlations between accuracy and reaction time when performing classical regression tests, the Pearson coefficients were inverted, indicating an opposite effect of any observable trade-off between accuracy and reaction time. This result can only be ascribed to varying levels of difficulty among each of the 3 temporal conditions.

5.3 Part 2.2

As we observed in part 2.1, there was no significant difference in response accuracy across all 3 temporal conditions, regardless of the change in direction of either the visual brightness or timbral brightness. The brighter visual stimulus displayed the same higher accuracy as part 2.1.

When measuring accuracy for each temporal condition, we again observed significant differences in accuracy responses, regardless of the change in brightness. Where part 2.2 differed from 2.1 is that the difference in accuracy between each of the temporal conditions was significant, whereas in 2.1 there was an insignificant difference between the Before and Same conditions. While the accuracy for the Same and After conditions remained consistent with those recorded in 2.1, with insignificant differences across both parts, scores fell significantly for the Before condition from 2.1 to 2.2. This result reinforces the observations in section 5.2 above that participants encountered significant cross-modal interference where sounds continued longer than the visual stimulus. The smaller time interval created significantly higher decisional difficulty.

When observing reaction times, results in part 2.2 reflected the similar insignificant differences in brightness direction as discussed in part 2.1, however, the difference in reaction time between Vu before and Vd Before returned a significant difference in reaction time, with Vu showing a faster RT. As previously stated, in reference to research by Wallmark (2021) we can attribute the faster RTs in response to stimuli modulating from dark to bright to higher levels of luminance intensity.

Where in part 2.1 the brightness factor returned similar results across all 3 temporal conditions, in 2.2 only 2 temporal RT comparisons returned significant results. This decrease in significant differences across conditions is evidence of an increased level of difficulty, resulting in higher error and hesitation among participants. As there is no significant correlation between the timbral pairings or brightness direction, we can not draw a correlation between the higher RT and any cross-modal interactions. There was also no correlation between musical sophistication and accuracy or RT in part 2.2, indicating that this shorter interval increased response difficulty in both the sensory and decisional dimensions for all participants.

As noted in 5.2 above, the correlations observed between accuracy and RT were inverted, indicating an opposite effect of any observable trade-off between accuracy and reaction time.

6. CONCLUSIONS

Our research aimed to further our understanding of how cross-modal sensory stimuli impact each other in interactive contexts. We specifically examined how modulations in timbre may affect visual perception of brightness when both modalities are modulated over short time periods, observing reactions in both the sensory (reaction time) and decisional (accuracy) domains as stimuli were altered from one state to another. Our research was divided into 2 parts.

In part 1 of our experiment we sought to measure the influence of timbral brightness on visual stimuli. We did this by asking participants to respond to congruent and incongruent visual/auditory pairings using the speeded response paradigm to measure reaction time and accuracy. The Audio Up stimulus, modulating from a dark to bright timbre, had a bigger impact on participants' perception of visual brightness in incongruent pairs, than auditory stimuli modulating darker or remaining neutral, indicating that visuals modulating darker were more susceptible to cross-modal interference from modulating timbre than congruent pairings. We can infer that participants were more confident identifying brighter visual changes than those remaining neutral or modulating darker. Similar to previous research, (Wallmark, 2021) we observed that the Vu stimulus, modulating brighter, generated higher accuracy and faster RTs regardless of cross-modal congruency than darker or neutral visual stimuli. While we set out to identify interference from timbral modulations in participants’ response time and accuracy, and were able to observe certain trends to this effect, the results were not significant. Further research is required to isolate specific levels of luminance intensity perception, along with correlations in timbral brightness intensity to test these two variables with each other.

In part 2 we aimed to measure if cross-modal congruency between visual and sound had an effect on temporal perception of visual changes. As in part 1, participants were asked to respond to visual/auditory stimulus pairings, however in this scenario, the auditory stimuli ended slightly before or after the visual stimuli. Part 2 was divided into two subsections, with part 2.1 at an interval of 0.5 seconds difference in duration between the visual and auditory stimuli, and part 2.2 with an interval of 0.3 seconds difference in duration. In both sub-sections, visual-auditory pairings were congruent, in each case modulating dark to bright or bright to dark. Part 2 showed significant dimensional interaction between visual auditory stimuli. In both subsections, participants found it significantly easier to identify stimuli pairs where the visual stimulus ended modulating after the auditory stimulus, as compared to when the visual stimulus ending before or at the same time as the auditory stimulus. Because our data in part 1 of the experiment did not return significant results, it is not possible to ascribe this dimensional interaction specifically to the timbral brightness effect. There is great opportunity here to do further research on our ability to identify visual events in time when paired with various auditory stimuli. If such an experiment proved that these interactions happened only with timbral modulations, we could definitively conclude that timbre impacts our ability to perceive visual events in time, however the current study's results are inconclusive.

When measuring speed-accuracy trade-off for both experiments, results were inconclusive. In part 1, there was no significant correlation between RT and accuracy due to a low variance in scores from both data sets. In part 2, the correlation between RT and accuracy was inverted, showing that the level of difficulty between the conditions within the experiment increased to such an extent that there was no correlation between datasets.

While part 1 of our experiment proved inconclusive, the trends in the data indicate evidence of our hypothesis that changes in timbre, from bright to dark or vice versa, could influence participants' perception of visual changes in brightness over a short period of time. Further research is required to identify levels of luminance intensity to isolate the correlating parameters.

Part 2 of our experiment showed significant cross-modal interaction, partially proving our hypothesis that asynchronous modulations in auditory and visual stimuli could influence participants' ability to observe the changes in one modality versus the other over a short period of time. These results open up numerous possible avenues for further research into the way sound and visuals interact with each other in a temporal context.

These findings hold implications for how we might design gaming environments, complex dashboard interfaces, educational and training tools, and ADA accessibility guidelines.

REFERENCES

Agrawal, S., Simon, A., Bech, S., Bæntsen, K., & Forchhammer, S. (2020). Defining immersion: Literature review and implications for research on audiovisual experiences. *Journal of the Audio Engineering Society*, *68*(6), 404-417.

Arieh, Y., & Marks, L. E. (2008). Cross-modal interaction between vision and hearing: A speed-accuracy analysis. *Perception & Psychophysics*, *70*(3), 412-421.

Canette, L.H., Lalitte, P., Tillmann, B., Bigand, E. Influence of Regular Rhythmic Versus Textural Sound Sequences on Semantic and Conceptual Processing. *Music Perception* 1 December 2021; 39 (2): 145–159.

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception & psychophysics*, *68*(7), 1191-1203.

Goss-Sampson, M. (2022). Statistical analysis in JASP: A guide for students.

Green, B., & Butler, D. (2002). From acoustics to Tonpsychologie. *The Cambridge history of Western music theory*, 246-271.

Guest, S., Catmur, C., Lloyd, D., & Spence, C. (2002). Audiotactile interactions in roughness perception. *Experimental Brain Research*, *146*(2), 161-171.

Head, V. (2016). Designing Interface Animation (Vol. 240). Rosenfeld Media.

Heitz, R. P. (2014). The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Frontiers in neuroscience*, *8*, 150.

Iwamiya, S. I. (2013). Perceived congruence between auditory and visual elements in multimedia. The psychology of music in multimedia, 141-164.

JASP Team (2023). JASP (Version 0.17.1)[Computer software].

Kailas, G., & Tiwari, N. (2021, January). An Empirical Measurement Tool for Overall Listening Experience of Immersive Audio. In *2021 IEEE International Conference on Consumer Electronics (ICCE)* (pp. 1-5). IEEE.

Marks, L. E. (1987). On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(3), 384.

Marks, L. E. (2004). 6 Cross-Modal Interactions. *The handbook of multisensory processes*, 85.

McAdams, S. (2019). The perceptual representation of timbre. In *Timbre: Acoustics, perception, and cognition* (pp. 23-57). Springer, Cham.

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. PloS one, 9(2), e89642.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., ... & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. Behavior research methods, 51(1), 195-203.

Saitis, C., & Weinzierl, S. (2019). The semantics of timbre. In *Timbre: Acoustics, perception, and cognition* (pp. 119-149). Springer, Cham.

Saitis, C., Weinzierl, S., von Kriegstein, K., Ystad, S., & Cuskley, C. (2020). Timbre semantics through the lens of crossmodal correspondences: A new way of asking old questions. *Acoustical Science and Technology*, *41*(1), 365-368.

Salselas, I., Penha, R., & Bernardes, G. (2021). Sound design inducing attention in the context of audiovisual immersive environments. *Personal and Ubiquitous Computing*, *25*(4), 737-748.

Shedroff, N., & Noessel, C. (2012). Make it so: interaction design lessons from science fiction. Rosenfeld Media.

Siedenburg, K., Saitis, C., & McAdams, S. (2019). The present, past, and future of timbre research. In *TIMBRE: acoustics, perception, and cognition* (pp. 1-19). Springer, Cham.

Stein, B. E., London, N., Wilkinson, L. K., & Price, D. D. (1996). Enhancement of perceived visual intensity by auditory stimuli: a psychophysical analysis. *Journal of cognitive neuroscience*, *8*(6), 497-506.

Wallmark, Z., & Kendall, R. A. (2018). Describing sound: The cognitive linguistics of timbre. *The Oxford handbook of timbre. Advance online publication. New York, NY: Oxford University Press.*

Wallmark, Z. (2019). Semantic crosstalk in timbre perception. *Music & Science*, *2*, 2059204319846617.

Wallmark, Z., Nghiem, L., & Marks, L. E. (2021). Does Timbre Modulate Visual Perception? Exploring Crossmodal Interactions. *Music Perception: An Interdisciplinary Journal*, *39*(1), 1-20.

Warrenburg, L. A. (2020). Choosing the right tune: A review of music stimuli used in emotion research. *Music Perception*, *37*(3), 240-258.

Whelan, R. (2008). Effective analysis of reaction time data. The psychological record, 58(3), 475-482.

Wood, C. C., & Jennings, J. R. (1976). Speed-accuracy tradeoff functions in choice reaction time: Experimental designs and computational procedures. *Perception & Psychophysics*, *19*(1), 92-102.

APPENDIX

Appendix A. Stimuli

Experiment 1 Data Table

Participant #	Visual Bright Audio Bright	Visual Bright Audio Dark	Visual Dark Audio Bright	Visual Dark Audio Dark	Visual Control Audio Bright	Visual Control Audio Dark
Sound 1						
Sound 2						
Sound 3						
Sound 4						
Sound 5						
Sound 6						

Experiment 2 Data Table

Participant #	Bright VisFast	Dark VisFast	Bright AudioFast	Dark AudioFast	Bright Same	Dark Same
Sound 1						
Sound 2						
Sound 3						
Sound 4						
Sound 5						
Sound 6						

Appendix B. Goldsmith Music Sophistication Index Questionnaire

AE_01	I spend a lot of my free time doing music-related activities.
AE_02	I enjoy writing about music, for example on blogs and forums.
AE_05	I often read or search the internet for things related to music.
AE_06	I don't spend much of my disposable income on music.
AE_07	Music is kind of an addiction for me - I couldn't live without it.
AE_09	I keep track of new music that I come across (e.g. new artists or recordings).
MT_01	I engaged in regular, daily practice of a musical instrument (including voice) for_ years.
MT_02	At the peak of my interest, I practiced my primary instrument for _ hours per day.
MT_03	I have never been complimented for my talents as a musical performer.
MT_06	I can play _ musical instruments.
MT_07	I would not consider myself a musician.
PA_01	I am able to judge whether someone is a good singer or not.
PA_02	I usually know when I'm hearing a song for the first time.
PA_03	I find it difficult to spot mistakes in a performance of a song even if I know the tune.
PA_04	I can compare and discuss differences between two performances or versions of the same piece of music.
PA_05	I have trouble recognizing a familiar song when played in a different way or by a different performer.
PA_06	I can tell when people sing or play out of time with the beat.
PA_07	I can tell when people sing or play out of tune.
PA_08	When I sing, I have no idea whether I'm in tune or not.
PA_09	When I hear a piece of music I can usually identify its genre.
SA_01	If somebody starts singing a song I don't know, I can usually join in.
SA_02	I can sing or play music from memory.
SA_03	I am able to hit the right notes when I sing along with a recording.
SA_04	I am not able to sing in harmony when somebody is singing a familiar tune.
SA_05	I don't like singing in public because I'm afraid that I would sing wrong notes.
BI_01	The instrument I play best (including voice) is:
ST_01	What age did you start to play an instrument?

Appendix C. Consent Form

Consent Form for IRB-FY2023-6963

You have been invited to take part in a research study to learn more about how different sounds influence our visual perception.

This study will be conducted by Emil Bergh, STEINHARDT - Music & Performing Arts Professions, Steinhardt School of Culture, Education, and Human Development, New York University, as a part of their Master's Thesis.

Their faculty sponsor is Professor Morwaread Farbood, Department of STEINHARDT - Music & Performing Arts Professions, Steinhardt School of Culture, Education, and Human Development, New York University.

If you agree to be in this study, you will be asked to do the following:

- You will be shown 2 experiments with 108 sound and visual prompts each (208 total).
- After each prompt you will be asked to respond to a question.
- At the end of the tasks you will be asked to complete a survey about your musical background.

Participation in this study will involve approximately 1hr of your time. 15 Minutes briefing on how the experiment will work, and a few practice rounds, 30 minutes performing the experiment tasks. 15 minutes of completing the musical background survey.

A potential risk of high volume will be mitigated by allowing each participant to establish a comfortable listening volume prior to the experiment commencing. There are no other known risks associated with your participation in this research beyond those of everyday life.

Although you will receive no direct benefits, this research may help the investigator understand how different sounds influence visual perception.

$20 cash payment. No payment will be made if the experiment is not started. If the participant withdraws having completed less than 50% of the experiment, they will receive a partial payment of $10 cash payment. Over 50% of completion will receive full payment. Confidentiality of your research records will be strictly maintained by - Data is never directly linked to individual identity. Keeping all completed forms in a locked cabinet only accessible to the investigator. Your information from this study will not be used for future research.

Participation in this study is voluntary. You may refuse to participate or withdraw at any time.. For interviews, questionnaires, or surveys, you have the right to skip or not answer any questions you prefer not to answer. Nonparticipation or withdrawal will not affect your grades or academic standing.

If there is anything about the study or your participation that is unclear or that you do not understand, if you have questions or wish to report a research-related problem, you may contact Emil Bergh at
(212) 998-5000
enb261@nyu.edu
82 Washington Square East, New York, NY, 10003,

or the faculty sponsor, Morwaread Farbood at
(212) 992-7680
mfarbood@nyu.edu
82 Washington Square E, New York, NY 10003

For questions about your rights as a research participant, you may contact the University Committee on Activities Involving Human Subjects (UCAIHS), New York University, 665 Broadway, Suite 804, New York, New York, 10012, at ask.humansubjects@nyu.edu or (212) 998-4808. Please reference the study # (IRB-FY2023-6963) when contacting the IRB (UCAIHS).

You have received a copy of this consent document to keep.

Agreement to Participate

____

Subject's Signature & Date

Appendix D. Participant Instructions

Welcome to our study on visual and audio perception. Today's process should take approximately 45 minutes and is divided into 3 sections: 1) Experiment 1, 2) Experiment 2, and 3) Questionaire.

The main portion of Experiment 1 consists of 3 sections with 24 tasks each. You are free to stop the experiment at any time should you feel uncomfortable or need to take a break.
During this experiment, you will be exposed to short segments of audio and visual cues. After each cue, you will submit a response by hitting a button on the controller in front of you. For each task, we would like you to focus on the visual cue you see on the screen and tell us whether the color is becoming a) Brighter, b) Darker, or c) Not sure.

There are no wrong answers. At the beginning of each task, there will be a 3-second countdown. After the cue, the screen will turn orange, after which you should respond with your answer as fast as possible. There are no wrong answers, but it is important that you go with your first reaction. To begin, we will do 3 practice rounds to get you familiar with the process.

Experiment 2 will follow the same format as Experiment 1, however, this time round we would like you to focus on both the visual and the audio clips. In each clip, either the visual or audio will end slightly before or after the other. We would like you to tell us which one (visual or audio) ended first. We would like you to respond as fast as possible with one of the following 3 options: a) Visual, b) Audio, and c) Not sure.

As before, there are no wrong answers, but it is important to go with your first reaction. To begin, we will do 3 practice rounds to get you familiar with the process.

Once you have completed both experiments, there will be an online survey for you to complete.

Appendix E. Call for participants

Hello,

We are seeking participants for our research study on visual and audio perception.

This experiment will be conducted in-person at the Research Lab located on the 6th floor (35 West 4th Street). Participants will be asked to perform speeded acuracy tests based on sound and color stimuli.

The study will take up to 60 minutes.

Your participation is highly valued and voluntary. Any participant will have the option to withdraw from the study before or during the experiment, without the need to provide an explanation.

This study will be conducted by Music Technology Master's candidate Emil Bergh, under the supervision of Dr. Morwaread Farbood.

Participants will be offered a $20 cash voucher as compensation for taking part in this study.

If you are interested in joining this study, please sign up by entering your name and NYU email address next to one of the time slots on the spreadsheet link found HERE: [link]

If you have any questions, please contact me at enb261@nyu.edu. Thank you!