Perception of Materials in Virtual Reality based on their Audiovisual properties

Harshitha Koppisetty

A Thesis submitted to the Faculty of Graduate Studies in Partial Fulfillment of the Requirements for the Degree of Master of Science.

Graduate Program in Electrical Engineering and Computer Science.

York University, Toronto, Ontario.

February 2025

Abstract

This study examined the effects of cue conflicts between auditory and visual material information in a virtual environment. All combinations of impact sounds and visual textures for four materials were paired, creating sixteen conditions. Participants, wearing a VR headset, viewed the rendered target object and heard the paired sound when it was struck with a virtual metal rod. To study the effect of agency, half the trials involved an agent striking the target (agent-interaction), while in the other half, participants struck it themselves (self-interaction). Once they classified the material of the target object, their responses and response times were recorded.

Results show that participants relied largely on auditory properties when classifying materials, no significant difference was found between agent-interaction and self-interaction modes, and in four conditions, potential audiovisual illusions were observed. These findings underscore the importance of high-quality auditory cues in VR, as discordant signals can distort perceived material properties.

Declaration

I declare that this thesis has been composed by myself and the results are my original work.

Name: Harshitha Koppisetty

Date: February 24, 2025

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Virtual reality headsets create a unique experience by allowing users to experience a computer-generated virtual simulation. To create an immersive virtual experience, users' senses need to be stimulated and engaged in order to mimic a real-world experience or create a plausible novel world experience. Current research focuses on the senses of vision, audition, and touch. Extensive research has been conducted on the visual aspects; however, recent years have witnessed an increase in auditory research, as well as in the field of haptics.

When we put on a VR headset, it takes a few moments for our senses to adjust and process the given stimuli. We automatically start comparing the displayed scene to what we would expect to see in reality. For example, the shadows may be pixelated or may not align with the direction of the light source. Perhaps the scenes appear flat and lack sufficient detail, or the colors are overly bright. Additionally, there may be slight lag when a user moves to pick up a virtual object, among other issues. Once we note any differences or points of similarity, we develop a clearer sense of what to expect and begin adjusting our expectations of the virtual world. We can then interact with the environment more confidently and comfortably as we continue adjusting to it (Seriès & Seitz, 2013).

In such an environment, if a given object exhibits all the expected properties of its material, then the experience of the interaction should be enhanced (Skarbez et al., 2021; Slater, 2009). If sound is considered a property of an object that varies based on its physical characteristics (such as shape and material), it should change accordingly. Simulating real-world sounds increases the likelihood that users will adapt to the simulation more quickly, as it reduces discrepancies between the simulation and real-world conditions. Minimizing these differences requires a deeper understanding of how an object's properties influence its associated sounds.

Human perception and identification of materials has been a subject of study for the past few decades (Adelson, 2001; Bi & Xiao, 2016; Durst & Krotkov, 1995; Faul, 2021; Giordano & McAdams, 2006; Hjortkjær & McAdams, 2016; Koumura & Furukawa, 2017; Sharan et al., 2014). Advances in computing, 2D display systems, data storage and processing, and virtual and augmented reality have expanded opportunities to study how different types of stimuli influence material perception. Visual, auditory, and audiovisual stimuli are the most commonly used signals in experimental models within this field.

Understanding which factors affect perception in a virtual environment will help identify the aspects of the simulation on which to focus efforts and how to reduce production costs while maintaining quality. Ultimately, this will make VR simulations more accessible for use in various industries. Additionally, this information can be leveraged in simulations and games to manufacture specific experiences—such as inducing confusion, distorting environmental perception, or providing cues to enhance it—for entertainment or to further research on human behavior under varying conditions.

1.1 Visual material perception

Visual cues such as smoothness, reflection, texture, and color provide information about the mechanical properties of objects by invoking learned associations accumulated over a person's lifespan through interactions with the natural environment (Adelson, 2001). These associations help identify and distinguish between materials, allowing us to intuitively understand how to interact with objects made of different materials even before touching them. For example, we instinctively handle sharp metallic objects or broken glass with caution, treat a ceramic vase or glass bottle delicately, expect to lift multiple pillows with ease, and anticipate the weight of furniture based on whether it is made of wood or plastic.

Fleming (2017) conducted an extensive review of various methods and studies that have explored visual cues for material classification. He categorized most studies in this field into two main groups: those focusing on optical properties (lighting, color, gloss, translucency, etc.) and those examining visually observable mechanical properties (viscosity, elasticity, shape, etc.).

For optical properties, the way light reflects off an object is considered one of the most critical features and has been studied through various methods, including bidirectional reflectance distribution functions (BRDF), Fresnel effects, and light map orientations (Faul, 2021; Zhang et al., 2020). Additionally, surface geometry and illumination affect how light is reflected and must be carefully controlled and analyzed. A wide range of metals and plastics can be accurately identified using a combination of these cues; however, many other materials, such as different types of cloth, matte-textured objects, and liquids, are much more difficult to distinguish. This is where mechanical properties play a critical role. For different types of cloth, density and stiffness serve as key identifiers (Bi & Xiao, 2016; Bouman et al., 2013), while for fluids, a combination of appearance and viscosity helps differentiate between various types (Van Assen et al., 2018; van Assen & Fleming, 2016).

Unlike most existing literature on the topic, Sharan et al. (2014) sought to demonstrate that material categorization is a distinct process, independent of reflectance properties (albedo, color, and surface gloss) and object recognition. They designed experiments to examine the difficulty of rapid material categorization tasks and the influence of color, texture, surface shape, and object identity on classification speed and accuracy using image datasets. Noting a lack of material-related datasets, they created their own, comprising real-life objects with materials spanning nine categories—fabric, glass, leather, metal, paper, plastic, stone, water, and wood—under varying illuminations, viewpoints, object geometries, and backgrounds.

First, they analyzed the difficulty of the standard material categorization task through two experiments. In the first one, observers were asked to categorize materials (e.g., plastic vs. non-plastic) while reaction times (RTs) and accuracy were recorded. In the second, observers viewed the same images for brief exposures (40, 80, or 120 ms), and their classification accuracy was measured to assess how quickly materials could be categorized. The results indicated that material categorization was both fast and accurate, with observers achieving 80.2% accuracy at 40-ms exposures, which is comparable to performance reported for object and scene recognition.

Next, they evaluated the importance of specific features by manipulating images to emphasize or de-emphasize particular surface properties (e.g., color, texture, gloss). Images were presented in grayscale, grayscale with blurring, negatives, silhouettes, varied shading, pixelated textures, and with altered color saturation. Performance on these modified images was then compared to performance on the original images.

The findings showed that removing individual surface cues (e.g., color or texture) had minimal effects on accuracy, but observers struggled when only one property (e.g., shape or color) was emphasized. Removing color had little impact, while contrast inversion caused a minor decline in accuracy. However, degrading texture significantly reduced accuracy. Performance was especially poor when participants were presented only with silhouettes and shading (i.e., shape, geometric information, or texture alone). The authors concluded that when one or a few cues are absent, participants can still rely on the remaining cues. However, when only a single cue is presented (e.g., negatives, silhouettes), the task becomes more challenging, as reflected by the drop in participant performance.

Finally, in their last experiment, observers were asked to classify objects as real or fake (e.g., genuine fruits vs. plastic fruits) to examine the relationship between material judgments and object recognition. While observers could distinguish real from fake objects, their performance was slower and less accurate compared to basic object categorization. This demonstrated that material categorization is not solely dependent on object recognition; shape-based object knowledge alone was insufficient to distinguish real from fake objects.

Overall, this study highlights the importance of combined cues in material recognition.

1.2 Auditory material perception

In addition to visual cues, auditory cues also play a crucial role in material identification. In daily life, people commonly recognize the material of a floor by the sound of footsteps or determine whether an object is hollow or solid based on the sound it produces when struck.

Before we delve deeper into this topic, a few key concepts need to be addressed first. The information regarding sound waves and signal processing are discussed in depth in Photinos (2021), Proakis & Manolakis (2013) and Lerch (2022). Concepts from these books that are relevant to this thesis are discussed in this section.

Sound is the propagation of energy, often in the form of vibrations, that travels through a medium (such as air, water, or solids) as pressure waves. These vibrations are typically generated by an oscillating source, such as a guitar string, vocal cords, or a loudspeaker diaphragm. Shorter, transient sounds are also generated by impact events (like 2 objects colliding), causing the objects interacting with each other to vibrate. When an object vibrates, it disturbs the surrounding air molecules (or the molecules of the medium), causing them to oscillate and initiate a chain reaction of sound wave propagation through the medium. The waves compress and rarefy the medium, creating alternating high- and low-pressure regions. Each region moves outward from the source, transmitting energy through the medium.

The propagation of sound through a medium depends heavily on the properties of the medium itself, specifically on its density, compressibility, and elasticity. Speed of sound in air at 20 °C is about 343 m/s, in water it is approximately 1,480 m/s and in solids like steel, it is 5,000 m/s or higher. Aside from the speed, other properties like the duration of propagation also change based on the medium, i.e. how quickly it loses energy or amplitude. Softer or more porous materials (foams, fabrics) absorb sound more, reducing its intensity over distance.

While the material of the medium is important, the materials of the objects generating the sound also influence its characteristics. Specifically, a material's elasticity and internal damping ability affect the sound it produces. Elasticity refers to a material's ability to return to its original shape after an impact, while internal damping refers to its ability to absorb or dissipate vibrational energy. For example, hard and brittle materials like metals, glass, and ceramics tend to have high elasticity and low internal damping, causing them to generate sharp, high-pitched ringing sounds. In contrast, softer materials like cloth or wood have low elasticity and high internal damping, meaning they are more likely to absorb the impact energy and produce lower-pitched, muffled sounds that last for a shorter duration. When collision events occur between objects of different materials, the resulting sounds differ from those produced by collisions between two objects made of the same material. The characteristics of both materials influence the sound produced. For example, a glass-on-glass collision generates a sharp, ringing sound, while a wood-on-wood collision produces a more muffled sound. However, placing a glass cup on a wooden table results in a brief but sharp 'click', combining the acoustic properties of both materials.

The unit of measurement used to measure sounds is usually a decibel. A decibel (dB) is a unit of measurement used to quantify the ratio of two values (in terms of power, intensity, or amplitude), i.e. it is a logarithmic unit. In the context of sound, 0 dB in most software, refers to the chosen reference level which acts as a reference point for measuring louder sounds, but it does not mean absolute silence.

With regards to signal processing, sound is treated as a time-varying analog signal (the continuous oscillation of pressure over time). These signals are often represented in the form of waves mapped on the basis of signal amplitude and time. To analyze these signals digitally, they need to be converted into discrete signals. This is done through the process of sampling, where instantaneous amplitude values are taken at fixed time intervals.

Fourier transforms are often used to analyze the content of a sound signal. It converts a time-based representation of information into a frequency-based representation and vice-versa. A Fourier series is a decomposition of a periodic wave into a series of sine and cosine transforms. This decomposition reveals frequency components (e.g., fundamental tones, harmonics) that shape what we perceive as pitch or timbre. A Fourier transform is the mathematical conversion of the time-domain signal into a collection of sinusoidal components of different frequencies, phases and amplitudes (i.e. the frequency domain representation). For a continuous time signal $f(t)$, it is defined as:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t}dt \tag{1.1}$$

$F(\omega)$ is the representation of the function in the frequency domain, ω is the angular frequency in radians per second, $-j = \sqrt{-1}$ and $e^{-j\omega t}$ represents a complex sinusoid, combining both a cosine and a sine wave and the integral computes the contribution of each frequency component in the signal. The output is a continuous function in the frequency domain.

A Discrete-Time Fourier Transform (DTFT), calculated for a discrete-time sequence, is defined as:

$$X(\omega) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n} \tag{1.2}$$

$X(\omega)$ is the frequency-domain function and $x[n]$ is the discrete-time signal.

Traditional amplitude vs. time graphs give us a lot of information in the time domain, like when the sound is generated, how the amplitude changes over time, when does it lose all energy (i.e. when does it stop), what type of temporal events occurred (like detecting the type and number of collisions and how far apart did they occurred), etc. However, many important sound properties, like pitch and timbre, relate to the frequency components (fundamental tones and harmonics). The time-domain representation of information often hides these details. These details also tell us how the energy of the sound wave is distributed across the different frequency bands. However, converting the information to the frequency domain would lead to obscuring temporal events and information instead. We can see this in both the continuous and discrete Fourier transform equations (equation 1.1 and 1.2), where the outputs are functions of frequency. These outputs does not keep track of time or time intervals.

To better understand both frequency and time-domain information, spectrograms often present a useful compromise. A spectrogram shows how the energy of a wave is distributed across the different frequencies over time. The y-axis of a spectrogram represents the different frequency levels, the x-axis depicts the progression of time and the colors/brightness at each point of the graph represents the power of the signal at that point in decibels.

A traditional Fourier transform would convert the entire sound wave into its frequency representation and would not allow us to observe the change in frequency over time. Instead we use a Short-Time Fourier Transform (STFT). An STFT is a finite-length discrete Fourier transform taken over a small segment of the signal (called a window) and can give a local measure of the frequency response over the length of signal considered. We calculate this value for the first window and then slide the window forward in time to repeat the process. This way, we compromise to present both the frequency and time-based information.

However, there are still trade-offs to consider when creating spectrograms. There is an implicit assumption of stationary behavior i.e. we assume that the signal is unchanging over the time window. But if the window is too large or the signal has multiple, quick transient bursts, then there may be errors in the represented spectrogram. Selecting the window size thus becomes an important but challenging decision. Using a shorter time window allows us to get a more detailed view of the temporal information but it blurs the frequency details. Conversely, a longer time window offers clearer frequency resolution but sacrifices precise timing information. Within the windows, a windowing function is used to smooth out the edges of the signal so as to avoid abrupt cuts between subsequent windows, but this could in turn introduce spectral leakage (spreading out frequencies) or other artifacts. Finally, spectrograms present a lot of information and can be hard to interpret.

Unlike time-domain graphs (which are helpful for determining the type of events that caused the sound's generation), frequency-domain graphs are better for identifying intrinsic properties such as the shape, size, or material of the sound-generating objects.

Over the past few decades, numerous studies have investigated the human ability to identify an object's material and its physical properties through its impact sounds. A typical experimental setup involves a participant observing a target object being struck and making judgments about its material and various physical properties (Giordano & McAdams, 2006; Hjortkjær & McAdams, 2016). Impact sounds convey valuable information, allowing for the identification of a range of physical characteristics.

Traer & McDermott (2016) found that judgments of the environment and sound source locations based on reverberations are generally accurate, provided the reverberations are not artificially altered. The authors conducted experiments using recordings of sounds played in various natural environments with different levels of reverberation. Participants were asked to identify sound sources and assess spatial properties (e.g., room size) of the environment. The study included both normal recordings and recordings with artificially induced reverberations to examine variations in participant responses. Acoustic features such as decay rates and frequency were analyzed to determine their influence on participant judgments. The results indicated that participants accurately identified sound sources and spatial environments. Early reflections of the sound contributed to source identification, while late reverberations informed spatial judgments. These findings suggest that participants could distinguish material properties from environmental influences based on impact sounds and reverberations, which also explains why artificially enhanced reverberations led to decreased performance.

Spectral and temporal cues, such as frequency, decay rate, and amplitude, help us make reliable judgments about object properties, including size. For example, based on sound alone, we can reasonably estimate the length of a dowel rod dropped on a hard surface (Carello et al., 1998) or the size of balls dropped on plates (Grassi, 2005). Further research by the same author demonstrated that ball size estimates were likely formed by integrating information derived from the amplitude and frequency of sounds (Grassi et al., 2013). Kunkler-Peck & Turvey (2000) analyzed the sounds generated by plates of different shapes and materials, while Tucker & Brown (2003) further examined the effects of plate shape when struck in air and underwater. Kunkler-Peck & Turvey (2000) found that the height and width of rectangular plates could be estimated when struck with a pendulum and that variations in shape also allowed for accurate shape identification purely based on sound. These experiments were repeated for a small variety of materials and showed consistent results, i.e., these physical properties can be estimated independent from the material of the object. Lakatos et al. (1997) found similar results when listeners had to estimate the width and height of struck bars.

Lutfi (2001) reported that on the basis of synthesized sounds generated from vibrating bars of different materials, listeners could distinguish between hollow and solid bars. In this experiment, they synthesized the impact sounds for solid and hollow clamped bars made of iron, aluminum, and wood. Participants completed a two-alternative forced-choice task where they identified which sound corresponded to a hollow bar. Regression analysis was conducted to see how much importance participants assigned to frequency, intensity, and decay rate. The researchers found two main strategies for decision making emerge, one that was termed the 'optimal strategy' where participants made judgments based on a combination of all cues and the other strategy was called the 'frequency bias strategy' where judgments were made solely on frequency. Participants were found to switch between strategies based on the material of the bar. For iron and aluminum, they relied more on the optimal strategy but for wood, they relied largely on the frequency to distinguish between hollow and solid bars.

Durst & Krotkov (1995) also found frequency and decay to be important factors for material classification. Klatzky et al. (2000) investigated how these shape-invariant factors affect material categorization for synthesized sounds in virtual environments. They designed a series of experiments to see how decay rate and frequency affect our judgments. Participants were asked to categorize sounds into four material classes (rubber, wood, glass, steel), rate the similarity between pairs of sounds (in terms of material properties) and lastly asked to rate the similarity between pairs of sounds (in terms of object length). Unlike Lutfi (2001), Klatzky et al. (2000) found that the decay rate of sound was the most important factor, followed by frequency. Decay rate was found to be a more robust, shape-invariant parameter for material perception, correlating with intrinsic damping properties of materials. Frequency had a smaller but consistent impact on judgments (for both materials and object lengths). The authors conclude that based on these two factors, a rudimentary model can effectively simulate material properties in virtual interactions.

Lemaitre & Heller (2012) researched how observers differentiated between and identified materials (e.g., metal, plastic, wood, glass) as well as the sound-generating action (e.g., scraping, rolling, hitting, bouncing) responsible for a presented sound. The results of their study indicate that identification of the actions was faster and more accurate than material classification. It was also harder for listeners to distinguish between certain material categories like wood-plastic and metal-glass. Material identification was also noted to be better when the sound was generated by an impact action as compared to scraping or rolling.

Their results were supported by Hjortkjær & McAdams (2016), who conducted a similar study investigating how spectral cues (frequency content, spectral centroid, and damping) and temporal cues (timing and pattern of sound events) influence listeners' identification of material properties and collision events. The study found that samples with similar spectral information and decay were more easily confused, whereas manipulating temporal information led to a decrease in the accuracy of action identification. This suggests that temporal cues—such as the timing and pattern of sound events—are more critical for identifying collision events, while spectral information is more useful for classifying materials. Furthermore, the study highlights that a combination of these cues is essential for accurately perceiving interactions in natural environments.

Further, research by Koumura & Furukawa (2017) showed how reflections and reverberations affect material perception. In their study they provided the participants with the auditory stimulus (impact sounds of wood, metal, and glass) under three different reverberant conditions - non reverberant condition, constant room conditions (where the levels of reverberation were kept constant within the same block) and varying room condition (where the level of reverberation was randomized and inconsistent). Participants categorized each sound as wood, metal, or glass in a three-alternative forced-choice task. Their responses were analyzed to measure accuracy and the effects of reverberation. Attack time, decay rate, spectral bandwidth, and spectral roughness were calculated for all sounds to assess their contributions to material perception under different conditions.

They found that while on average reverberation has little effect, there was a great deal of variation in individual responses. This led the authors to conclude that while each participant used different strategies, they all adapted to the reverberant conditions. In more reverberant conditions - attack time increased, decay rate decreased, spectral bandwidth and roughness decreased, likely due to smoothing effects of reverberation. Participants were found to rely more on spectral bandwidth as a cue to distinguish materials under the varying room conditions, where reverberation types changed frequently. They relied more on spectral roughness in the constant room conditions.

This study highlights the importance of the spectral features of the sounds. The spectral features were significantly altered by reverberation and allowed participants to use these features to compensate for reverberation's effects and make accurate material judgments.

Acoustic research on material perception is a newer field of study as compared to visual research on this topic. But, based on papers reviewed so far, we can see that frequency, decay rate and temporal information are important factors for material identification. Human audition is so sensitive to these cues that even subtle elements like reverberation, affect our perception.

1.3 Multisensory research

Numerous studies have examined unisensory visual and auditory cues in material and object classification. However, in real-world environments, multiple senses work together to provide information about objects and events. Multisensory integration plays a crucial role in material categorization, yet it is a topic that has only recently begun to receive significant research attention.

Malpica et al. (2020) ran a VR-based experiment to study crossmodal perception of materials. Visual cues like glossiness and texture were made as realistic as possible. Auditory samples were taken from the MIT hit sounds dataset. The visual textures were then rendered in both high and low quality. The specific sounds and materials that were paired together, remained fixed despite differences in visual quality. Participants were then shown the stimulus in a visual-only condition and an audiovisual condition and asked to rate the quality of the materials on the basis of glossiness, roughness, and realism. It was found that the addition of audio greatly improved perceived quality and realism. In their concluding remarks, the authors cautioned that these results were limited to the stimulus used, and future work should include different types and qualities of sounds alongside varied visual qualities.

However, a previous study by Bonneel et al. (2010) did consider different qualities of both sounds and visual cues to investigate how bimodal integration of information affected material perception in VR. They researched different levels of detail (LoD)/quality in both visual and auditory cues. Participants were shown 2 simulations of the same object falling on top of a surface. One of the two simulations had the highest quality of both the sound and visual cues while the simulation the quality of the cues were varied. The participants were not explicitly told which of the two simulations used the higher quality settings and were asked to compare the two simulations in terms of their overall quality. The results indicated that increasing the LoD of both the sound and the appearance individually, improved the perceived quality of the object. And, similar to Malpica et al. (2020), it was seen that pairing higher quality sounds with lower quality visuals also improved the perceived quality of the object while the reverse (high quality visuals paired with low quality sound) reduced perceived quality. Overall, for the same visual LoD, if sound LoD is increased, the perceived quality of the simulation also increases.

Both studies by Bonneel et al. (2010) and Malpica et al. (2020) suggest that sound can be used to enhance material perception in virtual environments while simultaneously reducing rendering costs by allowing for reducing the visual quality while maintaining the overall perceived quality. Thus, audio and video quality can be traded-off to produce the desired quality of simulation most efficiently.

While these studies showcase how the addition of sounds can aid our perception, if the auditory cues do not match the visual cues, then their addition could just as easily misguide or distort perception. There is less research on the suppressive or distortive effects of sound on perception.

One such study by Hidaka & Ide (2015) investigated the effects of bursts of white noise on visual perception. Participants were presented with Gabor patches (at 45 degrees) and asked to classify them as tilted to the right or left. Bursts of white noise of different intensities were played through a set of headphones. In one experiment, the noise was played through both the left and the right sides of the headphones, while in the second experiment, the noise was played on the side congruent with the tilt in the visual stimulus (for example, when the top of the Gabor patches tilt 45 degrees to the right, the sound is played on only the right side).

They found that the white noise reduced the accuracy of discrimination, and this effect was enhanced when the burst of noise was played congruent to the tilt in the patches. An interesting observation was that the suppressive effects weren't seen when the sounds were played using loudspeakers. A previous study by the same authors (Ide & Hidaka, 2013) showed a similar suppressive effect when the auditory cues were switched with tactile cues. It is possible that these effects of noise were due to attention or semantic interference rather than multisensory object perception.

Lastly, a study by Teramoto et al. (2012) showcased how audio cues could alter the perceived direction of motion of a moving object, especially if the object is moving in an observer's periphery. The visual motion of the object was paired with a set of horizontally aligned or vertically aligned speakers. The sounds alternated between the speakers (either played through the vertically aligned speakers or horizontally aligned speakers at a time), and the direction of motion was always perpendicular to the direction of the alternating sounds. It was found that the auditory spatial information negatively affected the perceived motion at the observer's periphery and reduced the accuracy of judgments made. This could be attributed to the human brain's integration of sensory inputs to obtain more accurate estimates of motion and spatial positioning, in which case, presenting conflicting information reduced the performance.

Thus, auditory cues enhance perceptual experiences when they align with visual information and can inhibit or distort perception when they conflict with expected visual input in a given situation. However, relatively few studies have specifically addressed the multimodal perception of materials.

1.4 Perceptual illusions and how they come about

Suppressive effects that reduce task performance - like those discussed so far, are not the only effects of conflicting sensory information. Some good examples of how these sensory conflicts are processed come from studies that explore the various perceptual illusions observed. As mentioned earlier, visual cues such as smoothness, reflection, texture, and color invoke learned associations that have been collected over a person's lifespan and help identify and categorize materials (Adelson, 2001). These learned associations can skew perception when interacting with objects (Gau & Noppeney, 2016).

Classical examples include the material-weight illusions. Objects of different materials are made to have the same shape and hollowed out or stuffed with a filler, to have the same mass. Here, the objects made of materials like expanded polystyrene were reported to weigh less than objects made of metal or stone despite having the same mass. Buckingham et al. (2009) studied this phenomenon to see how expectations based on material and size of the object influence the amount of force used to lift it and the perceived weight of the object. Their study showed that despite using the same amount of force to lift blocks of identical mass, the blocks made of metal were thought to be heavier than polystyrene. When asked to lift blocks that were of different materials and the same size but were not hollowed out to have the same mass, the participants were able to accurately report on the weight of the objects. The visual input seems to invoke certain expectations that influence how the different properties of objects are judged.

Considering audio-visual mismatch studies, a few notable ones are the McGurk effect (McGurk & MacDonald, 1976; Moris Fernandez et al., 2017), sound-induced flash illusions (Hirst et al., 2020; Shams et al., 2002), and the fusion effect (Andersen et al., 2004).

The McGurk effect (McGurk & MacDonald, 1976) is a rather famous audiovisual illusion where mismatched vocal sounds and lip movements trick us into hearing a different sound altogether. A common example of this phenomenon is when the observer is shown a video of a person mouthing the syllable 'ga' but is played with the sound of the syllable 'ba', resulting in the observer typically hearing 'da' instead. A possible explanation for the phenomenon is that when the mismatched stimulus gets integrated in our brains, instead of causing a confusion, it compromises by mixing the two to give a third different result (Moris Fernandez et al., 2017).

Hirst et al. (2020) published an overview on the last few decades of research done on sound-induced flash illusions. In the original iteration of the illusion (Shams et al., 2000), observers were presented with a single visual flash paired with two auditory beeps, which was perceived as two flashes instead of one. It was one of the first studies to show that vision is not always the dominant sense. Since then, several studies have focused on the effects of different parameters (like contrast and duration of flashes and the frequency and amplitude of the sounds) on the illusion. All throughout, this illusion proves to be robust.

There is also a different variation to this illusion called the fusion effect (Andersen et al., 2004), where two flashes appear as a single flash when accompanied by a single auditory signal. These studies highlight how visual signals combined with differing auditory signals can create perceptual illusions. These sound-induced flash illusions provided a very malleable format for testing cross-modal associations, and following the publication of this research, several others started investigating how sounds alter visual perception (Adams, 2016; DeLoss et al., 2013; Jaekl & Harris, 2009; Shams et al., 2005, 2002, 2001; Williams et al., 2021).

When we consider audio-visual mismatch regarding material perception, very few studies can be found on this topic. The two most prominent studies by Fujisaki et al. (2014) and Anderson et al. (2016) have found incompatible results.

Fujisaki et al. (2014) conducted a study to investigate how auditory and visual information are integrated and to compare different strategies used for material property judgments. To achieve this, they generated 2D images of six materials (glass, ceramic, metal, stone, wood, and bark) with realistic surface properties and recorded impact sounds for eight materials (glass, ceramic, metal, stone, wood, plastic, paper, and peppers, representing soft organic matter). A total of 48 audiovisual combinations were created, and participants completed tasks under three conditions: visual-only, auditory-only, and audiovisual. The stimuli were presented on an LCD monitor with accompanying speakers. Participants performed two tasks: first, they rated the likelihood that a stimulus belonged to one of 13 material categories on a scale from 1 to 7; second, they evaluated specific material properties such as glossiness, hardness, and roughness. Category and property judgments were completed for all audiovisual combinations and conditions.

The results showed strong audiovisual interactions in material-category perception. When sounds and appearances matched, participants rated these consistent material combinations highly but during mismatch they were likely to rate it as belonging to neither category. For example, a glass image coupled with vegetable sound had people categorizing it as plastic.

Regarding signal integration strategies, participants followed a multiplicative approach, meaning they were most likely to classify an audiovisual stimulus as belonging to a particular material category only if both visual and auditory cues aligned well. However, regression analysis revealed that in cases of conflicting stimuli, participants relied more on auditory signals. For example, an auditory-only metal stimulus was primarily classified as metal, ceramic, or glass, while a visual-only wood stimulus was mostly categorized as wood. When these were combined (a metal sound with a wood texture), the stimulus was predominantly classified as metal, ceramic, or glass—similar to the auditory-only condition—suggesting that participants prioritized sound in their categorization.

Unlike the multiplicative strategy used for material classification, the integration of visual and auditory cues for material property judgments followed a weighted averaging rule. The weight assigned to visual or auditory information varied depending on the property being judged and the congruence of the stimuli. For incongruent stimuli (e.g., a glass appearance paired with a vegetable sound), auditory cues dominated more frequently, leading the authors to conclude that auditory information was perceived as more reliable in ambiguous scenarios. Overall, visual information primarily influenced judgments of surface properties, such as glossiness, whereas auditory information played a greater role in assessing internal properties, such as hardness.

Their results contrast with those of Anderson et al. (2016), who primarily investigated the differences between multimodal and unimodal perception in virtual reality. Their study reported that integrating both sound and visual appearance led to higher accuracy in material identification. In the multimodal condition, there were six cases where sounds and appearances matched and four cases where they did not. The researchers analyzed the incongruent conditions to determine which sense was more dominant in material classification. They tested nine participants and found that in two of the incongruent conditions, participants categorized the material based on sound, while in the other two, they relied on visual appearance. Based on these findings, they concluded that neither sense consistently dominated the other. However, due to the small number of conditions tested, they suggested that expanding the experiment would provide more conclusive results.

This difference between the results of the two studies might be attributed to the different modes by which the experiments were carried out and the rigorousness of the testing conditions. Fujisaki et al. (2014) presented the stimulus using an LCD monitor while Anderson et al. (2016) used a VR headset. The additional immersive cues provided by a VR HMD may have influenced the outcomes. As noted by the authors, further testing under more controlled and elaborate conditions is necessary to determine whether sensory dominance occurs in VR.

Based on the papers discussed so far, sensory conflicts cause perceptual illusions to occur because the brain consistently misinterprets information, in a specific, non-veridical way (i.e. not fully aligned with objective reality). In most of the cases examined here, particularly the sound-induced flash illusion, extensive research has demonstrated that these illusions are robust and lead to consistent and unexpected alternative interpretations. For example, altering parameters such as contrast and duration of flashes, as well as the frequency and amplitude of sounds, did not affect the persistence of the sound-induced flash illusion.

However, the studies conducted by Fujisaki et al. (2014) and Anderson et al. (2016) present conflicting and varied findings. In the study by Fujisaki et al. (2014), the perceptual outcomes for each sensory conflict were less well-defined. Participants exhibited confusion between multiple categories rather than selecting one decisively. For instance, when presented with a combination of metal sounds and wood textures, participants often categorized the material as metal, ceramic, or glass. Unlike the sound-induced flash illusion, this effect cannot be reliably classified as an illusion, as it has not been tested extensively enough to determine its generalizability. While this example provides evidence for a potential illusion, further research is needed under varying conditions (such as different levels of immersion, changes in the geometric properties of samples, variations in sound generation, and differences in texture rendering quality and techniques) to establish whether the observed phenomenon qualifies as an illusion.

1.5 Effect of self-motion in virtual environments

In sound source localization studies (Brimijoin & Akeroyd, 2014; Kondo et al., 2012), it is commonly observed that self-motion improves the performance of a task, i.e., allowing the participant to move around in the environment provides them subtle cues that allow for greater success in task completion. In terms of material perception, perhaps allowing the user to interact with the object themselves rather than asking them to observe a third party/agent do so may similarly provide additional information and a boost in task performance. This would involve comparing the results of material categorization under the two conditions.

1.6 Aim and research questions

Our aim is to investigate how auditory signals support or interfere with material classification in virtual reality. To explore this, we designed a cue-conflict task to examine how discrepancies between visual and auditory information influence material categorization. Would participants rely more on sound or appearance? Would the conflict cause enough uncertainty to result in random responses or the selection of an entirely different material category? Furthermore, if participants were allowed to interact with the object directly, would that influence their judgments?

In a virtual environment we are able to simulate events that are harder to replicate in reality. We leveraged this property to create a realistic cue conflict experiment. We wanted to investigate the following:

Question 1: Between audition or vision, which modality do we rely more on for material identification?

In a virtual environment, it is possible to induce a conflict between the visual and auditory stimulus delivered to an observer. In such a situation, would the material classification be based more on the auditory information or on the visual information?

Hypotheses: On average, participants will classify more on the basis of sound than on visual properties to determine the material of an object.

Question 2: How much would self-interaction affect perception of a material?

If a participant is allowed to strike the target object themselves, they would be more conscious of factors like the amount of force used, the timing and speed with which the impact is made, etc., as opposed to watching an agent strike the object. Will this additional information affect a participant's judgments? Will it change how they respond or reduce the response time?

Hypotheses:

- Self-interaction will reduce response time as compared to observing another interact with the target object.
- Participant response accuracy will be higher for self-interaction than agent-interaction conditions.

Question 3: Is there any potential for audiovisual illusions in cue-conflict experiments that confuse visual and auditory material information?

Similar to the McGurk effect, are there any combinations of visual textures and impact noises that confuse the participants enough that they consistently classify the material as a third different option (different from both the sound and appearance)?

For a given combination of audio-visual stimuli, if participants label it as one material significantly more often than all the other materials and if the material selected differs from both the impact sound and the visual texture of the object, then we can say that there is a potential for audiovisual illusions to occur in this type of a setup.

Hypotheses: When presented with conflicting audiovisual stimulus for material perception, audiovisual illusions are observed.

In this study, we designed an experiment to answer these questions and address previously identified research gaps. Participants were tasked with categorizing the material of a target object presented in virtual reality. The visual appearance and impact sounds were randomly paired and presented, allowing us to examine how a cue conflict between an object's appearance and the sound of a collision event influenced material classification.

Chapter 2

Methodology

2.1 Participant details

Fifty-three Participants were recruited from the Undergraduate Research Participant Pool (URPP) at York University. After screening the data, 42 of those participants were considered (18-42 years, mean = 22.21, SD = 6.38). The exclusion criteria is discussed in Section 4.1.1. Their Demographic data is shown in Table 2.1.

Participant Attributes	Number	Percentage (%)
Gender		
Female	35	83
Male	7	16
Experience with VR		
None	21	50
Casual	13	30.9
Regular	8	19.1
Experiences Motion Sickness		
None	36	85.7
Occasional	2	4.7
Easily motion sick	3	7.14
Experience with Music		
None	23	54.7
Some (during childhood)	11	26.19
Casual	4	9.5
Regularly practices	4	9.5

Table 2.1: Participant demographic data

2.2 Auditory stimuli and setup

2.2.1 Recording the sounds

Given that our primary focus is on audiovisual integration, the selection of impact sounds for this experiment required careful consideration. Online sound sources were not used, as the recording conditions were often unclear or inconsistent. Factors such as microphone quality, recording environment, and the objects used to generate sounds needed to be meticulously controlled to minimize external variables that could skew the data. To ensure consistency, we opted to record our own impact sounds.

The recordings were obtained in a nearly soundproof environment to eliminate noise and interference, thereby preserving the integrity of the sound samples. This approach also allowed us to avoid post-processing techniques that might alter the subtle material properties of the recorded sounds.

For this we arranged a recording studio with the desired environment and equipment available as well as our own apparatus. York University's Library had a studio space available for student usage. It included:

- Sound booth (6 x 6 ft, 65 dB of sound attenuation)
- Microphones (Audio-Technica AT2020) with boom stands
- Headphones (Sennheiser HD 280 Pro)
- Recording-editing software (Hindenburg Pro) available at a workstation outside the sound booth
- One USB Audio Interface (Focusrite Scarlett 18i8)

Bit resolution was set to 24-bits and Sampling rate used was 48000Hz.

2.2.2 Physical setup - Tapping mechanism

To reduce any dependency on loudness or volume being used as a means to differentiate between sounds, all impact sounds needed to be generated using the same amount of force and speed. This way the impact would be relatively similar and differences in the sounds samples would largely be due to the intrinsic material properties. Therefore, for the physical sound production apparatus, a tapping mechanism was built and used (Figure 2.1).

The tapping mechanism consisted of an electromagnet coupled with a spring that helped lever the rod. The square plate of the material corresponding to the required sound was fixed at the other end of the setup. Utilizing the magnet and the spring ensured that no additional sounds were made as it activated and pulled the rod upwards. A button was programmed so that once pressed, it activated the magnet with a delay of two seconds. This way there was a short period of time between the button press and movement of the rod. - giving enough time to anticipate the impact and reduce any additional additional sounds being generated. The rod was then dropped onto the plate thus generating an impact sound.

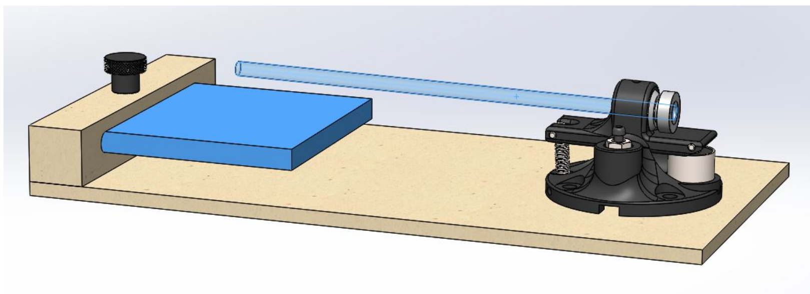

Figure 2.1: An illustration of the tapping mechanism

By using this mechanism, we were able to keep the force and speed of impact consistent across all our recordings and materials.

2.2.3 Physical setup - Materials

We required physical copies of the materials being presented in the experiment to generate equivalent sounds. Further, to diversify the sound pool, two 4”x4” square plates of different thickness (3/8” thickness and 1/4” thickness) were sourced per material. Similarly, rods of each material were also obtained (8 inches in length and 1/4th inch in diameter).

The shape and size of the samples were kept identical to reduce any dependence on geometric properties of the objects. We chose four common materials that most participants would have had large amounts of day-to-day interactions with. This was done so that none of the sounds or appearances would be new to them and their categorization of materials would be based on experiences built over time in real life. The samples used were made of glass, steel, acetal plastic and plywood and are shown in Figure 2.2. Sounds were recorded for each combination of target object and rod. However, the metal rod recordings were the most clear and were used in the experiment.

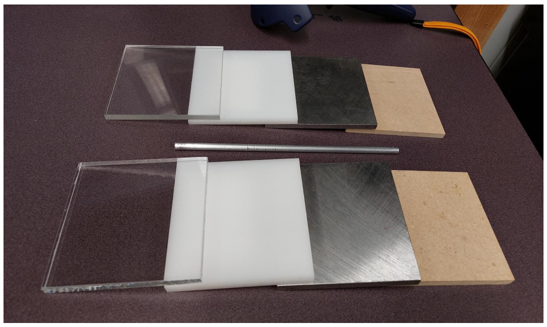


Figure 2.2: On the left are the plate materials sourced and on the right are the different rods.

2.2.4 Spectral analysis

We recorded a total of 16 sounds - 4 for each material. There were plates of 2 levels of thickness, and each plate was struck near the edge or at the center of the plate. This was done to diversify the sound samples. However, 16 sounds were later deemed to be too many and instead we used 8 sounds. Each of the 8 plates were struck near the edge (since that gave the loudest and clearest sound).

Figures 2.3-2.6 show the spectrograms of each of the sounds. The nomenclature of the sound files is: [Plate material]-[Rod material]-[thickness and position of impact]. Since we use a metal rod for this experiment, that remains constant. As for the number at the end of the file name, 1 = Thin plate struck at the edge, 2 = thin plate struck at the center, 3 = thick plate struck at the center, 4 = thick plate struck at the center.

Based on these spectrograms, we can observe that impacts on glass and metal exhibit two distinct peaks, likely due to the rod bouncing on impact. In contrast, impacts on plastic and wood show merged or less distinct peaks, possibly due to higher damping properties. Glass and metal, being high-density and rigid materials, allow for sharper, more defined impacts. On the other hand, plastic and wood, having lower densities and higher absorption, potentially caused the sound waves from the impacts to merge together.

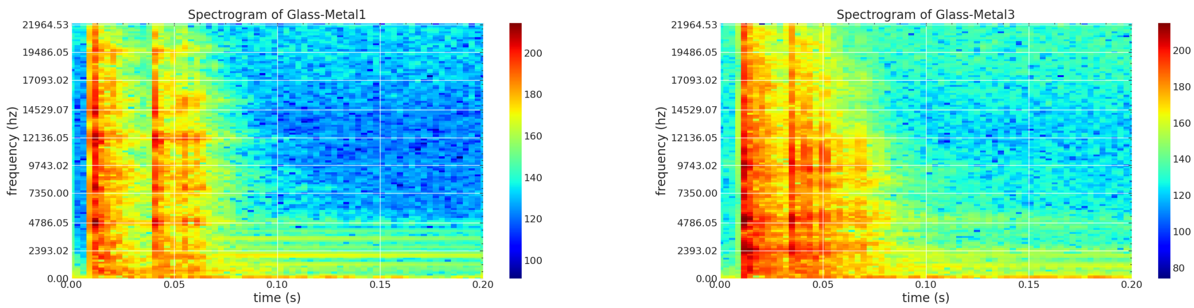

Figure 2.3: Glass spectrograms, Glass-Metal1 is the thinner plate and Glass-Metal3 is the thicker plate

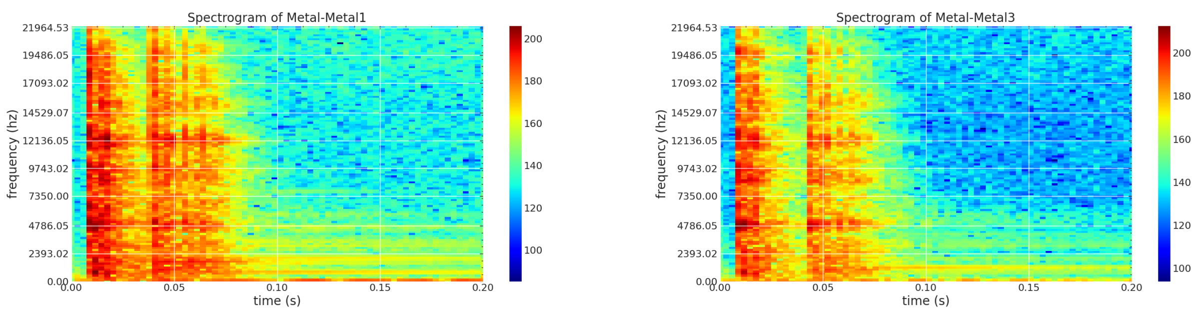

Figure 2.4: Metal spectrograms, Metal-Metal1 is the thinner plate and Metal-Metal3 is the thicker plate

Given the complex nature of sound waves and the vast amount of information one can obtain from it, no standard method exists to compare two sounds to each other. Comparing the data on the basis of spectrograms is one way to do so but obtaining something quantifiable like a similarity score is a harder task. Previous studies discussed, like Fujisaki et al. (2014); Hjortkjær & McAdams (2016) analyzed frequency specific information like the spectral centroid, spectral band-

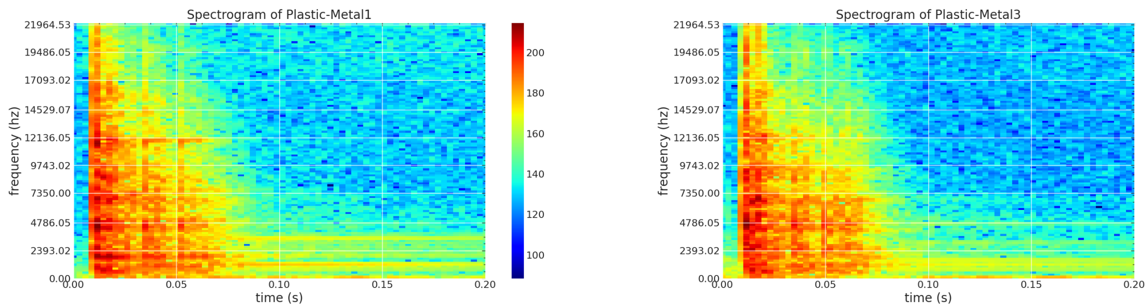

Figure 2.5: Plastic spectrograms, Plastic-Metal1 is the thinner plate and Plastic-Metal3 is the thicker plate

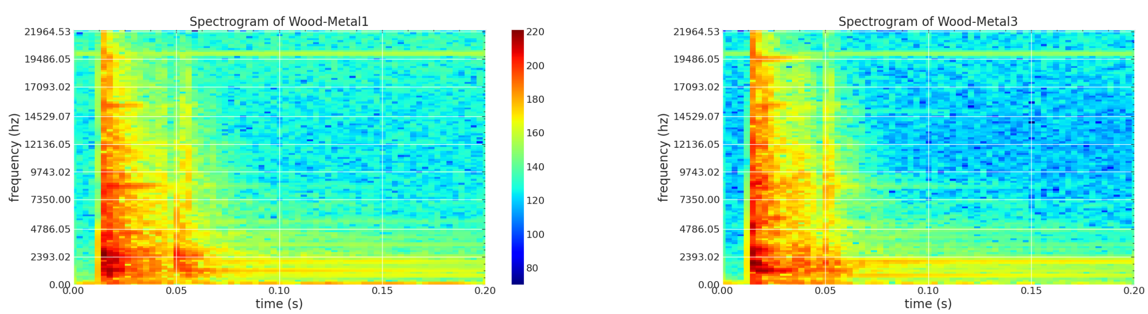

Figure 2.6: Wood spectrograms, Wood-Metal1 is the thinner plate and Wood-Metal3 is the thicker plate

width and spectral contrast.

The spectral centroid is often called the “center of gravity” of the magnitude spectrum calculated via the Short-Time Fourier Transform (STFT). The spectral centroid represents the weighted average frequency of the spectrum, with each frequency weighted by its corresponding amplitude over a short analysis window. Each individual window produces one single spectral centroid which is the weighted average frequency for that frame, giving us a series of centroid values over time. Low values indicate concentration of energy at lower frequencies; high values indicate stronger high-frequency energy. It often correlates to the perceived “brightness” or timbre of a sound and how it changes over time (since the calculations are done frame-by-frame).

The spectral bandwidth is more commonly called “spectral spread” or “instantaneous bandwidth”. Lerch (2022) defines spectral bandwidth as the weighted spread of frequencies around the spectral centroid i.e. it tells you the range of frequencies that carry most of the sound’s energy. It is almost like a standard deviation (spread) of the spectrum. A narrow bandwidth means most energy is clustered around a small frequency range (like a pure flute tone). A wide bandwidth suggests the sound spans a broad set of frequencies and imply a broad or noise-like distribution. Just like the centroid, spectral bandwidth is also computed frame by frame, thus giving us a time-varying bandwidth measure across the entire recording, describing how the spectral content changes over time.

Spectral contrast measures the difference between the loudest and quietest regions of the frequency spectrum within specific frequency bands. Instead of looking at the entire frequency range as a single unit, it is broken down into several frequency bands. For this study the bands chosen were logarithmically spaced, giving us 7 bands, with the band edges being (in Hz): 32.703 , 75.111, 172.510, 396.212, 909.998, 2090.033, 4800.273 and 11025. For every frequency band, the differences between the average of the most prominent frequency bins and average of the quieter bins in that band are calculated (generally the ratio is taken in decibels). This way, the contrast per frequency band is obtained, which collectively describe the overall shape or "texture" of the spectrum.

In order to obtain a quantifiable measure of similarity between sound samples, we combined the spectral centroid, contrast and bandwidth into feature vectors for each sound sample (one feature vector per sound) and computed the Euclidean distances between these vectors. The table with the feature values can be seen in the Appendix (Figure A.1). Figure 2.7 shows the heatmap based on the distance calculations

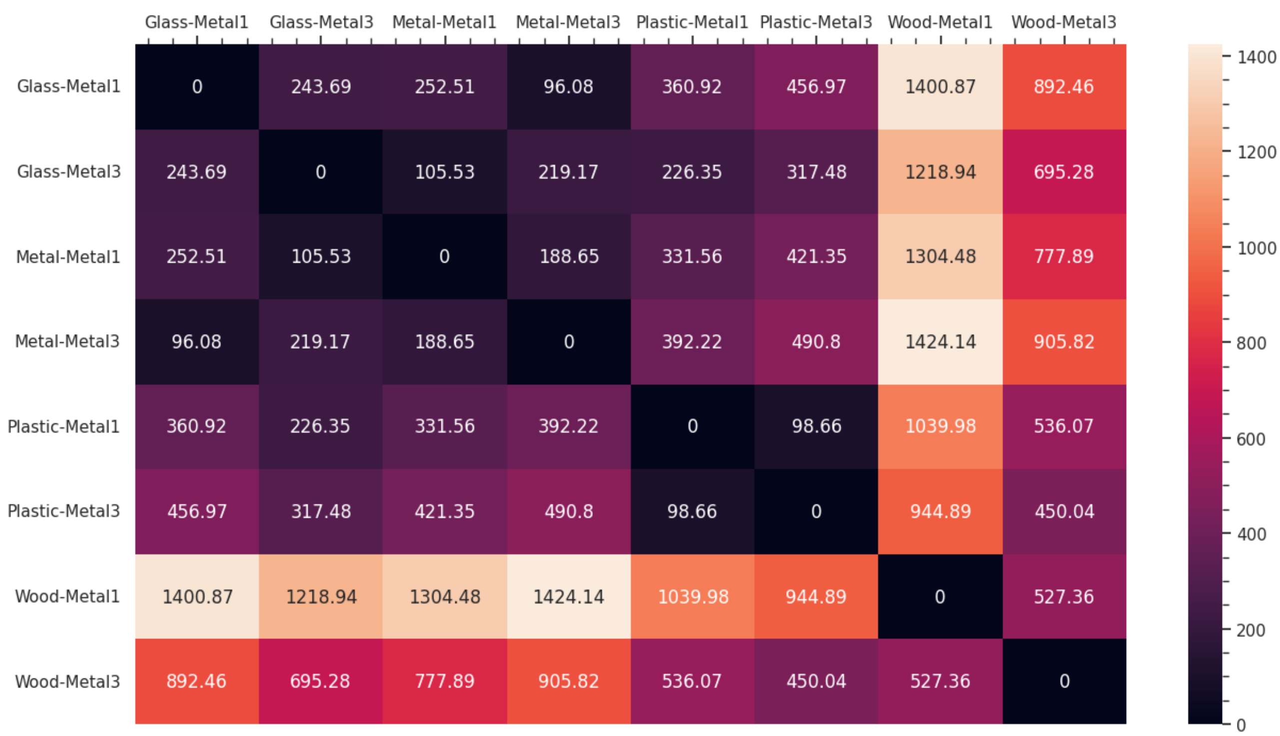

Figure 2.7: The heatmap of the calculated distance between the feature vectors of the sound samples. The colors represents how close the sounds are to each other and the numbers represent the numeric distance between the feature vectors. The more similar the sounds the darker the shade assigned.

Analyzing column 1 with the distance measures calculated for Glass-Metal1, we see that glass is closest to metal. In fact, Glass-metal1 is more similar to Metal-metal3 than Glass-metal3. Further away are plastic followed by wood. Looking at all of the distance measures calculated for any sound with glass or metal, we see that the values are very similar. Further, in last column of Wood-Metal3, plastic is the most similar to wood followed by metal and glass. Thus, if we were to group similar sounds based on this, they would be [glass and metal], [plastic and wood].

2.3 Visual stimuli and setup

2.3.1 Using AmbientCG converter tool

The experiment was coded in Unity (version 2021.3.47f1). The texturing tool used for the purpose of this study is called the AmbientCG to Unity converter 2.0 developed by Burning Mime software [https://gitlab.com/burningmime/ambientcg-to-unity/-/tree/master]. This tool allowed us to select textures available at the AmbientCG website [https://ambientcg.com/] and apply it in Unity to make realistic looking looking objects as seen in Figure 2.13.

2.3.2 Glass Shader

The AmbientCG website did not have a suitable glass texture available. Other available glass shaders were not satisfactory - some did not reflect properly, some were tinted, some did not cast shadows and some did not show other glass objects behind them. To overcome these limitations, we created a custom shader as seen in Figure 2.8.

To achieve this, we created individual nodes that calculated the reflection, refraction and fresnel effect. The inputs to each of these nodes were fed using the values from the camera and normal vector. For refraction, we needed a user defined variable IOR (index of refraction) to control the amount of refraction seen through the glass object. Then the values from these 3 nodes were combined using a lerp function (linear interpolation) and used as input to shader's 'base color field'. As the name suggests, this field decides the main color of the shader. Since we need it to be transparent, reflecting and refracting light, we used the stated combination of nodes as the inputs. All the other values of the shader were left as the defaults. The result can be seen in Figure 2.8.

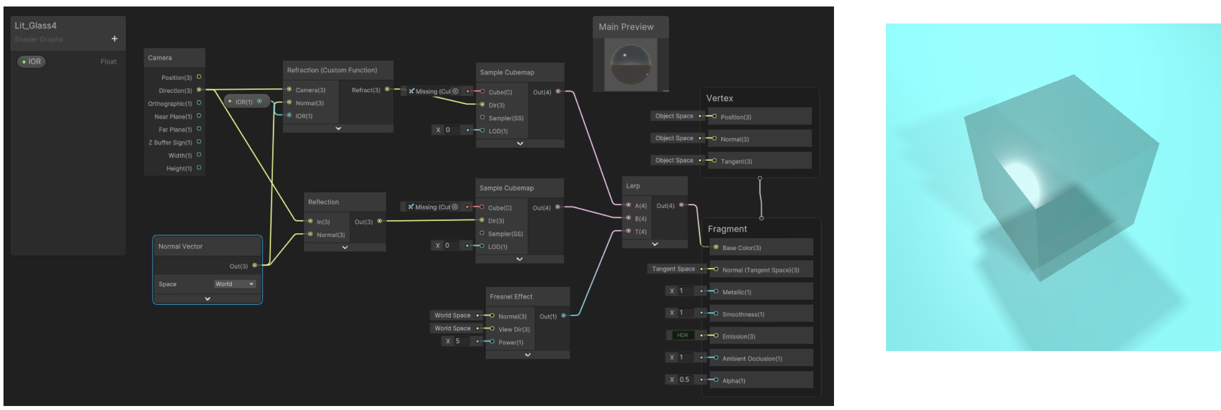

Figure 2.8: The custom glass shader created in Unity and how it looks in the scene.

2.3.3 Agent/bot

An integral part of the experiment was setting up an agent character which would hold the metal rod and execute the animation of striking the surface of the target object. We used a model from the Mixamo website because of its ease of integration in Unity projects as well as the wide variety of different models and animation already available for import. We applied a breathing idle animation on the model to give it a little natural sway while it moved but could not find a ready-made animation that matched the striking action we required for our experiment.

After importing the model and breathing animation to Unity, we then used Unity's animator and animation system to design the interaction. Since Mixamo animations are not editable, we duplicated the animation file to make it mutable. The animation needed was broken down into 5 poses or keyframes, as seen in Figure 2.9, to model the transition. Unity's animator interpolated between them and produced a smooth animation.

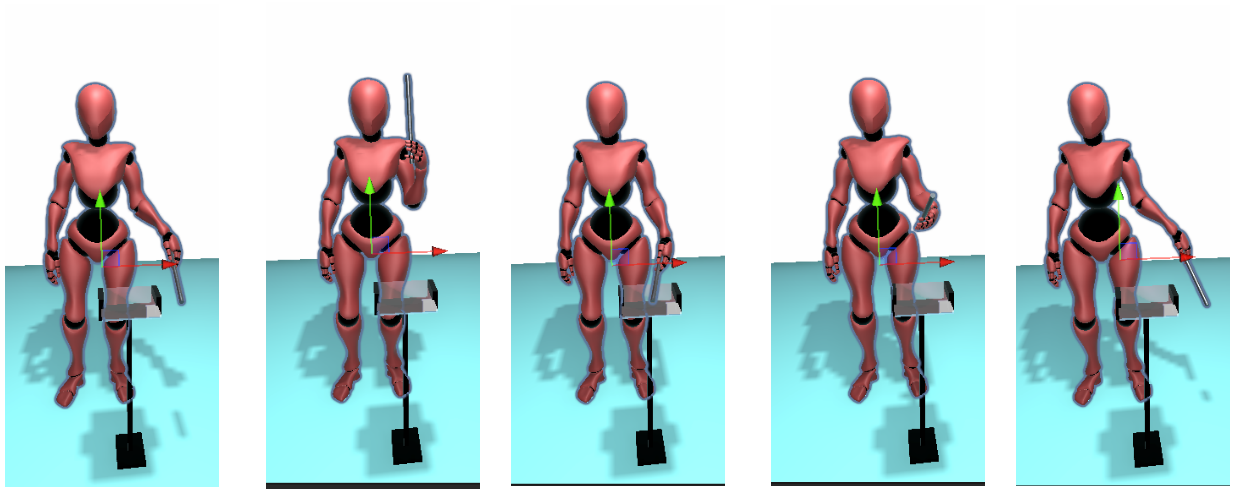

Figure 2.9: Keyframes created for smooth animation of striking motion

2.4 Programming in Unity and SteamVR

We used Unity and Visual Studio (C #) for development and the HTC Vive Pro Eye headset to carry this experiment out. Vive headsets require the SteamVR package for development. The following subsections detail the smaller working parts required to put this experiment together.

2.4.1 Velocity Function

For self-interaction, we also needed to measure the force with which the participant struck the target object. To calculate the force, we required velocity of impact as well. Using Unity's physics engine and the rigidbody's properties, velocity could be easily obtained. But, these measures were very inaccurate and oftentimes gave values of the order 10^{-15} to 10^{-22} even for quick and big displacements. We instead coded a custom function that calculated the position of the tip of the rod per frame, its velocity per frame and then averaged the values obtained from the five frames before the impact. Averaging was done to reduce noise and avoid data loss during random frame drops.

2.4.2 Friction coefficient and physic materials

In Unity, several behaviors of physical objects can be simulated including the surface properties that guide how objects interact with one another. This is done by manipulating the rigidbody's physic material properties. By changing the various friction coefficients and bounciness, more realistic impact interactions can be simulated. We set the physic properties as shown in Figure 2.10.

In addition to this, we attached the rod to the controller using a spring joint so as to reduce the number of times the rod got detached from the controller due to the force of the motion. However, for situations where it would still get detached, a re-spawn function was mapped to the controller's trigger button. This reset the position of the rod to the end of the controller whenever pressed.

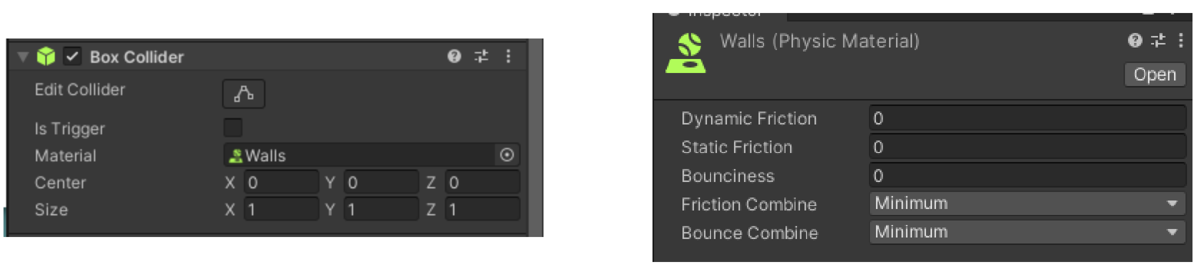

Figure 2.10: On the left is the target object's rigidbody properties and on the right is the specific settings of the physic material chosen

2.5 Experimental Design

2.5.1 Variables

The basic setup of the task involved a target object being struck by either a third party/agent or by the participant. The appearance and associated impact sounds were used by the participant to identify the material of the target object.

Independent Variables that were considered in the study:

- Target object's visual identity (Glass, Metal, Plastic or Wood)
- Target object's auditory identity (Glass, Metal, Plastic or Wood)
- Agency (Agent strikes the object/Participant strikes the object)

Visual identity of the object refers to what the object looks while the auditory identity of the object refers to what the object sounds like.

Dependent Variables here were:

- Response - the option selected for the presented combination of the visual and auditory properties
- Speed/force of impact
- Response time - the amount of time taken for them to select an option

2.5.2 Conditions

For this study, we considered four materials—wood, metal, glass, and plastic—resulting in 16 experimental conditions. Among these, four conditions featured matching auditory and visual identities, while the remaining 12 involved mismatched pairings.

Additionally, we aimed to investigate whether agency influenced material classification. To examine this, the experiment was divided into two parts. In one part, an agent held the rod and struck the target object, with participants merely observing the interaction before selecting the most suitable option. In the other part, the virtual rod was attached to a VR hand controller, allowing participants to strike the object themselves. These two conditions are referred to as agent-interaction

	Agency	
Audio-visual Identity	**Agent**	**Self-Interaction**
Matched	Condition set 1	Condition set 2
Mismatched	Condition set 3	Condition set 4

Table 2.2: Summary of different condition sets in the experiment

and self-interaction, respectively.

The 16 conditions were repeated under the 2 modes of interactions giving rise to 32 conditions (4 materials x 4 sound types x 2 modes of interaction) which can be divided into 4 sets of conditions as summarized in Table 2.2. During all conditions, the participants were presented with a target object and asked to identify its material. The target object was struck by either a third party/agent (Conditions sets 1 and 3) or by the participants themselves (Condition sets 2 and 4) using a rod. The sound upon impact would either match the target visually (Condition sets 1 and 2; for example: it looks like wood and sounds like wood) or differ from it (Condition sets 3 and 4; for example: it looks like wood but sounds like glass).

2.5.3 Procedure

Using a VR headset, in each trial, we presented a thin square plate/slab (target object) to the participant. In the simulation, a virtual metal rod was used to strike the target object and an impact sound played on collision. Based on the sound and appearance, they were asked to categorize the material of the target object as one of four options: Glass, Metal, Plastic or Wood.

The participants were briefed on the experiment and were instructed to use the familiarization task (which is discussed more in detail later in this section) to get used to the virtual environment, the sound samples and the task. For the agent interaction task, they had to carefully observe the agent striking the object. In the self-interaction task, where the metal rod was attached to the VR hand controller, they were asked to strike the object just once (to mirror the other condition). Figure 2.12 shows the different elements of these two modes of interactions.

After each trial where the target object was struck, a dialog box appeared (Figure 2.11), prompting the participant to use a virtual laser pointer to select which option would best categorize the material. The option selected for the particular stimulus and the decision time i.e. the time it took for them to categorize the material, was recorded.

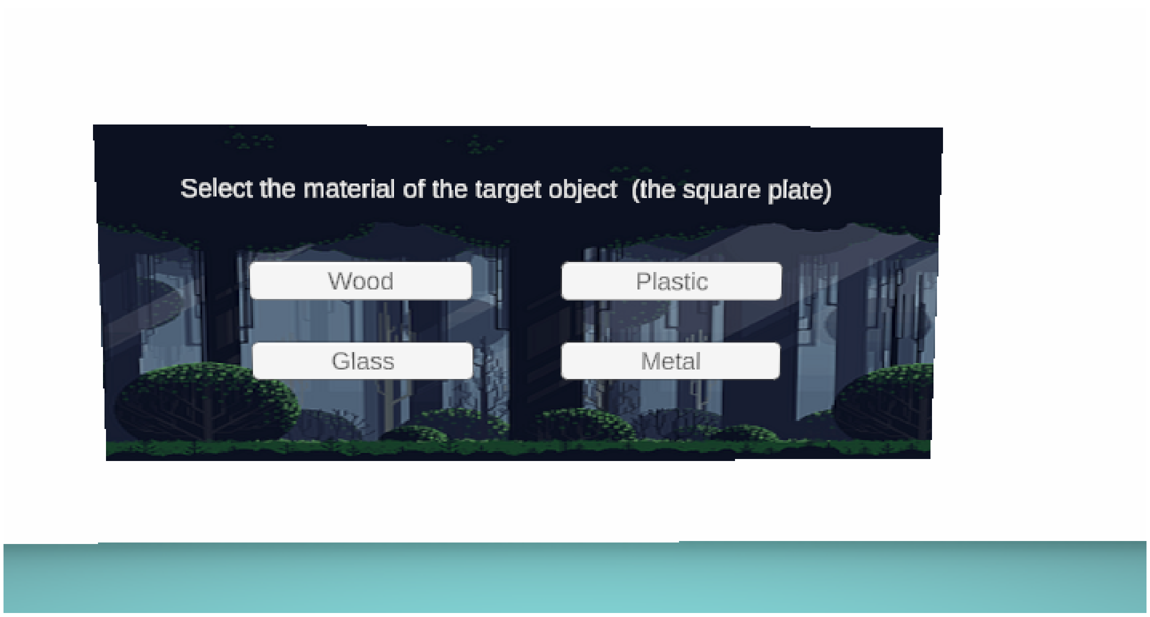

Figure 2.11: The dialog box prompting the user to classify the previously shown object's material

After all the trials were completed, the participants were asked a few questions about their experience. Those questions were:

- How did you find the experiment?
- Were there any sounds you found distinct or any pairs you found very confusing to distinguish between?

- Between the two modes of interaction, did you find one to be easier than the other? Do you think you did better in one as compared to the other?

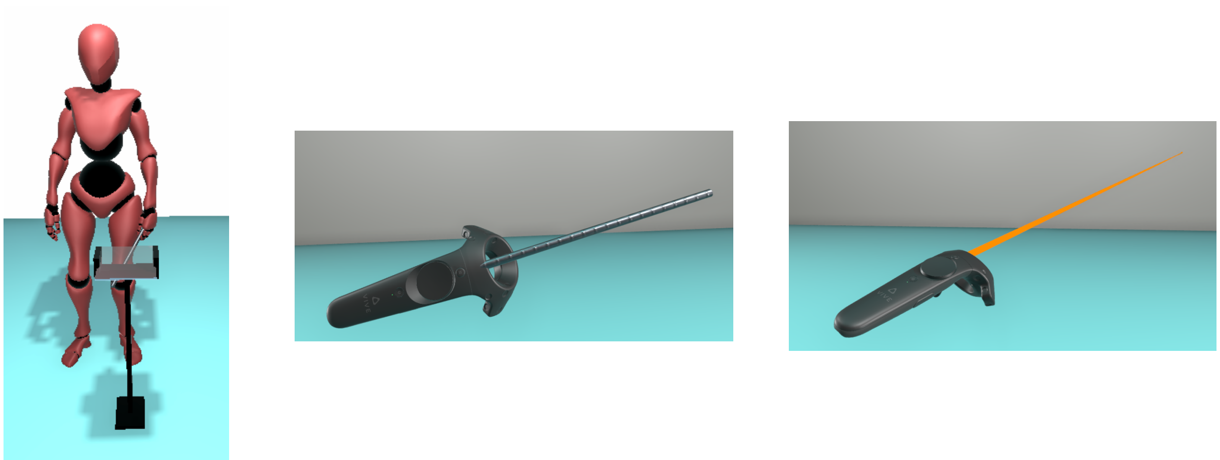

Figure 2.12: Elements of the agent-interaction and self-interaction parts. On the left, is the agent used in the agent-interaction parts. In the middle is the virtual rod attached to the hand controller as seen by the participant. On the right is the laser pointer attached to the other hand controller.

Block order and distribution

Each of the 16 condition was repeated 10 times to get a good sample of data leading to a total of 160 trials. Since a block of 160 trials is too long and prone to get tiring, the trials were split into two blocks of 80 trials presented in an randomized order, where each condition appeared five times. This was done for both agent-interaction and self-interaction i.e. 2 blocks of 80 trials for agent-interaction and 2 blocks of 80 trials for self-interaction with each block taking approximately 5-8 minutes.

During pilot experiments, participants commented that the self-interaction tasks were more engaging as it involved active participation as opposed to mere observation. So, when considering the block order, we decided to alternate between the two types of blocks. Half the participants were tested in the order of agent-self-agent-self while the other half were tested in the order of self-agent-self-agent blocks. This would also balance any order effects seen.

Familiarization task

Before testing, each participant completed a familiarization task followed by a practice task. In the familiarization task, four target objects with different visual textures were arranged in front of the participant (Figure 2.13). The impact sounds associated with each object matched its texture—for example, a wood texture was paired with a wood impact sound, and a glass texture with a glass impact sound. The purpose of this task was to familiarize participants with the experimental setup, the sounds used, and the visual appearances of the materials. During this phase, participants used a rod to strike the target objects and acclimate themselves to the corresponding sounds. This was followed by a short practice session consisting of 48 trials, with each condition presented three times.

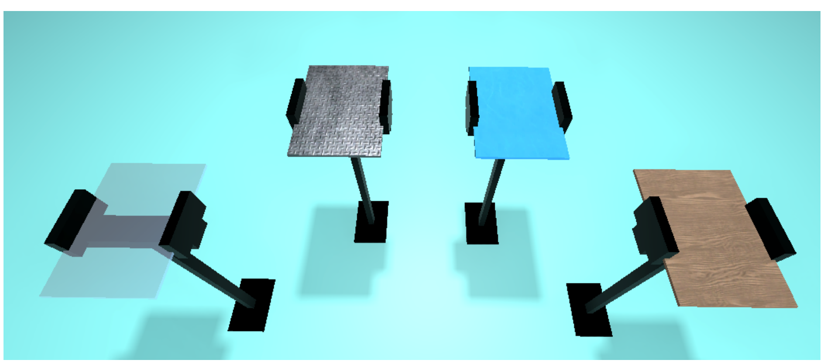

Figure 2.13: The Familiarization task. It shows the target objects used during the blocks and allows the users to familiarize themselves with both the appearances and associated impact sounds.

Given the confusing nature of the mixed audio-visual stimuli, the participant spent a minute or two between the blocks, in the familiarization task to reacquaint themselves with the reference sounds and materials to avoid any incorrect associations that may have formed during the trials.

Chapter 3

Results

To recap, for the four materials being considered, we obtained sixteen conditions comprising of the various combinations of four textures and four impact sounds. Each condition was then repeated twenty times, for a total of 320 trials. In half the trials, the participant observed an agent interact with the target object, and in the remaining trials the participant interacted with the object directly. A total of fifty-three participants were tested.

3.1 Data Pre-processing

3.1.1 Elimination criteria for dropping datasets

During data collection, it became apparent that some participants were not attentive during the trials while others struggled to perform the task. In order to segregate the datasets of the inattentive participants from data of the attentive participants (whether they were struggling or not), an elimination criteria was developed.

Out of the 16 experimental conditions, four involved matching auditory and visual properties (the matched condition). Each of these conditions was presented 20 times, resulting in a total of 80 trials where the impact sound and visual appearance were congruent—40 trials during agent-interaction and 40 during self-interaction.

Since participants had four response options per trial, random guessing would yield a 25% probability of selecting the correct answer. To exclude participants who were either guessing or not paying sufficient attention, we set a performance threshold of 15 correct responses out of 40 trials (slightly above chance). Participants who scored below this threshold in either of the two interaction conditions were eliminated from the dataset. Based on this criterion, 11 datasets were removed, leaving data from 42 participants for analysis. Appendix A.2 provides plots illustrating individual participant data distribution.

3.1.2 Outlier analysis and Winsorizing

Participants were informed that the time spent on each trial was being recorded, which led to shorter trial durations. However, analysis revealed the presence of outliers, typically resulting from misclicks, technical issues with the hand controller, or similar factors. Before conducting the analysis, these outliers in the recorded decision times needed to be addressed.

To handle these outliers, the data was winsorized - i.e. values beyond the 95^{th} percentile were replaced by the value of the 95^{th} percentile and the values below the 5^{th} percentile were also replaced such that they equaled to the value of the 5^{th} percentile. This way the extreme values on both ends were replaced with less extreme values from the same participant's dataset.

3.2 Self-interaction vs Agent interaction

To study the effect of interaction, we recorded participant responses and decision time, in both modes of interaction. After they completed the experiment, they were asked a few follow-up questions. One of them was if they found one mode of interaction easier than the other.

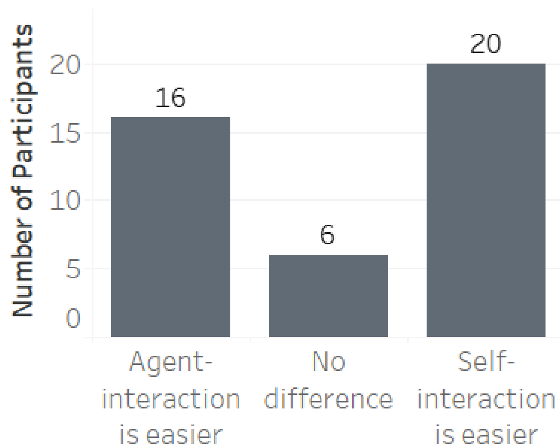

Figure 3.1: Bar chart illustrating the distribution of participant preferences after testing.

As shown in Figure 3.1, 16 participants found agent interaction easier than self-interaction, noting that they did not have to divide their attention between performing the action and hearing the sound. In contrast, 20 participants preferred self-interaction, stating that their active involvement made it easier to focus on the task. They also reported that having control over the rod allowed them to better anticipate the impact event and adjust parameters such as timing and volume. The remaining six participants did not perceive a difference between the two interaction modes.

We performed a set of binomial tests (with the confidence interval set at 95%) to determine whether the overall preference was statistically significant for any category. When testing for preference for self-interaction (where the alternative hypothesis stated that participants were more inclined to prefer self-interaction), we obtained a p-value of 0.0383. Thus, we can conclude that, statistically, participants demonstrated a significant preference for self-interaction.

The following subsections compare the obtained data to examine how participants' responses varied across different modes of interaction.

3.2.1 Results of the matched condition

For the matched conditions, we compared the overall response by calculating the percentage of trials where each participant categorized the target correctly and the average time taken to do so. This data is represented in Figures 3.2 and 3.3.

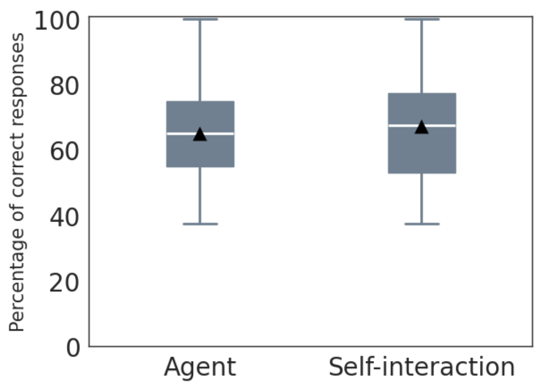

Figure 3.2: Percentage of correct responses for matched condition

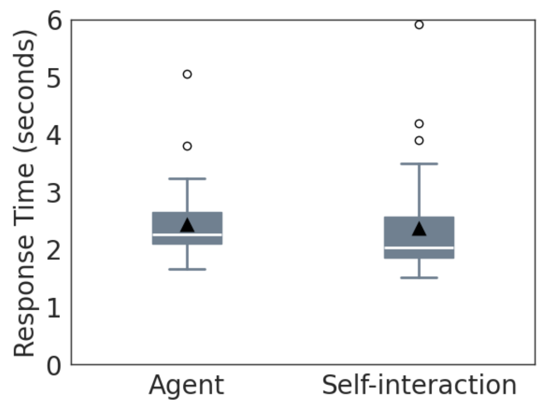

Figure 3.3: Average time for matched condition

The Box plots here (and throughout this thesis) represent the percentage of responses correctly categorizing the matched conditions across all participants. The line across the boxes show the median and the triangle represents the mean. The bottom of the box is the 25^{th} percentile and the top is the 75^{th} percentile. The whiskers typically extend to 1.5 x the Interquartile Range, beyond which data points are considered outliers (represented by circles).

While there is variation per participant, based on Figure 3.2 and 3.3, we can see that about 70% of the trials were correctly matched and the average decision time was approximately 2.5 seconds - for both agent-interaction and self-interaction. The statistical analysis can be seen in section 3.2.3

below.

3.2.2 Results of Agent v/s Self-interaction when option selected matched the audio played

Unlike in matched conditions, when the appearance and sound conflict with each other, there is no correct response to select. One way to analyze how agent interaction differs from self-interaction is to examine how participants' responses vary when they categorize the target object's material based on the impact sound played during these mismatched conditions. Figure 3.4 compares the percentage of times participants classified the material based on sound for the two modes of interaction, while Figure 3.5 compares their decision time. Based on these charts, we can see that there were no major differences between the two modes of interaction.

We see a similar result when comparing the responses of participants when they classified the object's material based on its visual appearance (please see Appendix A.3 for these graphs).

In all cases, the percentage of responses matching the auditory cue remained consistent across both interaction modes, despite participants' reported preferences after the experiment. In this study, the action performed remained the same for both modes of interaction. This suggests that as long as the observed or executed action is identical, the presence of agency in the interaction does not influence the perception of the object or its material.

3.2.3 Statistical comparison between Agent interaction and Self-interaction

Since our variables are categorical, we can analyze whether the frequency of responses (classifications based on auditory properties, visual properties, or neither) differed significantly between agent-interaction and self-interaction trials. To assess whether the mean response varied between

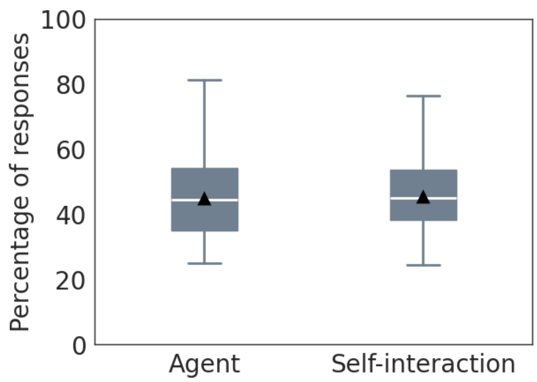

Figure 3.4: Percentage of responses when selected option matches the audio

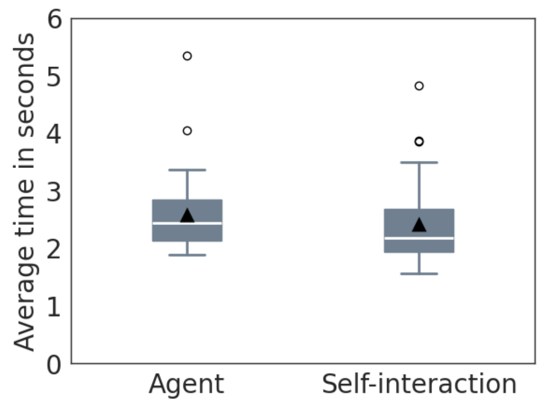

Figure 3.5: Average time taken when selected option matches the audio

these two interaction modes, we conducted a paired-samples t-test. Using Python's Pandas and Scipy packages, three paired samples t-tests were performed which found no significant difference in participant response when they classified based on the sound, appearance or neither between the two modes of interaction ($t_{(41)} = -0.575, -0.230, 0.187$; p-value = 0.568, 0.819, 0.853, respectively). Thus, there is no evidence that interaction affected material judgments.

Further, observing Figures 3.3 and 3.5, we can see that the average decision times do not depend on interaction or condition. Thus, we do no further investigations based on decision times in this thesis.

3.3 Auditory Properties v/s Visual Appearance

When the auditory properties do not match the visual appearance, a person can respond in one of four ways when classifying the target object's material. They may match the material based on its appearance or its sound, select a third option randomly, or choose a third option consistently for a given combination (based on neither the sound played nor the appearance shown).

To evaluate how materials are perceived when there is an audiovisual mismatch, we plotted the data to check if participants favored one modality over the other or if they selected an option based neither on the visual component nor the auditory component.

3.3.1 Results for when option selected matched audio/visual/neither

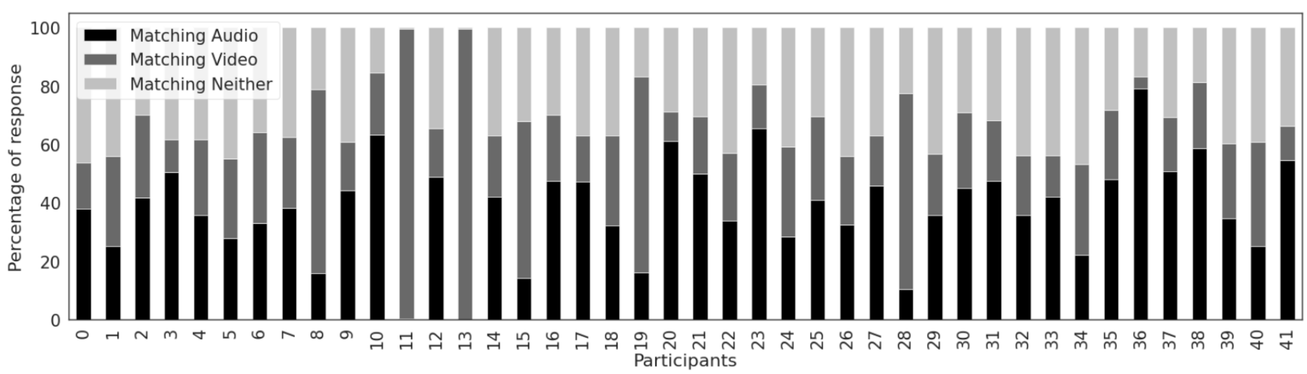

Figure 3.6: Percentage distribution of responses show the percentage of times each participant has classified the material based on either auditory information, visual information or selected a third unrelated option

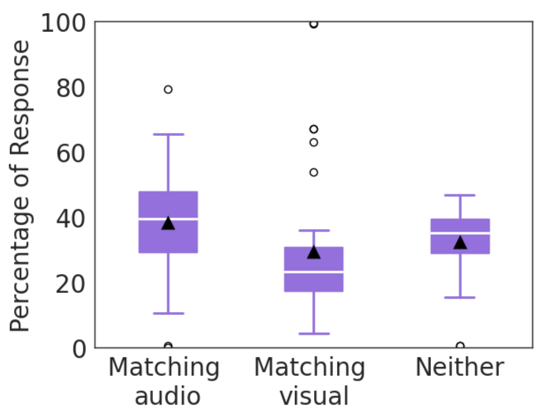

Figure 3.7: Distribution of participant response when divided based on whether it matched the sound, appearance or neither of the two

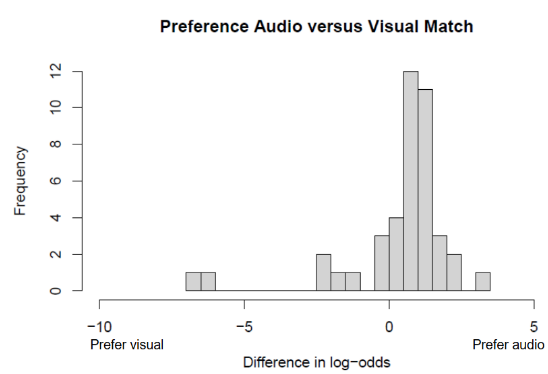

Figure 3.8: Difference in Log-odds of participant preference

Figure 3.6 shows the percentage of responses matched to the sound, to the appearance or to neither of the two options for each participant. The average response of all participants is summarized in Figure 3.7, which shows that they tend to classify based on the audio signal first, followed by a preference towards an unrelated third option over the visual component.

Signal preference can also be analyzed by calculating the log-odds. Per participant differences in log-odds were obtained by calculating: log_e(odds choosing audio) - log_e(odds choosing video). The log-odds were calculated for each participant from the random intercepts of logistic regression models for audio and visual matches. Figure 3.8 shows a histogram depicting this information, where the y-axis is the number of participants and the x-axis is the difference in log-odds of participant responses between responses matched to sound and matched to appearance. We can see that very few participants tend towards visual information when conflicting auditory cues were presented.

3.3.2 Statistical analysis of signal preference

Since interaction caused no significant differences in responses, we can analyze the data without considering it by combining observations across interaction modes. This results in 20 data points per condition per participant, for a total of 320 trials. We performed two repeated-measures two-way ANOVAs, with texture and sound as the independent variables. The dependent variable for the first ANOVA was the percentage of responses in which the participant categorized based on sound, while for the second ANOVA, it was the percentage of responses based on visual texture.

We observe high F-values and very low p-values in both cases. For example, when examining the effect of the interaction between the two input stimuli, $F_{(9,369)} = 29.3483$, p-value = 2.0263 x 10^{-38}, and $\eta_p^2 = 0.417$ when the dependent variable is the percentage of participant responses matched to audio. Similarly, when the dependent variable is the percentage of participant responses matched to video, $F_{(9,369)} = 63.4160$, p-value = 1.6870 x 10^{-69}, $\eta_p^2 = 0.607$.

The high f-value and small p-value could be due to violation of the assumption of sphericity of a repeated measures ANOVA. We ran the Mauchly's sphericity test and confirmed that this assumption is not being met. Thus, using the Greenhouse-geisser correction, we calculated the epsilon correction and recalculated p-values but found no major changes in the outcomes. For all corrected values see appendix section A.4.

These ANOVAs tell us that there are significant effects of texture and sound on response. To better understand the effects, we then broke down the two-way ANOVA into a sets of 4 one-way ANOVAs, where the appearance was kept constant while the audio factor was varied.

Consider one such one-way ANOVA case, where we keep the glass texture fixed and see how varying the sound effects how participants categorized the materials. The independent variable is the audio played (ignoring the matched condition we get 3 sounds: metal, plastic and wood) and dependent variable is the percentage of times participants matched the presented stimulus to the sound. Figures 3.9 shows the ANOVA results ($F_{(2,82)} = 39.8627$ and p-value = 8.0623 x 10^{-13}) and below that is the Tukey pairwise comparisons between each of the categories (using Family-wise error rate corrections with $\alpha = 0.05$). We see that wood is selected significantly more than metal and plastic (which is also visualized in Figure 3.10). This means that participants are more likely to classify based on the sound when presented with a wood impact sound despite the object appearing to be made of glass.

For cases where participants matched responses to sound, appearance, or neither, we performed one-way ANOVAs and found that the main effect was significant in all cases (the associated tables and pairwise comparisons are provided in Appendix A.4). Thus, for every texture or sound, the likelihood of an auditory or visual match in a conflict scenario depended on the identity of the

```
When checking the responses matched to the Sound
                 Anova
===================================
      F Value Num DF  Den DF Pr > F
-----------------------------------
Audio 39.8627 2.0000 82.0000 0.0000
===================================
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Metal	Plastic	0.1667	0.9884	-2.5482	2.8816	False
Metal	Wood	7.0476	0.0	4.3327	9.7625	True
Plastic	Wood	6.881	0.0	4.166	9.5959	True

Figure 3.9: Repeated measure one-way ANOVA and post-hoc pairwise test when texture is fixed to glass and sound is varied

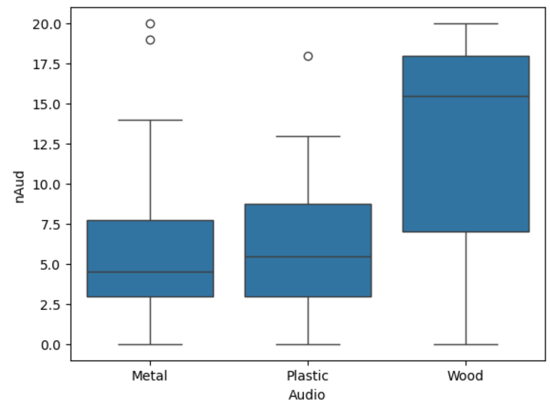

Figure 3.10: The frequency of responses matched to audio (nAud) when texture is glass and sound is varied

conflicting cues. Therefore, we will focus on post-hoc comparisons for each condition to better understand the observed effects, which are summarized in the next section.

3.4 Analysis of individual testing conditions

When participants select a third, unrelated option for a target's material, they may either choose randomly or consistently select the same alternative for a given combination. The latter resembles the McGurk effect, in which hearing one sound while seeing a different mouth movement reliably leads to the perception of a third, distinct sound.

To determine whether a similar phenomenon occurs in material perception, we analyzed participant responses across all 16 conditions. Figure 3.11 illustrates the percentage distribution of responses for each condition, while detailed box plots for individual conditions are provided in Appendix A.5.

Based on the graphs in Figure 3.11, we see that the likelihood of a response being based on the

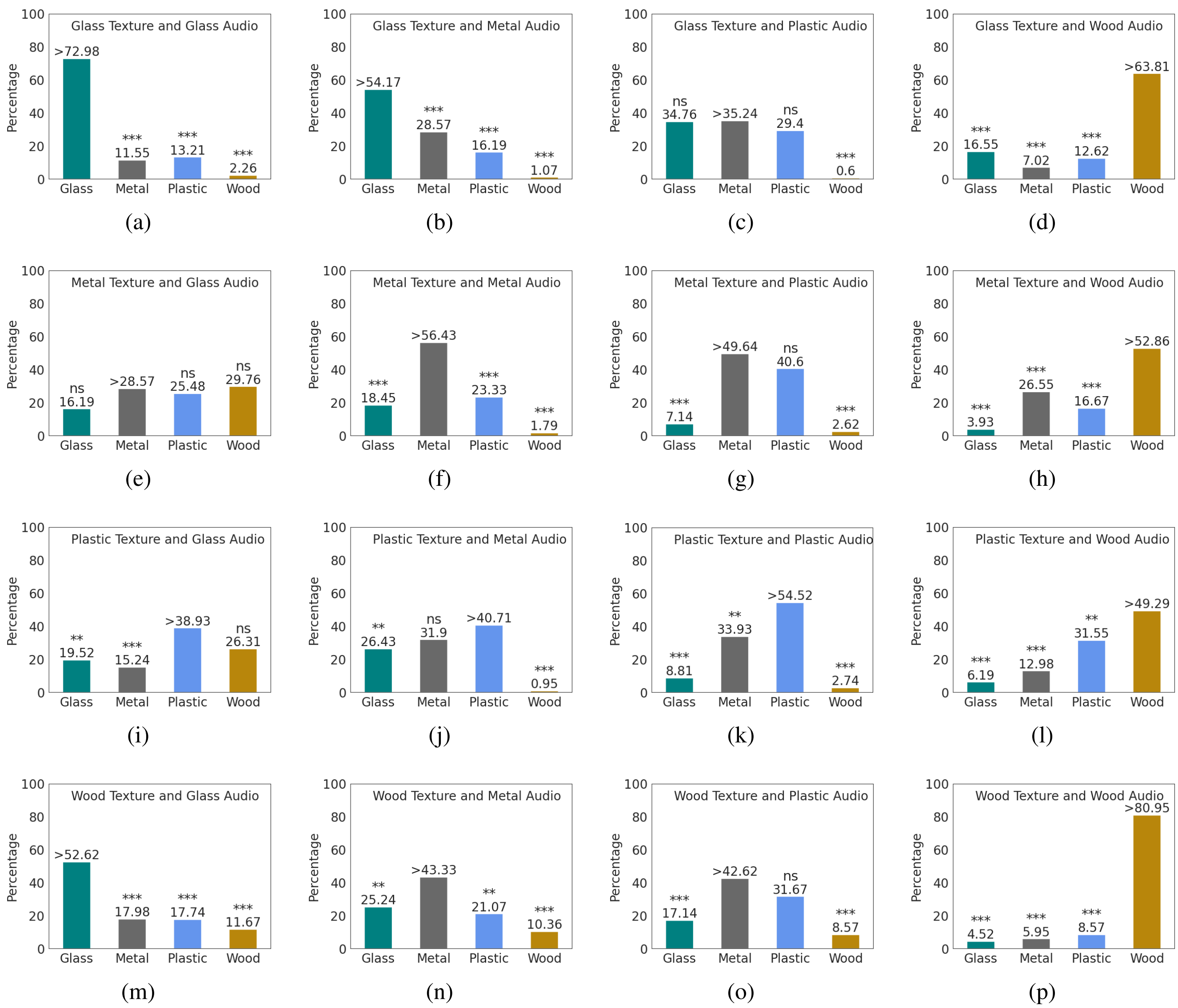

Figure 3.11: Participant responses in all conditions. The peak response for a given condition is indicated using - “ $>$ ” symbol before the percentage label. The significance values of the differences between the peak response and the other categories have been depicted using: ns as $p >= 0.05$, “$*$” $= p-value < 0.05$, “$**$” $= p-value < 0.01$ and “$***$” $= p-value < 0.005$. The pairwise test tables are in Appendix A.5.

	Glass Audio	Metal Audio	Plastic Audio	Wood Audio
Glass texture	-	Vision Dominant	Anomaly	Audio Dominant
Metal texture	Anomaly	-	A/V Tie	Audio Dominant
Plastic texture	Anomaly	A/V Tie	-	Audio Dominant
Wood texture	Audio Dominant	Audio Dominant	Anomaly	-

Table 3.1: Results of the statistical analysis. The cases where participants were equally likely to select based on the sound or appearance are labeled 'A/V Tie'.

audio, visual or other option vary on a case-by-case basis. Using the Tukey HSD pairwise tests we checked to see if the differences in responses are statistically significant. The patterns of the significant results are summarized in table 3.1 and the individual statistical tests for each condition can be found in Appendix A.5.

For each condition, we generated bar charts and box plots and conducted pairwise Tukey HSD tests with Familywise error rate corrections ($\alpha = 0.05$). We then analyzed pairwise comparisons to identify which response option was selected significantly more frequently than the others. Based on the comparisons with the predominant choice, we categorized each condition as vision/texture dominant, audio dominant, audio-visual tie, or ambiguous/anomalous.

Vision-dominant cases were those in which participants categorized based on appearance significantly more often than other options. Audio-dominant cases were those where participants relied significantly more on sound. Additionally, we introduced two other labels, Ambiguous/anomalous cases and 'Audio Visual Tie' cases. The Ambiguous/anomalous cases exhibited greater variability, often reflecting confusion between different categories. The 'Audio-Visual Tie' cases were where classification rates based on sound and texture did not differ significantly.

A few interesting observations can be immediately inferred from the table. In most cases where wood was an element—either in sound or appearance—participants tended to categorize based on sound. In contrast, most cases involving plastic as an element resulted in some degree of ambiguity. Notably, there was only one strongly visually dominant case: glass texture coupled with a metal impact sound.

3.4.1 Audio dominant cases

A closer examination of the results reveals that all audio-dominant cases involved the wood material. When presented with a wood impact sound, participants predominantly classified the target as wood, regardless of its visual appearance.

Examining graphs (e), (i), and (j) from Figure 3.11, we identify two anomaly cases and one audio-visual tie case, where some degree of confusion between glass and metal sounds is evident. This observation is supported by the spectral analysis and distance calculations in Section 2.2.4 and Figure 2.7, which indicate a high degree of spectral similarity between glass and metal sounds. These findings highlight the difficulty in distinguishing between the two materials based solely on auditory cues. However, as seen in graphs (m) and (n), when a wood texture is presented, participants' ability to differentiate between glass and metal impact sounds improves significantly, leading to the classification of these cases as audio-dominant.

3.4.2 Anomaly cases

As mentioned earlier, anomaly cases refer to instances where the predominant response did not align with either the visual appearance or the sound, often resulting in considerable confusion among multiple options. Figure 3.12 presents the box plots for these anomaly cases. Each anomaly case is unique:

- Figure 3.12 (a) - when presented with glass texture and plastic audio, participants were uncertain on whether to classify the material as glass, metal or plastic

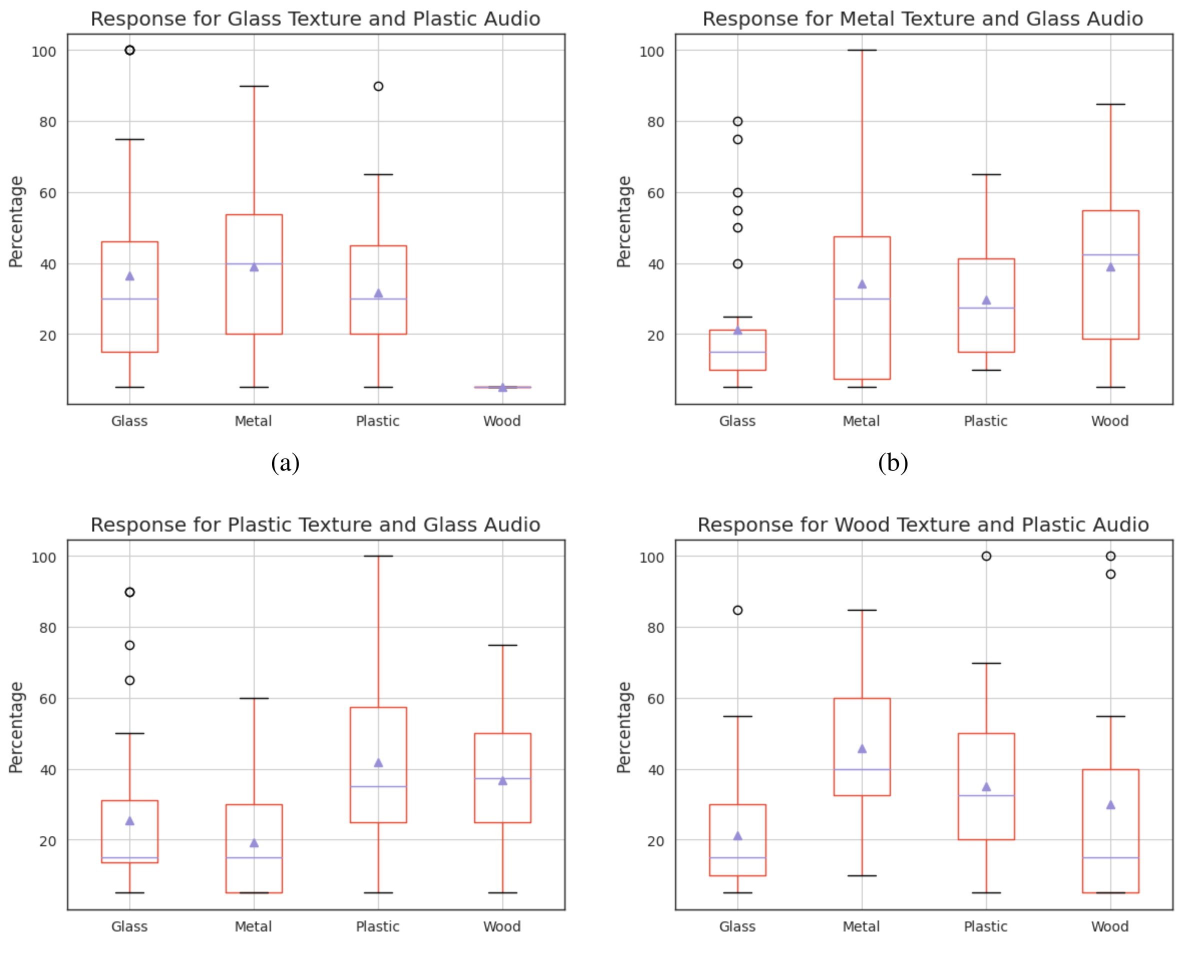

Figure 3.12: Box plots for the Anomaly cases

- Figure 3.12 (b) - when presented with metal texture and glass audio, participants were equally likely to categorize the material metal, plastic or wood.
- Figure 3.12 (c) - when presented with plastic texture and glass audio, participants were confused between all the categories.
- Figure 3.12 (d) - when presented with wood texture and plastic audio, participants selected metal and plastic the most often.

Figures 3.12 (a), (b) and (c) show that participants response was scattered across multiple categories, where as figures 3.12 (d) shows a tie between an expected option and an unexpected option.

Figure 3.12 (d) shows the strongest case of a potential audiovisual illusion, where despite being presented with a combination of wood texture and plastic audio, participants categorized it as metal the most often (42.62%) followed by plastic. We also see metal being selected more often than expected when presented with glass texture and plastic audio (Figure 3.12 (a)).

But this is not simply metal being selected as a type of default option in times of confusion, as this behavior is not reflected in Figure 3.12 (c). In fact, in Figures (b) and (c), we see wood being selected more often then expected instead.

The presence of these anomaly cases, as well as the audio-visual tie cases, provide evidence that audio-visual illusions can occur in terms of material categorization.

Chapter 4

Discussion and Conclusion

4.1 Summary

In this study, a VR headset was used to present a virtual target object that was struck with a rod while an associated impact sound was played. The participant's task was to classify the material of the object. We aimed to investigate three key questions: whether participants favor one sense over the other when presented with mismatched audiovisual stimuli, whether there is potential for audiovisual illusions to occur, and whether agency influences material judgments.

Based on the data obtained and the analysis performed in the previous section, we can summarize our findings as follows. First, there was no difference in classification response or classification time between the agent interaction trials and the self-interaction trials. Second, initial graphical analysis suggests that participants most often classified materials based on sound. Statistical analysis confirmed this trend and indicated that cue preference depended on the specific combinations of audio and texture presented. However, there were more audio-dominant cases than visual-dominant ones, leading to the conclusion that material perception is influenced more by auditory cues than by visual cues. Finally, in four of the sixteen conditions, we observed behaviors suggesting the potential occurrence of audiovisual illusions, similar to the McGurk effect, when auditory properties did not match visual properties.

4.1.1 Effect of Agency

Research on sound source localization (Brimijoin & Akeroyd, 2014; Kondo et al., 2012) suggests that allowing participants to move freely in virtual space typically improves accuracy of the tasks performed. Inspired by these findings, we sought to determine whether permitting participants to interact with an object themselves would similarly enhance their material judgment.

After completing the task, participants were asked whether they had a preference between the two modes of interaction. Some preferred the agent-interaction mode, as the agent consistently struck the rod with the same force and speed, ensuring uniform sound volume and making it easier to anticipate the impact event. Additionally, some participants noted that in self-interaction mode, they had to divide their attention between performing the action and observing the object, which they found challenging. Others favored the self-interaction mode, appreciating the freedom to control the force, speed, and timing of their strikes. Given that the force of the impact also determined the loudness of the sound, they were able to choose how loud or soft the impact had to be (and thus select their desired volume for the audio stimulus). And finally, there were a few participants who did not have a preference and found both modes of interaction to be equivalent.

Our data indicated that although most participants had a personal preference for a particular mode of interaction, their performance did not vary between conditions. We compared the percentage of correctly identified responses in the matching condition (Figure 3.2), as well as the percentage of responses categorized based on auditory cues, visual cues, or neither during mismatched conditions (Figures A.10 and A.11 in the appendix). Additionally, we analyzed average response times for these cases. Both graphical and statistical analyses revealed that interaction mode had no significant effect, as the average response accuracy and response times were statistically indistinguishable between agent-interaction and self-interaction trials. Furthermore, an analysis of individual responses, grouped according to participants' stated preferences, showed no differences in performance between the two modes of interaction."

This could be because the action being performed and observed were very simple and similar, in which case agency may not matter. That is, to keep the agent interaction and self-interaction trials comparable, we restricted the degree of freedom that free interaction with an object usually entails. Even though the goal was to compare the sounds made by the target object, participants were not allowed to tap multiple times or do any other action like scraping the rod across the surface, dropping it on the floor etc. It is possible that allowing participants to interact with the object with more freedom may lead to different results.

4.1.2 Auditory dominance over visual

Given the multi-sensory nature of material perception, it is hard to say if one sense is inherently more important than another. Fujisaki et al. (2014) extensively studied this with 48 audiovisual combinations made from 13 material categories (six visual stimuli and seven auditory stimuli). They concluded that auditory inputs had more dominance over visual inputs during cue-conflicts. However, their study was performed using an LCD monitor - a non-immersive device. This makes it hard to generalize their results. The lack of depth cues could have caused the visual stimulus to be less informative and reliable. This could be one of the reasons why participants relied more on the auditory stimulus.

Anderson et al. (2016) conducted experiments focused on understanding multimodal integration of audiovisual information for material perception in VR. The results for their cue conflict conditions suggested that neither of the two senses have dominance over the other (except when visual cues were diminished, in that case the participants relied on the auditory cues). However, given the small number of cases tested, they stated that expanding on their experiment would give a clearer idea.

Our study addressed these limitations and helped reconcile the findings of the two previous studies. By conducting our experiment in VR, we created a more immersive environment than that of Fujisaki et al. (2014). The inclusion of depth cues and 3D spatial sound better simulated real-world conditions, enhancing the generalizability of our results. Furthermore, our experiment was more extensive and comprehensive than that of Anderson et al. (2016). We tested four materials and created 16 conditions by examining all possible combinations of sounds and textures, allowing for a thorough investigation of the effects of mismatched audiovisual stimuli under various conditions.

All audio-dominant cases involved the wood material, likely because participants found the wooden impact sounds to be the most distinctive. The distinctiveness was so strong that even when a transparent glass plate or a metal plate was paired with a wood impact sound, participants showed a strong tendency to categorize the object as wood.

Further, based on our spectral analysis, we could see that glass and metal sounds were very similar, while plastic and wood shared more similarities. In our data, we see this manifesting as confusion when distinguishing between glass and metal sounds (as seen in graphs 3.11 (e), (i), (j)). This is similar to results reported in the study by Lemaitre & Heller (2012) where they found that it was harder for listeners to distinguish between similar material categories like wood-plastic and metal-glass.

Despite this confusion, participants were able to distinguish between glass and metal sounds when presented with a wood texture. As shown in Figures 3.11 (m) and (n), when a wood texture was paired with either glass or metal impact sounds, participants were more likely to categorize the material based on the auditory cue. This suggests that although differentiating between glass and metal sounds alone was challenging, participants were able to do so when the visual element was distinctly different from their expectations of what the sounds should match.

The only visually dominant case was when glass texture was paired with a metal impact sound. As previously noted, glass and metal sounds were the most difficult to distinguish. This difficulty may explain why, in this case, participants relied more heavily on visual appearance for classification.

There were two cases where participants were equally likely to categorize based on sounds or appearances - plastic texture combined with metal audio and metal texture combined with plastic audio.

Re-examining Table 3.1, we observe a high amounts of ambiguity associated with the plastic material. With the exception of the matched condition and the case where plastic texture was paired with wood audio, all conditions involving plastic were either classified as audio-visual ties or anomaly cases. Notably, the only anomaly or audio-visual tie case that did not involve plastic—either as the auditory or visual component—was the condition where a metal texture was paired with glass audio. A possible explanation for this ambiguity is that, despite plastic being a common material in everyday life, it exists in many different forms, leading to significant variability in its appearance and sound. The plastic sample used in this experiment was acetal plastic in the form of 1/4" and 3/8" thick slabs, whereas plastic is more commonly encountered in thinner variations, such as bottles, cutlery, and plates. As a result, the plastic sounds used in this study may have differed from participants' expectations, contributing to increased confusion during mismatch conditions.

Overall, we see more audio dominant cases than visual dominant conditions. Based on this, we can conclude that audio has more dominance when matched with conflicting visual textures for material categorization. Thus, our results are in-line with the conclusions drawn by Fujisaki et al. (2014) who investigated the same phenomena using non-immersive displays (LCD monitor).

There are several possible reasons why sound may have dominance over visual inputs in material categorization. One explanation is a lower trust in visual appearances due to the increasing prevalence of faux materials in real life, which are often used to disguise lower-quality materials or conceal faults in products. Another possibility is the lower level of immersion in VR compared to reality, where underperforming graphics may lead participants to rely less on visual information and compensate by relying more on other senses.

4.1.3 Potential for Audiovisual illusions

In Chapter 1.4 we discussed several cases of how incongruent sounds and visual appearances can skew perception (Gau & Noppeney, 2016; Hidaka & Ide, 2015; Teramoto et al., 2012) and also give rise to audiovisual illusions (Hirst et al., 2020; McGurk & MacDonald, 1976; Moris Fernandez et al., 2017; Shams et al., 2005, 2000). We see a similar potential in this study when incongruent sounds and visual textures are presented for material categorization tasks. This is demonstrated by the existence of our anomaly cases. The anomaly cases we observed in our study were:

- Metal texture and glass audio (where participants were equally likely to categorize it as metal, plastic and wood)

- Plastic texture and glass audio (where participants were equally likely to categorize it as any of the presented options)
- Glass texture and plastic audio (where participants were equally likely to categorize it as glass, metal or plastic)
- Wood texture and plastic audio (where participants were equally likely to categorize it as metal or plastic)

These cases were classified as anomaly cases because participants did not exhibit a clear preference for any single material, and often their selections did not consistently align with either the sound or the appearance. Participant responses were distributed across multiple options, with the exception of the wood texture paired with plastic audio (where it was a tie between 2 options). This condition presented the strongest evidence for a potential audiovisual illusion, as participants most frequently categorized it as metal, followed by plastic. One possible explanation for this preference toward metal is the use of a metal rod to generate the impact sounds in this experiment. Although the rod remained constant across all recorded sounds, it may have subtly influenced the final acoustic properties. In this case, participants may have been uncertain enough to resort to classifying on the rod's material instead. This could also explain why metal was selected more frequently than expected when glass texture was paired with plastic audio (Figure 3.15 (a)).

But this does not justify the behavior seen for the other two anomaly cases where we see participants selecting wood more often than one would assume. This is a point of interest because looking at the audio dominant cases, it is evident that wooden impact sounds are very distinct. Despite this, as shown in Figures 3.15 (b) and (c), when presented with glass impact sounds coupled with plastic or metal textures, participants categorized the material as wood more frequently than expected.

One thing to note about the effects seen in this study is that they may be specific to the materials used to generate the sounds. If we change factors like the dimensions, hollowness, the specific type of metal/plastic/wood used etc, it will cause a change in the impact sounds generated and may also change the effects seen here. While the specific effects may shift and vary, based on the results of this study and the study conducted by Fujisaki et al. (2014), it is likely that we will continue to witness some variation of audiovisual illusions. Further, our study is specifically about understanding how audio-visual cue conflicts affect our material categorization in VR. While we managed to replicate one of the major findings of the non-immersive version of this experiment (material categorization being more audio dominant), if we were to change mediums to non-immersive or semi-immersive displays, the exact illusions seen in our experiment may change. More research is needed to understand interactions between signals of different material combinations under different immersive conditions.

4.1.4 Relevance of the study

Real-time visual graphics quality has already been pushed to the limits of current technology, and further advancements will require significant investments in terms of cost, hardware, computational power, and memory. Our research highlights the crucial role of sound in virtual simulations, demonstrating that effective soundscaping can reinforce visual information and enhance immersion. Since improving graphics often necessitates costly hardware upgrades, investing in high-quality sound design presents a more cost-effective alternative that is likely to yield substantial improvements in user experience.

Audiovisual illusions can also be employed for entertainment purposes in the gaming industry, including inducing confusion, distorting perception of the environment, enhancing immersion through cues, or providing hints to players (Garner et al., 2010; Zlobin, 2024). One example is their use in horror or mystery games to create a sense of 'wrongness' by making subtle changes to the environment that unsettle or confuse the player. With respect to our findings, one such case can be altering the sound of a wooden door to instead sound like plastic, making it difficult for players to determine whether they can break it with enough force, as they are likely to perceive it as a metal door. Current research on effective soundscaping is regularly applied to enhance the gaming experience (Cao et al., 2023; Guillen et al., 2021). However, VR game development remains a relatively new field with limited corresponding research. Understanding how to create a more immersive experience is essential, and insights into perceptual manipulation can serve as valuable narrative tools.

Lastly, knowing how sounds interfere or integrate during VR simulations can guide the technical hardware specifications of newer headsets, leading to better hardware, specialized recording equipment, and new sound generation techniques.

4.2 Limitations and future work

Material perception is a multisensory phenomenon influenced by the combined processing of multiple sensory signals. Additionally, individual perception is shaped by various biological factors and personal experiences accumulated over a lifetime. The following subsections discuss important factors that may affect an individual's perception. Since our study does not account for these variables, they also represent limitations that highlight opportunities for future research.

4.2.1 Age and gender

The dataset in this study, is unbalanced between male and female participants. We had an overrepresentation of female participants in our study. This is a potential concern as it is currently unknown if there are any gender based effects with regards to material perception in VR. Additionally, the majority of the participants were undergraduate students around the age of 18-20 years. This is a very young demographic and the results can thus, not be generalized for all age groups.

The affect of age on perception is a an area of research that has always been of interest in the scientific community. The brain relies on information that it has previously encountered to contextualize events it perceives in real time. This means that perception relies on integration of sensory input with the prior knowledge gathered through past experiences. The commonly agreed upon theory is that with increasing age, susceptibility to audio-visual illusions also increases and the intensity or magnitude of the illusions perceived tells us the relative degree by which the brain relies on prior knowledge (Chan et al., 2021; Moran et al., 2014; Petzschner et al., 2015; Seriès & Seitz, 2013). However, recent studies on the effects of age on various visual phenomena have proven that this hypothesis is over-generalizing the effects of age.

Mazuz et al. (2024) published a recent study on the effect of age on three different visual illusions - the Ponzo illusion, the Ebbinghaus, and the Height-width illusions. They concluded that these illusions were the results of different perceptual mechanisms, leading to differences in the overall effect of age across the illusions.

Additionally, existing studies centered on the effects of aging on perception of audiovisual stimuli are usually with regards to speech (Avivi-Reich et al., 2018; Boncz et al., 2024; Füllgrabe et al., 2015; Rahne et al., 2023; Yang et al., 2022). To better understand the impact of age on material perception in VR, future studies should include a more diverse range of age groups.

4.2.2 Level of musical experience and training

Given that audio was found to be a dominating factor in our study, another point of interest is the affect of musical training on our perception. The dataset collected for our study, did not have an equal distribution of musicians and non-musicians which prevented us from analyzing the effects of musical training. Previous studies (Burns & Houtsma, 1999; Hennessy et al., 2022; Liang et al., 2016; Spiegel & Watson, 1984) have reported that musicians outperformed non musicians in their respective experiments related to frequency discrimination, harmonics and identifying speech in noise. However, there are also studies where there was no difference between the performance of the two groups for their given tasks (Boebinger et al., 2015; Eagleson & Eagleson, 1947; Lutfi et al., 2005).

These mixed results and our limited data prevent us from generalizing the effect of musical training on material perception. Further, little to no studies exist that test the difference in perception of musicians and non musicians in virtual reality or in the field of material perception and is thus an area that that future research can explore.

4.2.3 Quality of the signal used

Prior work has also shown that equipment, environment and recording techniques affect our perception of sound and its quality (Peters et al., 2013). This is important as higher sound quality leads to better immersion and higher reliability - be it in games (Moreira et al., 2019), in virtual simulations (Bonneel et al., 2010) or in assimilating information (Newman & Schwarz, 2018). Given that studies have also shown that sound can both aid and hinder visual perception (Hidaka & Ide, 2015; Teramoto et al., 2012; Williams et al., 2021), it is possible that the sound quality in an audiovisual virtual simulation may affect material perception. Thus, follow-up experiments can be designed to see how differing sound quality affects our material perception in VR.

4.2.4 Physical samples and sound generation

As discussed in Chapter 1.1 we can infer a lot of information about an object's physical properties like material, shape, size, thickness, hollowness etc. from sound (Dimiccoli et al., 2022; Fujisaki et al., 2014; Giordano & McAdams, 2006; Grassi, 2005; Grassi et al., 2013; Klatzky et al., 2000; Kunkler-Peck & Turvey, 2000; Lutfi, 2001). Further, we know that humans are also able to tell apart different materials within the same category based on sound (El-Hadad et al., 2018; Wegst, 2006). And finally, based on the experiments conducted in this study, we can say that mismatch between the auditory properties and visual elements can lead to audiovisual illusions.

This means that if recording of impact sounds are used, then they have to be generated from objects of physical properties similar to what is being simulated in VR. For controlled studies like ours, recording the auditory stimulus under carefully controlled circumstances is necessary but more importantly, it is feasible. It is relatively easy to source material samples of the desired dimensions, creating a tapping mechanism to control and keep constant the force and speed of impact and selecting a suitable recording environment with appropriate recording equipment. However, for large-scale studies requiring a greater variety of sound samples, it would be more effective to use modern sound synthesis methods, which can generate sounds that are nearly indistinguishable from recorded ones (Aramaki & Kronland-Martinet, 2006; Lutfi et al., 2005; Su et al., 2023).

Lastly, stereo sounds were recorded and presented using Unity's sound spatialization technology, which took into account things like source of sound, position, distance and direction with respect to the headset and scaled sounds accordingly. In order to allow some degree of freedom, the force used by the participant at the time of impact was taken into account and the volume of the impact was adjusted with respect to that. This means that the more force used to strike the object, the louder the sound.

For this study we assumed a linear relationship between the impact sound and force used to generate it. When we strike an object harder, the force applied increases, leading to higher vibration amplitudes in the material. In the ideal cases, the sound intensity (power) would be proportional to the force used but perceived loudness is based on a logarithmic scale and is thus, not linear. If both the sound-producing medium (e.g., a speaker or vibrating surface) and the perceptual system (our hearing) were to behave linearly, then doubling the amplitude would indeed double the perceived amplitude. However, human hearing responds approximately logarithmically to amplitude (which is why we use decibels to measure loudness), purely linear changes in signal amplitude won't match our sense of how the loudness of a sound sample should scale. Long samples give more context, making potential mismatches in perceived loudness more noticeable. However, with short, transient sounds (like those used in our study), it may be harder for listeners to identify unnatural loudness scaling because there's less time for the ear/brain to form an expectation or compare.

Follow-up experiments are required to investigate whether taking this into consideration (for impact sounds) and simulating loudness with a higher degree of realism could lead to different results.

4.3 Work in progress

Currently, follow-up experiments are being carried out to see how sensitive human perception is when we change the material of the rod used to generate the sound. This involves carrying out the same experiment but with sound samples generated using a glass rod, metal rod and plastic rod. Further, for each rod, the trials are split into three blocks during which different levels of background noise is being induced (no noise, low levels of noise - 0.05 dB and high levels of noise - 0.1 dB). This is being done to see how reduced sound quality affects our.

4.4 Conclusion

Material perception in VR based on conflicting audiovisual stimuli remains a relatively unexplored field with significant potential for further research. The results of this study suggest that auditory cues are relied upon more frequently than visual cues when classifying materials. Additionally, observing an interaction performed by another person, as opposed to actively participating in the same interaction, did not influence material perception of the object involved. Finally, the findings indicate a potential for audiovisual illusions to occur when conflicting stimuli are presented in a material categorization task. While the specific nature and effects of these illusions may vary depending on several factors, the results of this study suggest that when presented with conflicting audiovisual cues in material perception tasks, some variations of these illusions are likely to emerge.

Appendix A

A.1 Spectral analysis

In order to analyze and compare the sounds, we calculated the spectral centroid, spectral bandwidth and spectral contrast of each of these sounds. These features were then combined into feature vectors and used to calculate the euclidean distance as a way of comparing the similarity between the sound samples. Figure A.1 details these features.

Index	Glass-Metal1	Glass-Metal3	Metal-Metal1	Metal-Metal3	Wood-Metal1	Wood-Metal3	Plastic-Metal1	Plastic-Metal3
Mean Centroid (Hz)	3641.722537	3411.724145	3459.200903	3630.288393	2317.177816	2768.226388	3300.144364	3218.066436
Mean bandwidth (Hz)	2544.865511	2625.233641	2719.246766	2640.055721	2088.820555	2362.016036	2428.371654	2373.747271
Spectral contrast (dB) in band 1	13.955582	11.805072	18.361965	15.651539	17.204523	13.857448	15.093078	16.766470
Spectral contrast (dB) in band 2	11.121187	12.955633	13.230781	8.310192	13.893145	14.124700	11.942895	11.872316
Spectral contrast (dB) in band 3	15.828813	14.826963	15.024644	16.269846	17.714172	16.795357	14.169178	12.265933
Spectral contrast (dB) in band 4	18.352024	18.213338	18.405231	20.549310	20.183605	23.266735	20.876788	21.347961
Spectral contrast (dB) in band 5	22.215478	20.296771	21.198480	18.765622	20.603308	22.335268	21.065314	20.068623
Spectral contrast (dB) in band 6	22.756551	18.891596	19.335469	19.491878	18.284958	18.324629	21.117546	18.742344
Spectral contrast (dB) in band 7	41.143496	40.976382	41.753292	40.650931	42.165736	41.566381	43.078546	42.596227

Figure A.1: The features calculated for checking similarity of sound samples

A.2 Individual participant data

Figures A.1 and A.2 show how individual participant response varied during the matching condition. We can also see that while individual response varies, participant response across agent interaction and self-interaction remains constant.

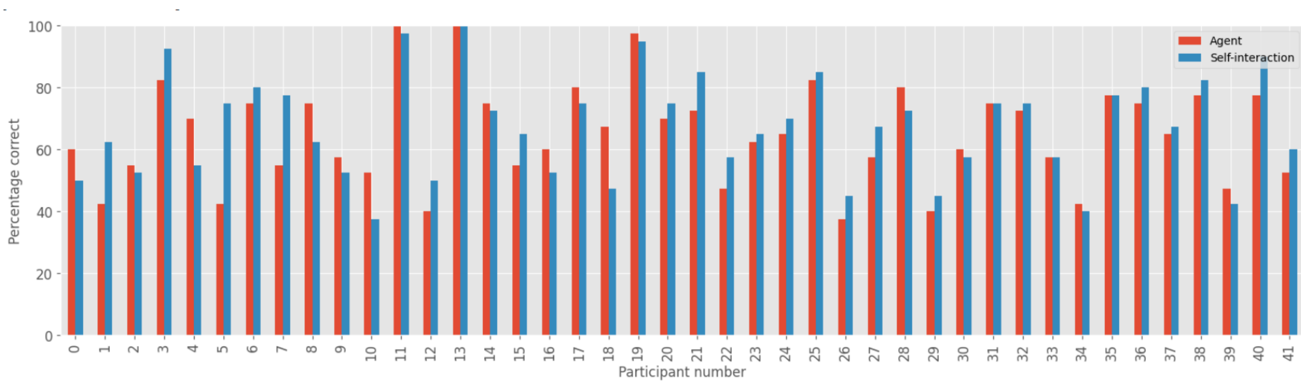

Figure A.2: Percentage of responses where the options selected matched the audio and appearance of the target - after the datasets from the less attentive participants were dropped

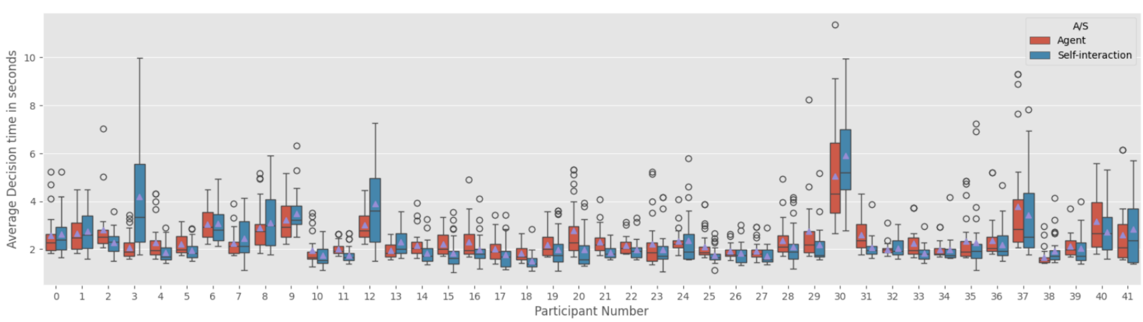

Figure A.3: Average time taken to select an option during the matching condition

We can observe a few things:,

- Average decision times vary greatly between participants (see Figure A.1) but when we observe graphs plotting averages across any of the groups (self interaction vs. agent interaction in Figure 3.2, 3.3. 3.4, 3.5; responses categorized based on sound, sight or neither of the two in Figures 3.6, 3.7, A.10, A.11) we see that there are no noticeable patterns.

- While there are individual differences in the number of responses a participant gets correct during the matching condition, even on an individual level, there is little difference between the number of correct responses correctly identified in the self interaction trials vs. the agent interaction trials (as seen in Figures A.2 and A.3).
- While there are a few participants that have a high tendency to rely on the visual information, most participants tend towards the auditory cues.

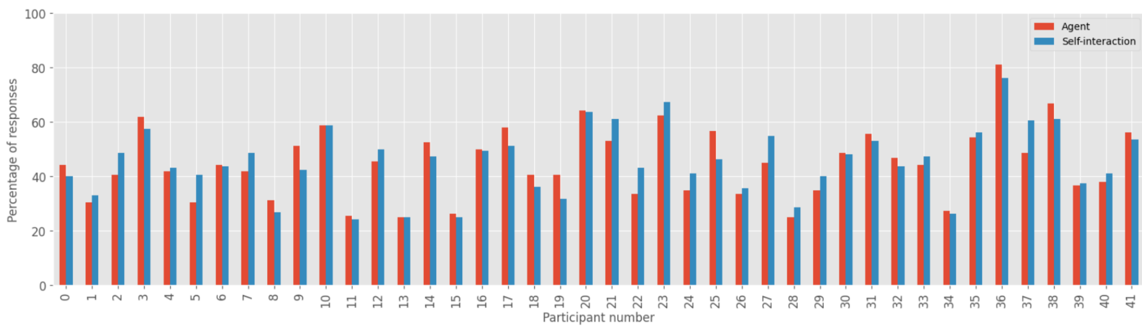

Figure A.4: Percentage of responses where the options selected matches the audio of the target

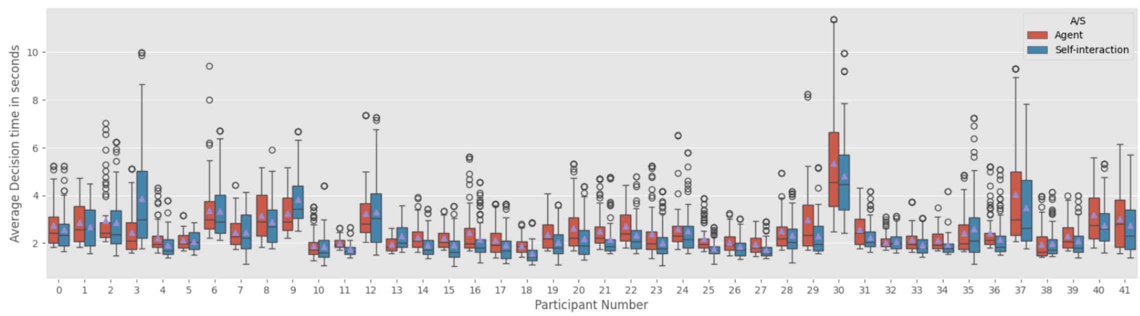

Figure A.5: Average time taken when the option selected matches the audio

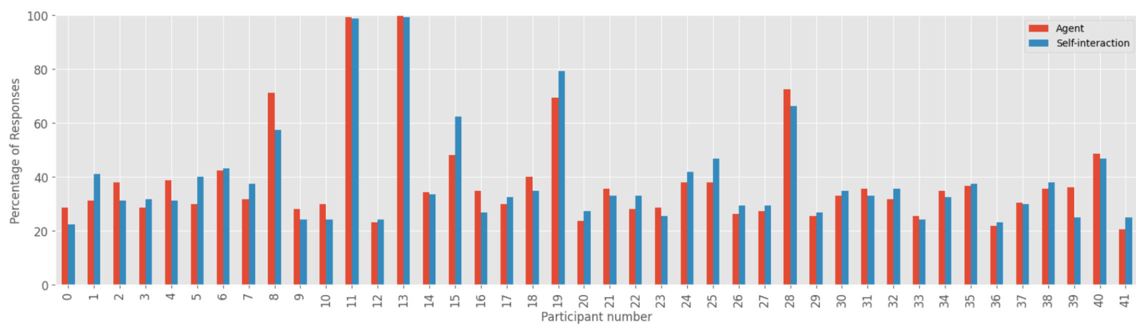

Figure A.6: Percentage of responses where the options selected matches the appearance of the target

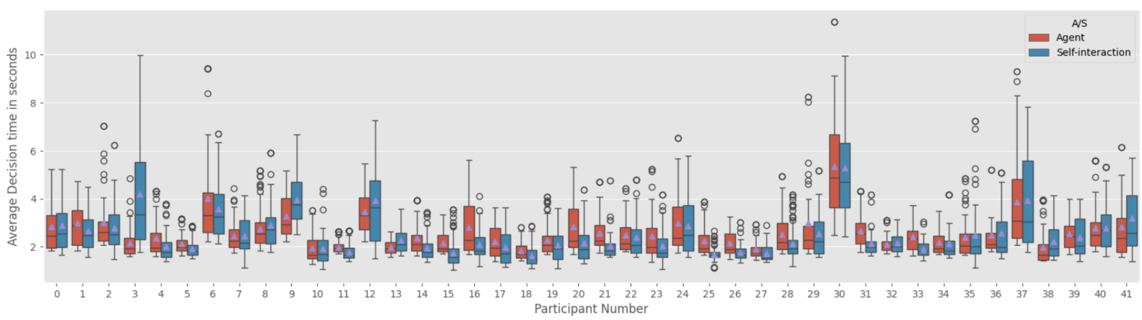

Figure A.7: Average time taken when the option selected matches the appearance

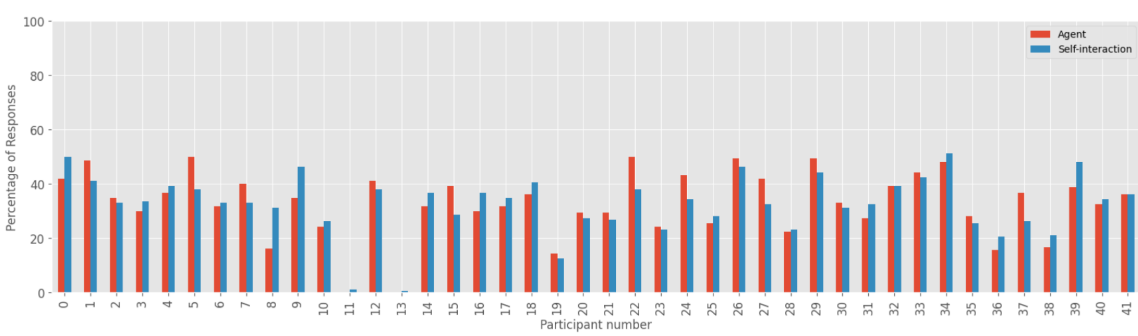

Figure A.8: Percentage of responses where the options selected matches neither the audio nor the appearance of the target

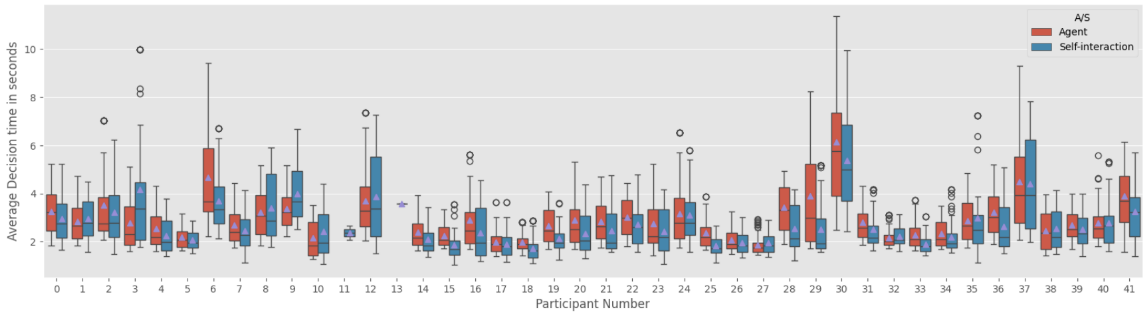

Figure A.9: Average time taken when the option selected matches neither the audio nor the appearance

A.3 Comparing agent-interaction with self-interaction when option selected matches audio/visuals/neither

In section 3.2 we compared how response did not vary between self-interaction blocks and agent-interaction blocks. Some more graphs and comparisons can be seen in this subsection.

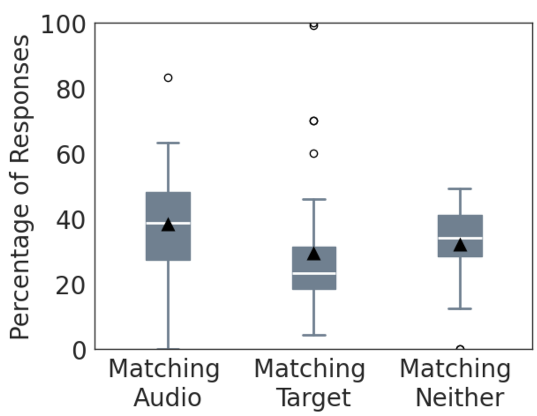

Figure A.10: Percentage of response distribution for agent interaction trials

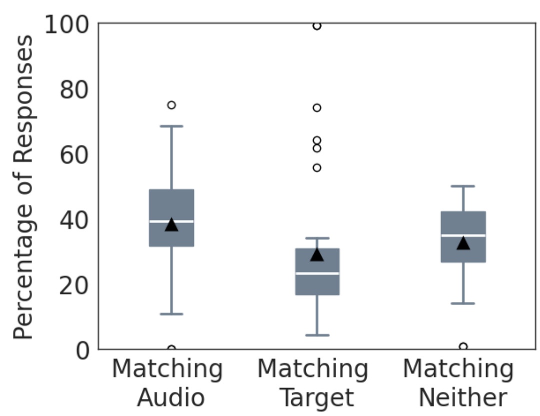

Figure A.11: Percentage of response distribution for self-interaction trials

Looking at Figures A.10-A.13, we see comparable results for Self-interaction and Agent-interaction thus supporting the conclusions from Section 3.2.

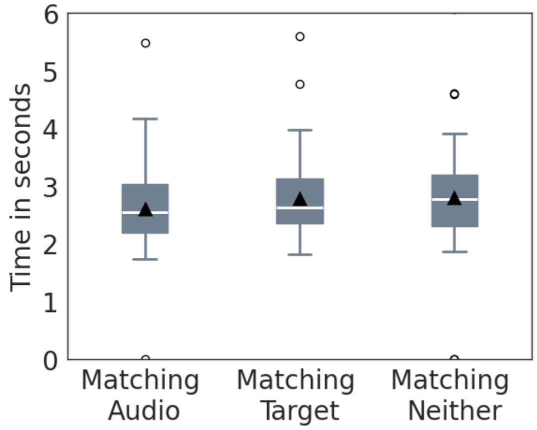

Figure A.12: Average time taken for the three categories during the agent interaction trials

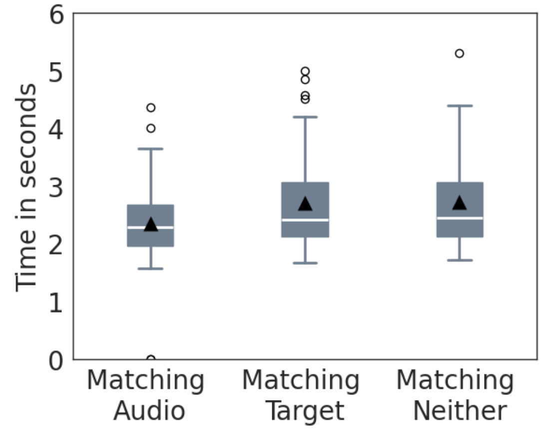

Figure A.13: Average time taken for the three categories during the self-interaction trials

A.4 ANOVA and post-hoc analysis for signal preferences

In section 3.3.2, we performed 2 two-way repeated-measures ANOVAs. The independent variables were the audio played and the visual texture shown. The dependent variables were the percentage of responses participants matched to either the sound or the visual appearance. Since the assumption of Sphericity was not being met, we also performed the Greenhouse-geisser corrections for adjusting the degrees of freedom and p-values. Figures A.14 and A.15 show this data in tabular format.

DV: Percentage of responses matched to the Sound

	F Value	Num DF	Den DF	Pr > F	E_gg	Adj DF_Num	Adj DF_Den	Adj p-value
Audio	23.577573	3.0	123.0	4.007173e-12	0.854074	2.562222	105.051115	1.117542e-10
Video	23.494332	3.0	123.0	4.333325e-12	0.552108	1.656324	67.909289	1.183215e-07
Audio:Video	29.348327	9.0	369.0	2.026335e-38	0.317573	2.858159	117.184500	8.248957e-14

Figure A.14: The results of the two way repeated measures ANOVA where the dependent variable is the percentage of participant responses that match the sound played. E-gg is the epsilon correction to be applied.

We also performed a series of repeated measures one-way ANOVAs to analyze signal preference. The analysis and graphs not included in that section are added here. In the graphs of A.16 -

DV: Percentage of responses matched to the visual texture

	F Value	Num DF	Den DF	Pr > F	E_gg	Adj DF_Num	Adj DF_Den	Adj p-value
Audio	2.279256	3.0	123.0	8.277998e-02	0.397997	1.193991	48.953645	1.333503e-01
Video	26.669666	3.0	123.0	2.341851e-13	0.463857	1.391572	57.054433	2.232664e-07
Audio:Video	63.416045	9.0	369.0	1.687081e-69	0.215513	1.939616	79.524244	1.110223e-16

Figure A.15: The results of the two way repeated measures ANOVA where the dependent variable is the percentage of participant responses that match the texture shown. E-gg is the epsilon correction to be applied.

A.27, on the y-axis, 'nAud' is the number of times a participant has selected the option on the basis of the sound, 'nVid' is the number of times they have selected on the basis of the appearance and 'nNei' is the number of times they matched the option selected to a third different option (unrelated to the sound or the appearance presented).

When checking the responses matched to the Sound

Anova

	F Value	Num DF	Den DF	Pr > F
Audio	39.8627	2.0000	82.0000	0.0000

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Metal	Plastic	0.1667	0.9884	-2.5482	2.8816	False
Metal	Wood	7.0476	0.0	4.3327	9.7625	True
Plastic	Wood	6.881	0.0	4.166	9.5959	True

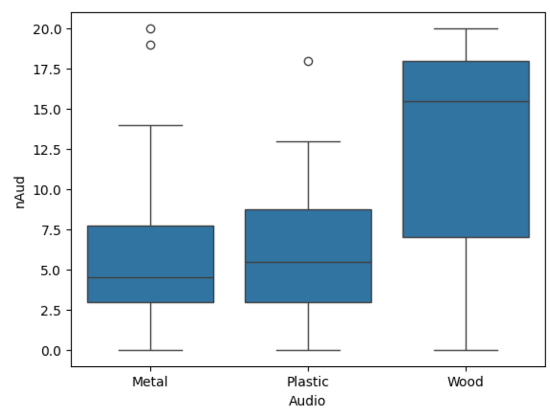

Figure A.16: Repeated measures one-way ANOVA when texture is fixed to glass and sounds are varied. The dependent variable is the frequency of the responses matched to the sound. Followed by the associated post-hoc Tukey HSD test and graphical representation

```
When checking the responses matched to the Appearance
                Anova
=====================================
      F Value Num DF  Den DF Pr > F
-------------------------------------
Audio 51.5654 2.0000 82.0000 0.0000
=====================================
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Metal	Plastic	-3.881	0.0045	-6.7298	-1.0321	True
Metal	Wood	-7.5238	0.0	-10.3727	-4.6749	True
Plastic	Wood	-3.6429	0.0082	-6.4918	-0.794	True

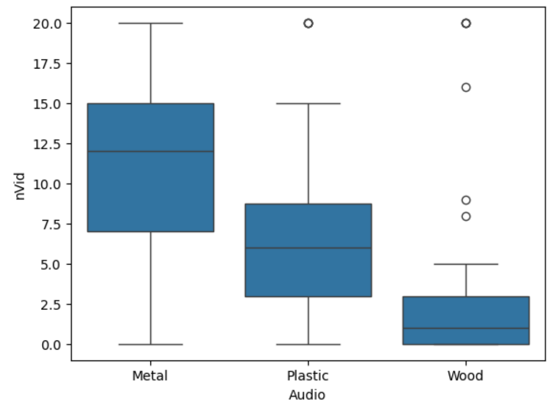

Figure A.17: Repeated measures one-way ANOVA when texture is fixed to glass and sounds are varied. The dependent variable is the frequency of the responses matched to the texture. Followed by the associated post-hoc Tukey HSD test and graphical representation

```
When checking the responses matched to neither
                Anova
=====================================
      F Value Num DF  Den DF Pr > F
-------------------------------------
Audio 10.4342 2.0000 82.0000 0.0001
=====================================
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Metal	Plastic	3.7143	0.0004	1.477	5.9515	True
Metal	Wood	0.4762	0.8691	-1.761	2.7134	False
Plastic	Wood	-3.2381	0.0023	-5.4753	-1.0009	True

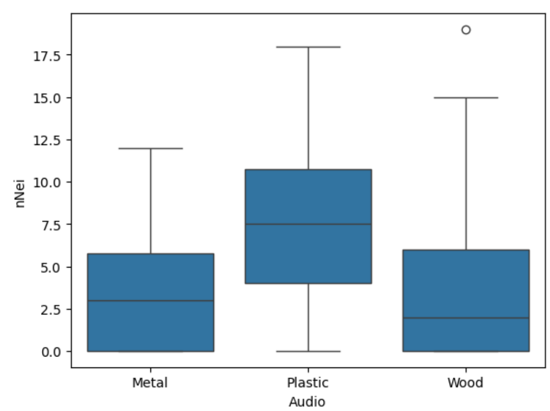

Figure A.18: Repeated measures one-way ANOVA when texture is fixed to glass and sounds are varied. The dependent variable is the frequency of the responses matched to neither the sound nor the texture. Followed by the associated post-hoc Tukey HSD test and graphical representation

```
When checking the responses matched to the Sound
                Anova
===================================
      F Value Num DF  Den DF Pr > F
-----------------------------------
Audio 28.6019 2.0000 82.0000 0.0000
===================================
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Plastic	4.881	0.0002	2.0684	7.6935	True
Glass	Wood	7.3333	0.0	4.5208	10.1459	True
Plastic	Wood	2.4524	0.1007	-0.3602	5.2649	False

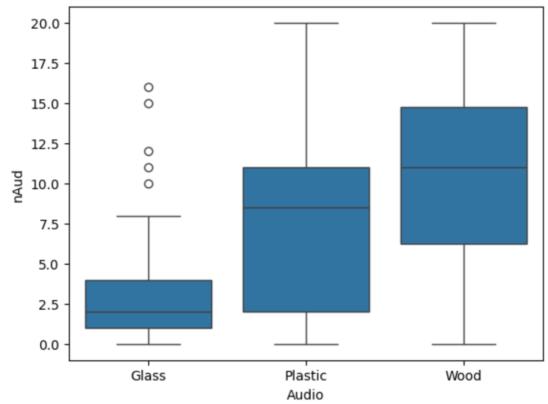

Figure A.19: Repeated measures one-way ANOVA when texture is fixed to metal and sounds are varied. The dependent variable is the frequency of the responses matched to the sound. Followed by the associated post-hoc Tukey HSD test and graphical representation

```
When checking the responses matched to the Appearance
                Anova
===================================
      F Value Num DF  Den DF Pr > F
-----------------------------------
Audio 13.9074 2.0000 82.0000 0.0000
===================================
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Plastic	4.2143	0.0041	1.1482	7.2804	True
Glass	Wood	-0.4048	0.9474	-3.4709	2.6613	False
Plastic	Wood	-4.619	0.0015	-7.6852	-1.5529	True

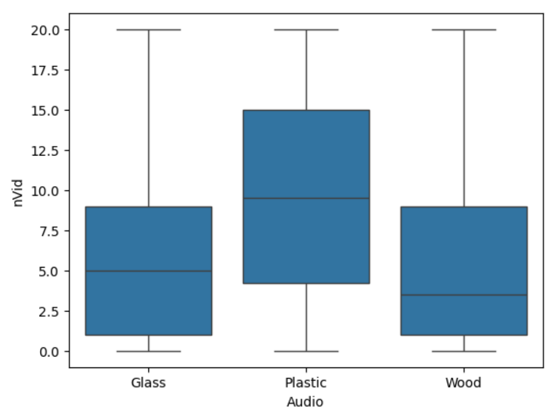

Figure A.20: Repeated measures one-way ANOVA when texture is fixed to metal and sounds are varied. The dependent variable is the frequency of the responses matched to the texture. Followed by the associated post-hoc Tukey HSD test and graphical representation

When checking the responses matched to neither

```
                Anova
===================================
      F Value Num DF  Den DF Pr > F
-----------------------------------
Audio 74.9776 2.0000 82.0000 0.0000
===================================
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Plastic	-9.0952	0.0	-11.1429	-7.0476	True
Glass	Wood	-6.9286	0.0	-8.9762	-4.8809	True
Plastic	Wood	2.1667	0.0354	0.119	4.2143	True

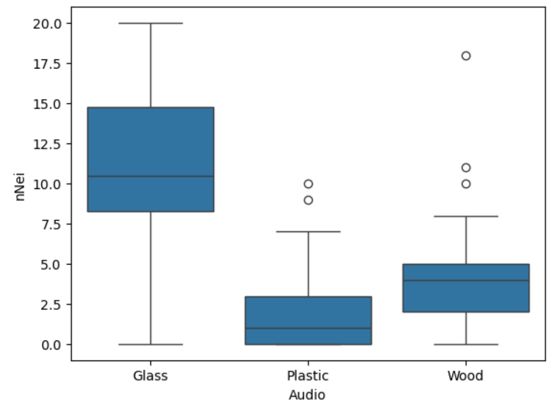

Figure A.21: Repeated measures one-way ANOVA when texture is fixed to metal and sounds are varied. The dependent variable is the frequency of the responses matched to neither the sound nor the texture. Followed by the associated post-hoc Tukey HSD test and graphical representation

When checking the responses matched to the Sound

```
               Anova
===================================
      F Value Num DF  Den DF Pr > F
-----------------------------------
Audio 18.5542 2.0000 82.0000 0.0000
===================================
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	2.4762	0.0608	-0.0877	5.04	False
Glass	Wood	5.9524	0.0	3.3885	8.5162	True
Metal	Wood	3.4762	0.0047	0.9123	6.04	True

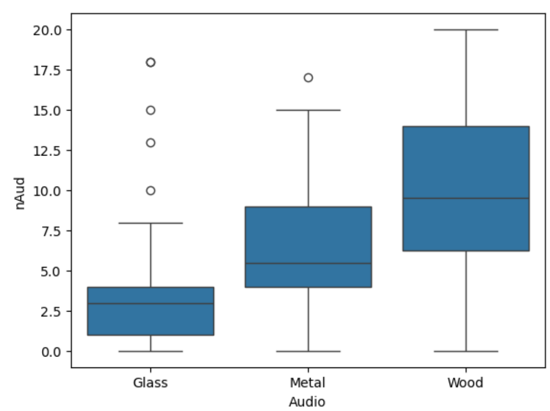

Figure A.22: Repeated measures one-way ANOVA when texture is fixed to plastic and sounds are varied. The dependent variable is the frequency of the responses matched to the sound. Followed by the associated post-hoc Tukey HSD test and graphical representation

```
When checking the responses matched to the Appearance
                    Anova
===================================
       F Value Num DF  Den DF Pr > F
-----------------------------------
Audio  2.6696 2.0000 82.0000 0.0753
===================================
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	0.3571	0.9468	-2.3322	3.0465	False
Glass	Wood	-1.4762	0.3967	-4.1655	1.2132	False
Metal	Wood	-1.8333	0.2422	-4.5227	0.856	False

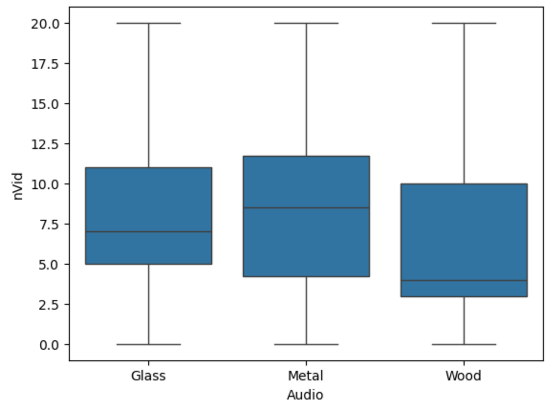

Figure A.23: Repeated measures one-way ANOVA when texture is fixed to plastic and sounds are varied. The dependent variable is the frequency of the responses matched to the texture. Followed by the associated post-hoc Tukey HSD test and graphical representation

```
When checking the responses matched to neither
                    Anova
===================================
       F Value Num DF  Den DF Pr > F
-----------------------------------
Audio 19.2356 2.0000 82.0000 0.0000
===================================
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	-2.8333	0.0027	-4.8167	-0.85	True
Glass	Wood	-4.4762	0.0	-6.4596	-2.4928	True
Metal	Wood	-1.6429	0.1253	-3.6262	0.3405	False

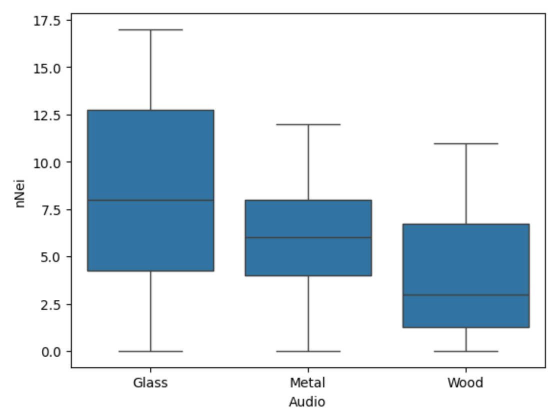

Figure A.24: Repeated measures one-way ANOVA when texture is fixed to plastic and sounds are varied. The dependent variable is the frequency of the responses matched to neither the sound nor the texture. Followed by the associated post-hoc Tukey HSD test and graphical representation

```
When checking the responses matched to the Sound
                 Anova
===================================
      F Value Num DF  Den DF Pr > F
-----------------------------------
Audio  9.3290 2.0000 82.0000 0.0002
===================================
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	-1.8571	0.2351	-4.5529	0.8387	False
Glass	Plastic	-4.1905	0.001	-6.8863	-1.4947	True
Metal	Plastic	-2.3333	0.104	-5.0291	0.3625	False

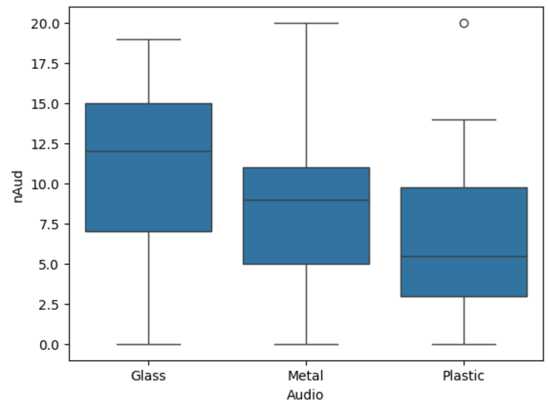

Figure A.25: Repeated measures one-way ANOVA when texture is fixed to wood and sounds are varied. The dependent variable is the frequency of the responses matched to the sound. Followed by the associated post-hoc Tukey HSD test and graphical representation

```
When checking the responses matched to the Appearance
                 Anova
===================================
      F Value Num DF  Den DF Pr > F
-----------------------------------
Audio  4.6260 2.0000 82.0000 0.0125
===================================
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	-0.2619	0.9659	-2.7376	2.2138	False
Glass	Plastic	-0.619	0.824	-3.0947	1.8566	False
Metal	Plastic	-0.3571	0.9375	-2.8328	2.1185	False

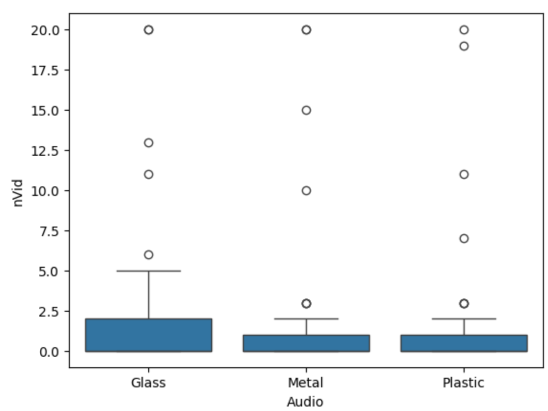

Figure A.26: Repeated measures one-way ANOVA when texture is fixed to wood and sounds are varied. The dependent variable is the frequency of the responses matched to the texture. Followed by the associated post-hoc Tukey HSD test and graphical representation

```
When checking the responses matched to neither
               Anova
===================================
      F Value Num DF  Den DF Pr > F
-----------------------------------
Audio 12.9379 2.0000 82.0000 0.0000
===================================
```

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	2.119	0.1313	-0.4695	4.7076	False
Glass	Plastic	4.8095	0.0001	2.2209	7.3981	True
Metal	Plastic	2.6905	0.0397	0.1019	5.2791	True

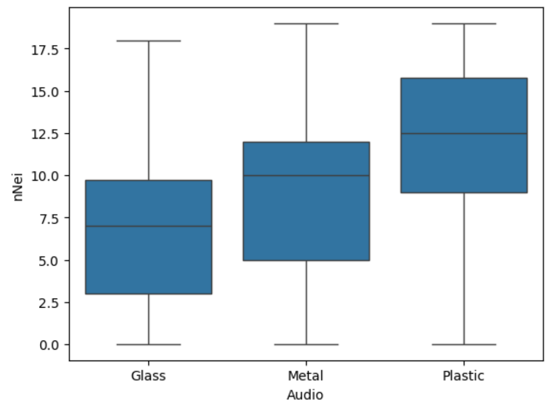

Figure A.27: Repeated measures one-way ANOVA when texture is fixed to wood and sounds are varied. The dependent variable is the frequency of the responses matched to neither the sound nor the texture. Followed by the associated post-hoc Tukey HSD test and graphical representation

A.5 Pairwise comparison to statistically check the significance of the different conditions

In section 3.4 we analyzed the response distribution of each of the 16 testing conditions. Figure 3.11 represented this information in a grid of charts. Statistical pairwise Tukey HSD tests were carried out to check the significace of this grid of graphs and ascertain what options were selected the most for each condition. Familywise error rate corrections (with $\alpha = 0.05$ were carried out during the analysis using Tukey's procedure. All conditions' pairwise analysis is added in this section and the summary of these results is included in Table 3.1 in the section 3.4.

We can see that:

- In Figures A.28, A.33, A.38 and A.43, for the matching conditions, the material being displayed is always the predominant option selected.
- In Figure A.29, for the case of Glass texture coupled with Metal audio, responses to all options were statistically different from each other with Glass being the most selected and metal being the second most selected options, making it a visual dominant case.

For Glass Texture and Glass Audio:
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	-12.2857	0.0	-14.084	-10.4874	True
Glass	Plastic	-11.9524	0.0	-13.7507	-10.1541	True
Glass	Wood	-14.1429	0.0	-15.9411	-12.3446	True
Metal	Plastic	0.3333	0.9632	-1.4649	2.1316	False
Metal	Wood	-1.8571	0.04	-3.6554	-0.0589	True
Plastic	Wood	-2.1905	0.01	-3.9888	-0.3922	True

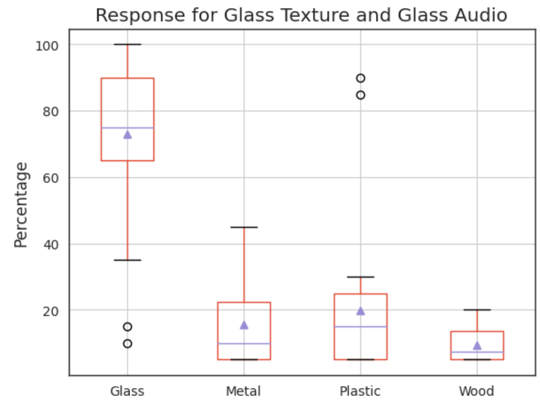

Figure A.28: Box plot of the glass texture and glass audio condition with the associated Tukey HSD pairwise test

For Glass Texture and Metal Audio:
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	-5.119	0.0	-7.3894	-2.8487	True
Glass	Plastic	-7.5952	0.0	-9.8656	-5.3249	True
Glass	Wood	-10.619	0.0	-12.8894	-8.3487	True
Metal	Plastic	-2.4762	0.0266	-4.7466	-0.2058	True
Metal	Wood	-5.5	0.0	-7.7704	-3.2296	True
Plastic	Wood	-3.0238	0.0038	-5.2942	-0.7534	True

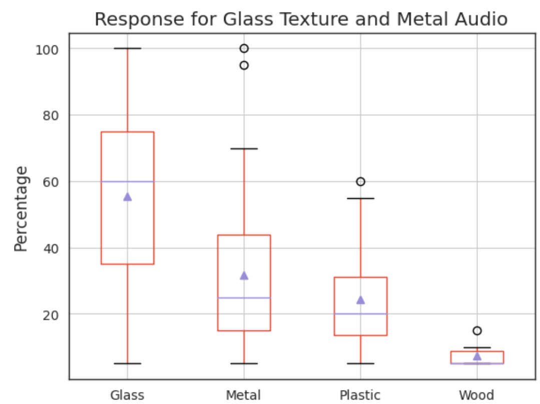

Figure A.29: Box plot of the glass texture and metal audio condition with the associated Tukey HSD pairwise test

For Glass Texture and Plastic Audio:
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	0.0952	0.9996	-2.2303	2.4207	False
Glass	Plastic	-1.0714	0.6303	-3.3969	1.2541	False
Glass	Wood	-6.8333	0.0	-9.1588	-4.5078	True
Metal	Plastic	-1.1667	0.5629	-3.4922	1.1588	False
Metal	Wood	-6.9286	0.0	-9.2541	-4.6031	True
Plastic	Wood	-5.7619	0.0	-8.0874	-3.4364	True

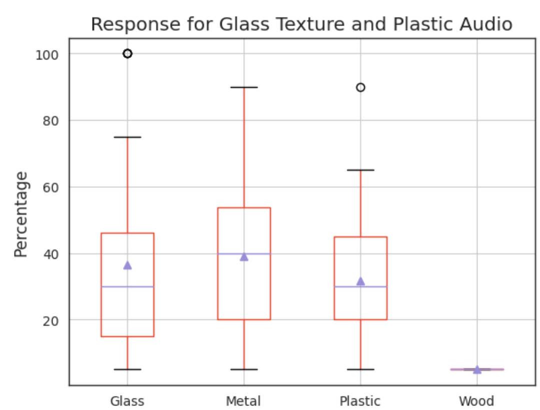

Figure A.30: Box plot of the glass texture and plastic audio condition with the associated Tukey HSD pairwise test

For Glass Texture and Wood Audio:
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	-1.9048	0.282	-4.6672	0.8577	False
Glass	Plastic	-0.7857	0.8815	-3.5482	1.9768	False
Glass	Wood	9.4524	0.0	6.6899	12.2149	True
Metal	Plastic	1.119	0.7194	-1.6434	3.8815	False
Metal	Wood	11.3571	0.0	8.5947	14.1196	True
Plastic	Wood	10.2381	0.0	7.4756	13.0006	True

Figure A.31: Box plot of the glass texture and wood audio condition with the associated Tukey HSD pairwise test

For Metal Texture and Glass Audio:
Multiple Comparison of Means - Tukey HSD,
FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	2.4762	0.0767	-0.1759	5.1283	False
Glass	Plastic	1.8571	0.2688	-0.795	4.5093	False
Glass	Wood	2.7143	0.0427	0.0622	5.3664	True
Metal	Plastic	-0.619	0.9301	-3.2712	2.0331	False
Metal	Wood	0.2381	0.9955	-2.414	2.8902	False
Plastic	Wood	0.8571	0.8359	-1.795	3.5093	False

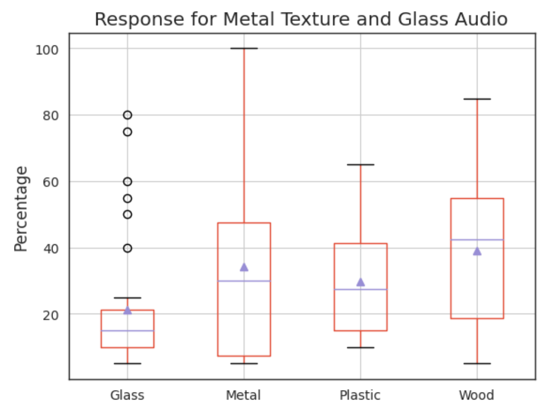

Figure A.32: Box plot of the metal texture and glass audio condition with the associated Tukey HSD pairwise test

For Metal Texture and Metal Audio:
Multiple Comparison of Means - Tukey HSD,
FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	7.5952	0.0	5.7607	9.4298	True
Glass	Plastic	0.9762	0.513	-0.8584	2.8107	False
Glass	Wood	-3.3333	0.0	-5.1679	-1.4988	True
Metal	Plastic	-6.619	0.0	-8.4536	-4.7845	True
Metal	Wood	-10.9286	0.0	-12.7631	-9.094	True
Plastic	Wood	-4.3095	0.0	-6.1441	-2.475	True

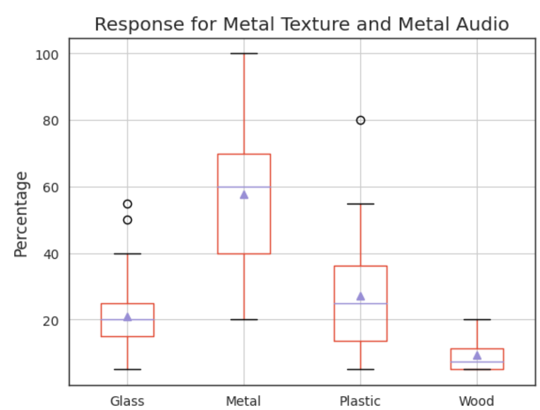

Figure A.33: Box plot of the metal texture and metal audio condition with the associated Tukey HSD pairwise test

For Metal Texture and Plastic Audio:
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	8.5	0.0	5.8626	11.1374	True
Glass	Plastic	6.6905	0.0	4.0531	9.3279	True
Glass	Wood	-0.9048	0.8098	-3.5422	1.7327	False
Metal	Plastic	-1.8095	0.2863	-4.4469	0.8279	False
Metal	Wood	-9.4048	0.0	-12.0422	-6.7673	True
Plastic	Wood	-7.5952	0.0	-10.2327	-4.9578	True

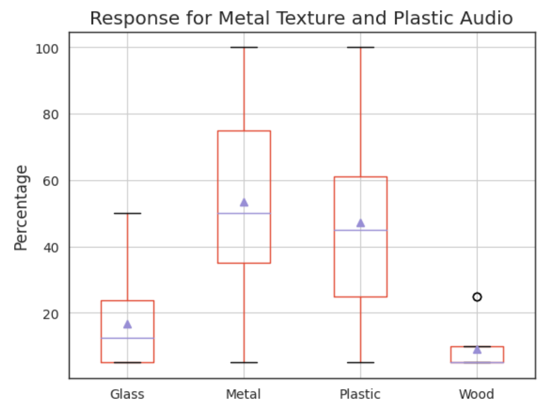

Figure A.34: Box plot of the metal texture and plastic audio condition with the associated Tukey HSD pairwise test

For Metal Texture and Wood Audio:
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	4.5238	0.0	1.9817	7.0659	True
Glass	Plastic	2.5476	0.0493	0.0055	5.0898	True
Glass	Wood	9.7857	0.0	7.2436	12.3279	True
Metal	Plastic	-1.9762	0.1857	-4.5183	0.5659	False
Metal	Wood	5.2619	0.0	2.7198	7.804	True
Plastic	Wood	7.2381	0.0	4.696	9.7802	True

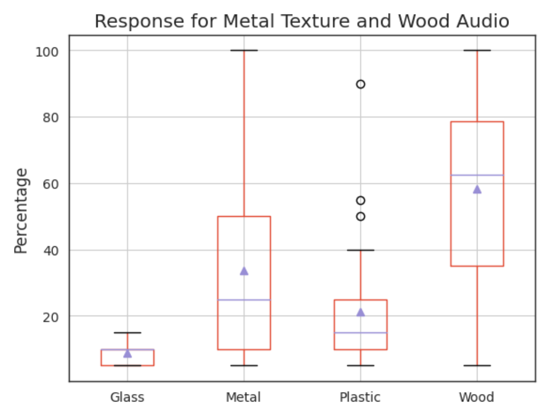

Figure A.35: Box plot of the metal texture and wood audio condition with the associated Tukey HSD pairwise test

For Plastic Texture and Glass Audio:
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	-0.8571	0.8207	-3.4166	1.7023	False
Glass	Plastic	3.881	0.0007	1.3215	6.4404	True
Glass	Wood	1.3571	0.516	-1.2023	3.9166	False
Metal	Plastic	4.7381	0.0	2.1787	7.2975	True
Metal	Wood	2.2143	0.1155	-0.3451	4.7737	False
Plastic	Wood	-2.5238	0.0548	-5.0832	0.0356	False

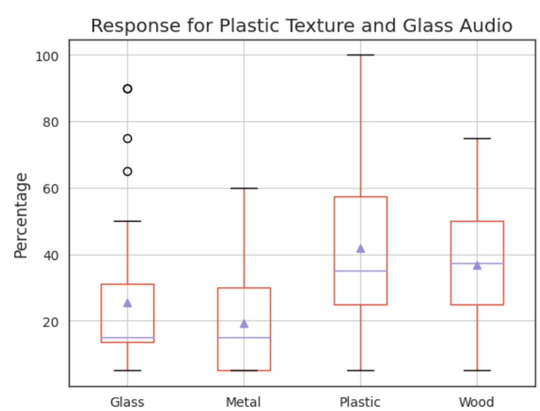

Figure A.36: Box plot of the plastic texture and glass audio condition with the associated Tukey HSD pairwise test

For Plastic Texture and Metal Audio:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	1.0952	0.5106	-0.9575	3.148	False
Glass	Plastic	2.8571	0.0023	0.8044	4.9099	True
Glass	Wood	-5.0952	0.0	-7.148	-3.0425	True
Metal	Plastic	1.7619	0.12	-0.2908	3.8146	False
Metal	Wood	-6.1905	0.0	-8.2432	-4.1377	True
Plastic	Wood	-7.9524	0.0	-10.0051	-5.8996	True

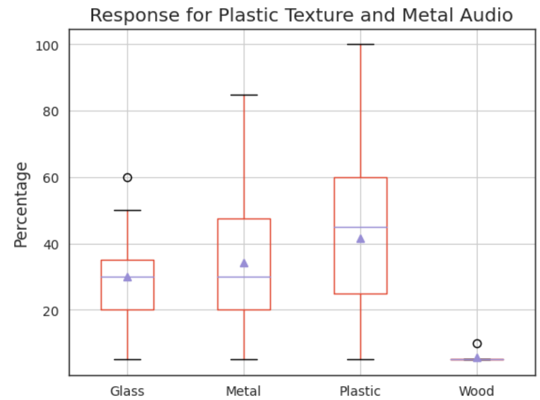

Figure A.37: Box plot of the plastic texture and metal audio condition with the associated Tukey HSD pairwise test

For Plastic Texture and Plastic Audio:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	5.0238	0.0	2.5349	7.5127	True
Glass	Plastic	9.1429	0.0	6.654	11.6317	True
Glass	Wood	-1.2143	0.5856	-3.7032	1.2746	False
Metal	Plastic	4.119	0.0002	1.6302	6.6079	True
Metal	Wood	-6.2381	0.0	-8.727	-3.7492	True
Plastic	Wood	-10.3571	0.0	-12.846	-7.8683	True

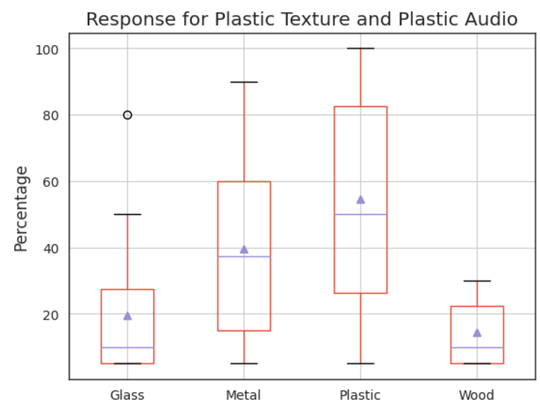

Figure A.38: Box plot of the plastic texture and plastic audio condition with the associated Tukey HSD pairwise test

For Plastic Texture and Wood Audio:

Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	1.3571	0.4614	-1.0483	3.7626	False
Glass	Plastic	5.0714	0.0	2.666	7.4768	True
Glass	Wood	8.619	0.0	6.2136	11.0245	True
Metal	Plastic	3.7143	0.0005	1.3089	6.1197	True
Metal	Wood	7.2619	0.0	4.8565	9.6673	True
Plastic	Wood	3.5476	0.001	1.1422	5.953	True

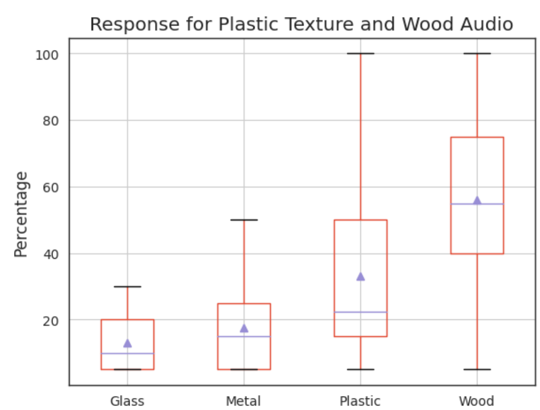

Figure A.39: Box plot of the plastic texture and wood audio condition with the associated Tukey HSD pairwise test

For Wood Texture and Glass Audio:
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	-6.9286	0.0	-9.5956	-4.2615	True
Glass	Plastic	-6.9762	0.0	-9.6432	-4.3091	True
Glass	Wood	-8.1905	0.0	-10.8575	-5.5234	True
Metal	Plastic	-0.0476	1.0	-2.7147	2.6194	False
Metal	Wood	-1.2619	0.6099	-3.929	1.4052	False
Plastic	Wood	-1.2143	0.6392	-3.8813	1.4528	False

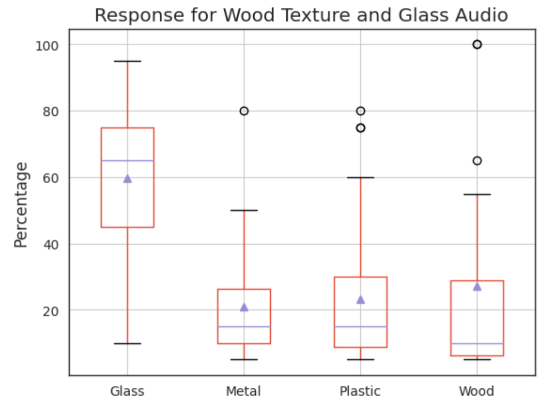

Figure A.40: Box plot of the wood texture and glass audio condition with the associated Tukey HSD pairwise test

For Wood Texture and Metal Audio:
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	3.619	0.0016	1.091	6.1471	True
Glass	Plastic	-0.8333	0.8276	-3.3614	1.6947	False
Glass	Wood	-2.9762	0.0138	-5.5042	-0.4482	True
Metal	Plastic	-4.4524	0.0001	-6.9804	-1.9244	True
Metal	Wood	-6.5952	0.0	-9.1233	-4.0672	True
Plastic	Wood	-2.1429	0.1275	-4.6709	0.3852	False

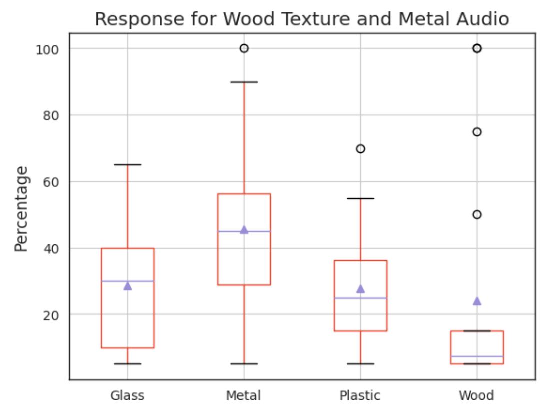

Figure A.41: Box plot of the wood texture and metal audio condition with the associated Tukey HSD pairwise test

For Wood Texture and Plastic Audio:
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	5.0952	0.0	2.642	7.5485	True
Glass	Plastic	2.9048	0.0131	0.4515	5.358	True
Glass	Wood	-1.7143	0.2706	-4.1675	0.739	False
Metal	Plastic	-2.1905	0.0982	-4.6437	0.2628	False
Metal	Wood	-6.8095	0.0	-9.2628	-4.3563	True
Plastic	Wood	-4.619	0.0	-7.0723	-2.1658	True

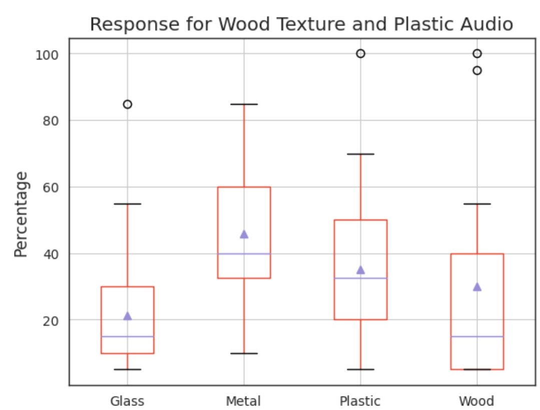

Figure A.42: Box plot of the wood texture and plastic audio condition with the associated Tukey HSD pairwise test

For Wood Texture and Wood Audio:
Multiple Comparison of Means - Tukey HSD, FWER=0.05

group1	group2	meandiff	p-adj	lower	upper	reject
Glass	Metal	0.2857	0.96	-1.211	1.7824	False
Glass	Plastic	0.8095	0.4987	-0.6872	2.3062	False
Glass	Wood	15.2857	0.0	13.789	16.7824	True
Metal	Plastic	0.5238	0.8004	-0.9729	2.0205	False
Metal	Wood	15.0	0.0	13.5033	16.4967	True
Plastic	Wood	14.4762	0.0	12.9795	15.9729	True

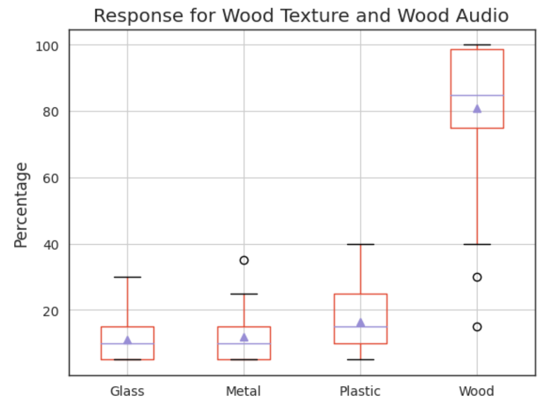

Figure A.43: Box plot of the wood texture and wood audio condition with the associated Tukey HSD pairwise test

- In Figure A.30, for the case of Glass texture coupled with Plastic audio, we have a tie across 3 categories: glass, metal and plastic. Thus, being labeled an anomaly condition. Similarly, in A.32 and A.36 we see ties between multiple responses and no clear predominant option selected, thus we mark them as anomaly cases as well.
- In Figures A.31, A.35, A.39, for the cases of Glass texture coupled with Wood audio, Metal texture coupled with wood and Plastic texture and Wood audio, wood is the predominant option selected, making them audio dominant cases.
- In Figure A.34, for the case of Metal texture coupled with Plastic audio, there is no statical difference between the responses for metal and plastic options, and is thus, an Audio-visual tie case. This is also seen in Figures A.37 (Plastic texture and Metal Audio) where the top-most selected options matched the sound and appearance and were not statistically different from each other.
- In Figures A.40 and A.41, for the cases of Wood texture coupled with Glass or Metal audio, the predominant option selected was glass and metal respectively, making them audio dominant cases.
- In Figure A.42, we see another anomaly case where Wood texture is coupled with plastic audio. The predominant response is a tie between Metal and Plastic. While plastic is a

given, metal is an unrelated response selected here.

References

Adams, W. J. (2016). The development of audio-visual integration for temporal judgements. *PLoS Computational Biology*, *12*(4), e1004865.

Adelson, E. H. (2001). On seeing stuff: the perception of materials by humans and machines. In *Human vision and electronic imaging vi* (Vol. 4299, pp. 1–12).

Andersen, T. S., Tiippana, K., & Sams, M. (2004). Factors influencing audiovisual fission and fusion illusions. *Cognitive Brain Research*, *21*(3), 301–308.

Anderson, R., Arro, J., Hansen, C. S., & Serafin, S. (2016). Audio-visual perception-the perception of object material in a virtual environment. In *Augmented reality, virtual reality, and computer graphics: Third international conference, avr 2016, lecce, italy, june 15-18, 2016. proceedings, part i 3* (pp. 162–171).

Aramaki, M., & Kronland-Martinet, R. (2006). Analysis-synthesis of impact sounds by real-time dynamic filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, *14*(2), 695–705.

Avivi-Reich, M., Puka, K., & Schneider, B. A. (2018). Do age and linguistic background alter the audiovisual advantage when listening to speech in the presence of energetic and informational masking? *Attention, Perception, & Psychophysics*, *80*, 242–261.

Bi, W., & Xiao, B. (2016). Perceptual constancy of mechanical properties of cloth under variation of external forces. In *Proceedings of the acm symposium on applied perception* (pp. 19–23).

Boebinger, D., Evans, S., Rosen, S., Lima, C. F., Manly, T., & Scott, S. K. (2015). Musicians and non-musicians are equally adept at perceiving masked speech. *The Journal of the Acoustical Society of America*, *137*(1), 378–387.

Boncz, Á., Szalárdy, O., Velősy, P. K., Béres, L., Baumgartner, R., Winkler, I., & Tóth, B. (2024). The effects of aging and hearing impairment on listening in noise. *Iscience*, *27*(4).

Bonneel, N., Suied, C., Viaud-Delmon, I., & Drettakis, G. (2010). Bimodal perception of audio-visual material properties for virtual environments. *ACM Transactions on Applied Perception (TAP)*, *7*(1), 1–16.

Bouman, K. L., Xiao, B., Battaglia, P., & Freeman, W. T. (2013). Estimating the material properties of fabric from video. In *Proceedings of the ieee international conference on computer vision* (pp. 1984–1991).

Brimijoin, W. O., & Akeroyd, M. A. (2014). The moving minimum audible angle is smaller during self motion than during source motion. *Frontiers in neuroscience*, *8*, 273.

Buckingham, G., Cant, J. S., & Goodale, M. A. (2009). Living in a material world: how visual cues to material properties affect the way that we lift objects and perceive their weight. *Journal of neurophysiology*, *102*(6), 3111–3118.

Burns, E. M., & Houtsma, A. J. (1999). The influence of musical training on the perception of sequentially presented mistuned harmonics. *The Journal of the Acoustical Society of America*, *106*(6), 3564–3570.

Cao, Z., Magalhães, E., & Bernardes, G. (2023). Sound design impacts user experience and attention in serious game. In *Joint international conference on serious games* (pp. 95–110).

Carello, C., Anderson, K. L., & Kunkler-Peck, A. J. (1998). Perception of object length by sound. *Psychological science*, *9*(3), 211–214.

Chan, J. S., Wibral, M., Stawowsky, C., Brandl, M., Helbling, S., Naumer, M. J., ... Wollstadt, P. (2021). Predictive coding over the lifespan: increased reliance on perceptual priors in older adults—a magnetoencephalography and dynamic causal modeling study. *Frontiers in aging neuroscience*, *13*, 631599.

DeLoss, D. J., Pierce, R. S., & Andersen, G. J. (2013). Multisensory integration, aging, and the sound-induced flash illusion. *Psychology and aging*, *28*(3), 802.

Dimiccoli, M., Patni, S., Hoffmann, M., & Moreno-Noguer, F. (2022). Recognizing object surface material from impact sounds for robot manipulation. In *2022 ieee/rsj international conference on intelligent robots and systems (iros)* (pp. 9280–9287).

Durst, R. S., & Krotkov, E. P. (1995). Object classification from analysis of impact acoustics. In *Proceedings 1995 ieee/rsj international conference on intelligent robots and systems. human robot interaction and cooperative robots* (Vol. 1, pp. 90–95).

Eagleson, H. V., & Eagleson, O. W. (1947). Identification of musical instruments when heard directly and over a public address system. *The Journal of the Acoustical Society of America*, *19*(1_Supplement), 291–291.

El-Hadad, A., Brodie, G. I., & Ahmed, B. S. (2018). The effect of wood condition on sound wave propagation. *Open Journal of Acoustics*, *8*(03), 37.

Faul, F. (2021). Perceived roughness of glossy objects: The influence of fresnel effects and correlated image statistics. *Journal of Vision*, *21*(8), 1–1.

Fleming, R. W. (2017). Material perception. *Annual review of vision science*, *3*(1), 365–388.

Fujisaki, W., Goda, N., Motoyoshi, I., Komatsu, H., & Nishida, S. (2014). Audiovisual integration in the human perception of materials. *Journal of Vision*, *14*(4), 12–12.

Füllgrabe, C., Moore, B. C., & Stone, M. A. (2015). Age-group differences in speech identification despite matched audiometrically normal hearing: contributions from auditory temporal processing and cognition. *Frontiers in aging neuroscience*, *6*, 347.

Garner, T., Grimshaw, M., & Nabi, D. A. (2010). A preliminary experiment to assess the fear value of preselected sound parameters in a survival horror game. In *Proceedings of the 5th audio mostly conference: A conference on interaction with sound.* New York, NY, USA: Association for Computing Machinery. Retrieved from `https://doi.org/10.1145/1859799.1859809` doi: doi: 10.1145/1859799.1859809

Gau, R., & Noppeney, U. (2016). How prior expectations shape multisensory perception. *Neuroimage*, *124*, 876–886.

Giordano, B. L., & McAdams, S. (2006). Material identification of real impact sounds: Effects of size variation in steel, glass, wood, and plexiglass plates. *The Journal of the Acoustical Society of America*, *119*(2), 1171–1181.

Grassi, M. (2005). Do we hear size or sound? balls dropped on plates. *Perception & psychophysics*, *67*, 274–284.

Grassi, M., Pastore, M., & Lemaitre, G. (2013). Looking at the world with your ears: How do we get the size of an object from its sound? *Acta psychologica*, *143*(1), 96–104.

Guillen, G., Jylhä, H., & Hassan, L. (2021). The role sound plays in games: a thematic literature study on immersion, inclusivity and accessibility in game sound research. In *Proceedings of the 24th international academic mindtrek conference* (pp. 12–20).

Hennessy, S., Mack, W. J., & Habibi, A. (2022). Speech-in-noise perception in musicians and non-musicians: A multi-level meta-analysis. *Hearing research*, *416*, 108442.

Hidaka, S., & Ide, M. (2015). Sound can suppress visual perception. *Scientific reports*, *5*(1), 10483.

Hirst, R. J., McGovern, D. P., Setti, A., Shams, L., & Newell, F. N. (2020). What you see is what you hear: Twenty years of research using the sound-induced flash illusion. *Neuroscience & Biobehavioral Reviews*, *118*, 759–774.

Hjortkjær, J., & McAdams, S. (2016). Spectral and temporal cues for perception of material and action categories in impacted sound sources. *The Journal of the Acoustical Society of America*, *140*(1), 409–420.

Ide, M., & Hidaka, S. (2013). Tactile stimulation can suppress visual perception. *Scientific reports*, *3*(1), 3453.

Jaekl, P. M., & Harris, L. R. (2009). Sounds can affect visual perception mediated primarily by the parvocellular pathway. *Visual neuroscience*, *26*(5-6), 477–486.

Klatzky, R. L., Pai, D. K., & Krotkov, E. P. (2000). Perception of material from contact sounds. *Presence*, *9*(4), 399–410.

Kondo, H. M., Pressnitzer, D., Toshima, I., & Kashino, M. (2012). Effects of self-motion on auditory scene analysis. *Proceedings of the National Academy of Sciences*, *109*(17), 6775–6780.

Koumura, T., & Furukawa, S. (2017). Context-dependent effect of reverberation on material perception from impact sound. *Scientific reports*, *7*(1), 16455.

Kunkler-Peck, A. J., & Turvey, M. T. (2000). Hearing shape. *Journal of Experimental psychology: human perception and performance*, *26*(1), 279.

Lakatos, S., McAdams, S., & Caussé, R. (1997). The representation of auditory source characteristics: Simple geometric form. *Perception & psychophysics*, *59*(8), 1180–1190.

Lemaitre, G., & Heller, L. M. (2012). Auditory perception of material is fragile while action is strikingly robust. *The Journal of the Acoustical Society of America*, *131*(2), 1337–1348.

Lerch, A. (2022). *An introduction to audio content analysis: Music information retrieval tasks and applications*. John Wiley & Sons.

Liang, C., Earl, B., Thompson, I., Whitaker, K., Cahn, S., Xiang, J., … Zhang, F. (2016). Musicians are better than non-musicians in frequency change detection: behavioral and electrophysiological evidence. *Frontiers in Neuroscience*, *10*, 464.

Lutfi, R. A. (2001). Auditory detection of hollowness. *The Journal of the Acoustical Society of America*, *110*(2), 1010–1019.

Lutfi, R. A., Oh, E., Storm, E., & Alexander, J. M. (2005). Classification and identification of recorded and synthesized impact sounds by practiced listeners, musicians, and nonmusicians. *The Journal of the Acoustical Society of America*, *118*(1), 393–404.

Malpica, S., Serrano, A., Allue, M., Bedia, M. G., & Masia, B. (2020). Crossmodal perception in virtual reality. *Multimedia Tools and Applications*, *79*, 3311–3331.

Mazuz, Y., Kessler, Y., & Ganel, T. (2024). Age-related changes in size perception: A tale of three illusions.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Moran, R. J., Symmonds, M., Dolan, R. J., & Friston, K. J. (2014). The brain ages optimally to model its environment: evidence from sensory learning over the adult lifespan. *PLoS computational biology*, *10*(1), e1003422.

Moreira, J., Gros, L., Nicol, R., Viaud-Delmon, I., Le Prado, C., & Natkin, S. (2019). Binaural sound rendering improves immersion in a daily usage of a smartphone video game. In *Eaa spatial audio signal processing symposium* (pp. 79–84).

Moris Fernandez, L., Macaluso, E., & Soto-Faraco, S. (2017). Audiovisual integration as conflict resolution: The conflict of the mcgurk illusion. *Human brain mapping*, *38*(11), 5691–5705.

Newman, E. J., & Schwarz, N. (2018). Good sound, good research: How audio quality influences perceptions of the research and researcher. *Science Communication*, *40*(2), 246–257.

Peters, N., Braasch, J., & McAdams, S. (2013). Recording techniques and their effect on sound quality at off-center listening positions in 5.0 surround environments. *Canadian Acoustics*, *41*(3), 37–49.

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A bayesian perspective on magnitude estimation. *Trends in cognitive sciences*, *19*(5), 285–293.

Photinos, P. (2021). *The physics of sound waves (second edition)*. IOP Publishing. Retrieved from `https://dx.doi.org/10.1088/978-0-7503-3539-3` doi: doi: 10.1088/978-0-7503-3539-3

Proakis, J. G., & Manolakis, D. G. (2013). *Digital signal processing: Pearson new international edition*. Pearson Higher Ed.

Rahne, T., Wagner, T. M., Kopsch, A. C., Plontke, S. K., & Wagner, L. (2023). Influence of age on speech recognition in noise and hearing effort in listeners with age-related hearing loss. *Journal of Clinical Medicine*, *12*(19), 6133.

Seriès, P., & Seitz, A. R. (2013). Learning what to expect (in visual perception). *Frontiers in human neuroscience*, *7*, 668.

Shams, L., Iwaki, S., Chawla, A., & Bhattacharya, J. (2005). Early modulation of visual cortex by sound: an meg study. *Neuroscience letters*, *378*(2), 76–81.

Shams, L., Kamitani, Y., & Shimojo, S. (2000). What you see is what you hear. *Nature*, *408*(6814), 788–788.

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive brain research*, *14*(1), 147–152.

Shams, L., Kamitani, Y., Thompson, S., & Shimojo, S. (2001). Sound alters visual evoked potentials in humans. *Neuroreport*, *12*(17), 3849–3852.

Sharan, L., Rosenholtz, R., & Adelson, E. H. (2014). Accuracy and speed of material categorization in real-world images. *Journal of vision*, *14*(9), 12–12.

Skarbez, R., Brooks, F. P., & Whitton, M. C. (2021). Immersion and coherence: Research agenda and early results. *IEEE Transactions on Visualization and Computer Graphics*, *27*(10), 3839-3850. doi: doi: 10.1109/TVCG.2020.2983701

Slater, M. (2009). Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1535), 3549–3557.

Spiegel, M. F., & Watson, C. S. (1984). Performance on frequency-discrimination tasks by musicians and nonmusicians. *The Journal of the Acoustical Society of America*, *76*(6), 1690–1695.

Su, K., Qian, K., Shlizerman, E., Torralba, A., & Gan, C. (2023). Physics-driven diffusion models for impact sound synthesis from videos. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition* (pp. 9749–9759).

Teramoto, W., Hidaka, S., Sugita, Y., Sakamoto, S., Gyoba, J., Iwaya, Y., & Suzuki, Y. (2012). Sounds can alter the perceived direction of a moving visual object. *Journal of Vision*, *12*(3), 11–11.

Traer, J., & McDermott, J. H. (2016). Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, *113*(48), E7856–E7865.

Tucker, S., & Brown, G. J. (2003). Modelling the auditory perception of size, shape and material: Applications to the classification of transient sonar sounds. In *Audio engineering society convention 114.*

Van Assen, J. J. R., Barla, P., & Fleming, R. W. (2018). Visual features in the perception of liquids. *Current biology*, *28*(3), 452–458.

van Assen, J. J. R., & Fleming, R. W. (2016). Influence of optical material properties on the perception of liquids. *Journal of vision*, *16*(15), 12–12.

Wegst, U. G. K. (2006). Wood for sound. *American journal of botany*, *93*(10), 1439–1448.

Williams, J. R., Markov, Y., Tiurina, N., & Störmer, V. (2021). What you hear is what you see: Sounds alter the contents of visual perception.

Yang, W., Guo, A., Yao, H., Yang, X., Li, Z., Li, S., … Wu, J. (2022). Effect of aging on audiovisual integration: Comparison of high-and low-intensity conditions in a speech discrimination task. *Frontiers in Aging Neuroscience*, *14*, 1010060.

Zhang, F., de Ridder, H., Barla, P., & Pont, S. (2020). Effects of light map orientation and shape on the visual perception of canonical materials. *Journal of vision*, *20*(4), 13–13.

Zlobin, D. (2024). Intrinsically unpleasant sounds and player experience.