Investigating Textual Visual Sound Effects in a Virtual Environment and their impacts on Object Perception and Sound Perception

Thibault Fabre* ESIEA, France
Adrien Verhulst† The University of Tokyo, Japan
Alfonso Balandra‡ The University of Tokyo, Japan
Maki Sugimoto§ Keio University, Japan
Masahiko Inami¶ The University of Tokyo, Japan

ABSTRACT

In comics, Textual Sound Effects (TE) can describe sounds, but also actions, events, etc. TE could be used in Virtual Environment to efficiently create an easily recognizable scene and add more information to objects at a relatively low design cost. We investigate the impact of TE in a Virtual Environment on objects' material perception (on category and properties) and on sound perception (on volume [dB] and spatial position). Participants (N=13, repeated measures) categorized metallic and wooden spheres and significantly changed their reaction time depending on the TE congruence with the spheres' material/sound. They then rated a sphere's properties (i.e., wetness, warmness, softness, smoothness, and dullness) and significantly changed their rating depending on the TE. When comparing 2 sound volumes, they perceived a sound associated with a shrinking TE as less loud and a sound associated with a growing TE as louder. When locating an audio source location, they located it significantly closer to a TE.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality

1 INTRODUCTION

If you ever saw the "Vapor cone" of a plane doing a Sonic Boom, you likely felt able to "Visualize the sound" of the Boom. Those cones help us better understand the "Pressure waves" pushed by a moving object. If we were able to see those cones from everyday sounds, we'd essentially be able to locate sounds more easily. Similarly, if we were able to see a descriptor (e.g., a subtitle) from a sound, we'd be able to better "describe" it (e.g., the object made a high-pitched "TAP" sound), and augment our environment's perception.

Because Virtual Reality (VR) allows us to add those descriptors (and developers are making use of it), we want to explore how a subset of those descriptors, Textual sound Effects (abbr. TEs) can enhance or modify the perception of a Virtual Environment (VE).

TEs are already used in numerous VR applications. Not necessarily because they are part of the graphics design (like in XIII [29]), but also because they have VR specifics use-cases, such as: Describing the VE (or even describing the interaction with the VE); increasing the immersion [4]; drawing attention; highlighting an interaction [23]; etc.

Outside of VR, TEs usually take the form of onomatopoeia. They are mostly used in comics (e.g., "BOOM" when an object explodes in American comics) and others visual arts (e.g., the movie "Scott Pilgrim vs. the World"). Even if TEs typically describe sounds, in the context of onomatopoeia, it is better to think of them as an *Environmental descriptors*, since they can describe things beyond sounds, such as actions (e.g., "ZZZ" for sleeping), events (e.g., "TA-DA" for a surprise), emotions (e.g., reddish texturing for anger), etc. [16]. Moreover, because TEs are part of the language, we can easily understand their meaning and what they describe [8].

TEs sit at the crossroad between Visual Effect and Sound Effect, and use the former to describe the latter. TEs are written symbols and characters illustrating a sound. Let's also note that others possible words for TEs are: Textualized sounds [1], Comic book soundtrack [28], or Visual Sound Effects.

Before presenting our hypothesis, we present topics related to TEs: Audiovisual cross-modality and Sound symbolism.

1.1 A Bias of Vision on Sound and Perception

Audiovisual cross-modality typically provides a better perception of the environment than the sole auditory or visual channel (e.g., a better material perception [5, 12]). But by introducing bias (a.k.a incongruities) in the audiovisual cross-modality perception, it is possible to change the auditory and visual perception. For example, it is possible to bias auditory localization with the *Ventriloquism effect* [5] (when there is a spatial bias between an auditory and visual stimulus, the auditory localization leans toward the location of the visual stimulus) or with the *Ventriloquism aftereffect* [18] (a re-calibration of the auditory localization to correct the ventriloquism effect). There are numerous other examples, such as: The object size can be influenced by loudness or frequency [7]; the object length by loudness [10]; the object weight by pitch [27]; etc. Therefore, adding a visual stimulus to an auditory stimulus can affect the auditory/visual perception.

A TE might also be able to affect a cross-modal visual/audio interaction. For example, a TE congruent with the visual/audio might help to perceive more precisely the VE (e.g., is this object made of wood?). Such a result would be useful for VR designers looking to make their objects easily identifiable.

1.2 A Mapping between Sound and Perception

Sound Symbolism (or phonosemantics) refers to the non-arbitrary mappings between the phonetic properties of sounds and their perceptual and/or semantic elements[1] [13,26]. For example, the *mil/mal effect* [22, 26], where the /mil/ and /mal/ non-words are respectively associated with small and big shapes; and the *bouba/kiki effect* [17, 26], where the /bouba/ and /kiki/ non-words are respectively associated with round and sharp shapes. Therefore, a sound can carry meaning by itself, and a sound-symbolic word (like an onomatopoeia) might also likely carry meaning by itself.

A TE might also be able to carry properties to affect an object's property perception. For example, a TE with "water-like properties" (blue colored onomatopoeia with a water-sound like text, water

*Join co-first author; e-mail: tfabre@et.esiea.fr
†Join co-first author; e-mail: adrienverhulst@star.rcast.u-tokyo.ac.jp
‡e-mail: alfonso@star.rcast.u-tokyo.ac.jp
§e-mail: sugimoto@ics.keio.ac.jp
¶e-mail: inami@star.rcast.u-tokyo.ac.jp

[1] Sound Symbolism is a broad field in linguistic. It also classifies several types of sound-symbolic words, among which: *onomatopoeia*, *phenomines* (words that depict states, conditions, or manners of the external world [24]) and *psychomimes* (words that depict psychological states [24]).

drops, etc.) might help perceive an object's wet/dry property. Such a result would be useful for VR designers looking to increase the range of perceived objects' properties with minimal design costs.

1.3 Hypothesis

From the previous subsections above, we formulate the following hypothesis:

- H1 A TE congruent with a visual/audio interaction will lead to an object classification (in category) faster than a TE non-congruent with a visual/audio interaction;
- H2 A TE congruent with a visual/audio interaction will lead to a difference in object classification (in category) compared to a TE non-congruent with a visual/audio interaction;
- H3 A TE associated with a material property affects the perception of an object material.
- H4 A TE has an impact on the perception of the sound volume;
- H5 A TE has an impact on the perception of the sound location.

In summary, H1 and H2 focus of the object category (e.g., wood, metal), and H3 on the object properties (e.g., wet, dry, soft, hard, etc.). H4 and H5 both focus on sound perception. While H1 and H2 are somewhat similar, H1 allows us to know if an identification can be made faster with congruent/non-congruent TE, and H2 if there is a difference in the classification with congruent/non-congruent TE.

1.4 Novelties

Our work focuses on exploring if and how TEs can be used to enhance or modify the perception of a VE in VR. It has the following novelties:

- We show that TEs can change the categorization speed of an object and are therefore taken into account when categorizing an object (although they do not impact the categorization result);
- We show that TEs can change the perception of an object's properties and that its impact is linked to the object/interaction that it describes. We also show that a TE's impact can be furthered by adding a visual descriptor;
- We show that TEs have an effect on sound volume perception and sound localization (albeit this is a small novelty).

2 Related Works

We present here recent works on text presentation in VR, on TE in VR, and on audiovisual cross-modality of material perception.

2.1 Subtitles in Virtual Reality

Texts in VR, and more generally in Mixed Reality, are often used to display subtitles. But because in VR the user can move their head, and therefore look at something else than a "point of interest", the subtitles' positioning has to be carefully considered [19, 20]; all while keeping text readability, text-speaker association, immersion, etc. in mind.

Rothe et al. [20] compared 2 types of subtitles' positioning in VR: static – the subtitles stay at the same screen position even if the user moves their head; and dynamic – the subtitles stay near the speaker. Their results show that dynamic subtitles have: Higher presence; lower VR sickness; and lower workload (on the NASA-TLX items [9]). However, static subtitles cannot be missed when looking around in the VE [3].

Rzayev et al. [21] compared 4 types of notification positioning in VR: P_a = Attached to the display (i.e., static); P_b = Attached to the VR controller; P_c = Floating in the VE, and instantiated at about 1.5m in front of the user; and P_d = Attached to the VE (e.g., a wall) and instantiated on the closest object in front of the user. Their results show that: P_a should be used for important notifications; P_d should be used for unimportant notifications; and P_b and P_c are preferred by the participants and should generally be used.

Texts in VR can also have depth issues (mainly occlusion with the VE and fatigue). To avoid those issues, Sidenmark et al. [25] suggested 3 techniques using eye-tracking in VR: Blurring the area behind the gazed-on subtitles; matching the subtitles depth with the gazed-on object; combining both. There was no evaluation of those techniques.

What we take away from those works is that dynamic text floating in the VE and in the direction of a sound are overall better for our use case. We will therefore display our TEs this way.

2.2 Onomatopoeia in Virtual Reality

Research on onomatopoeia in VR is rather scarce. Oh et Kim. [15] conducted a pilot study in VR to compare: C_a = onomatopoeia + sounds; and C_b = no onomatopoeia + sounds. The comparison was done in 2 different VEs: VE_a = A farm with noisy animals; and VE_b = A kitchen with noisy objects. Their results show that there are no significant differences in *Realism* for both VEs between C_a and C_b, and that the effect on presence/immersion might depend on the crowdedness of the scene. The authors also reported that onomatopoeia motivates the users to interact with the VE, although they did not report the results to support it. The same year, Choi et al. [4] (from the same laboratory than [15]) compared in VR the contact sound of 3 materials (metal, stone, wood) under the 2 following conditions: S_a = The sounds are specific for each material; S_b = The sounds are the same for each material. They compared S_a and S_a with and without onomatopoeia. Their results show that S_b with onomatopoeia and S_a without onomatopoeia were perceived as similar in realism/naturalness, meaning that adding an onomatopoeia to a non-realist/non-natural sound (here non-congruent sound) can match the realism/naturalness of a realist/natural sound (here congruent sound).

2.3 Audiovisual Material Perception

As noted in the introduction, audiovisual cross-modality is particularly useful in material perception.

Fujisaki et al. [5] compared the audiovisual interaction of different visual appearances associated with different impact sounds (+/- congruent). Their results show that: material-category ratings (e.g., "is the object wood?") follow a multiplicative integration rule, while material-property ratings (e.g., "what is the object's roughness?") follow a weighted average rule. Klatzky et al. [12] investigated the relation between material perception and the parameters of contact sounds synthesis. Giordano at al. [6] did a similar investigation, with also a variety of sizes for each material. Their results show that it is possible to discriminate objects along their material categories (like [12]), but not to discriminate objects along their sizes (a property here) when the objects are of the same material category.

Specifically in 3D. Bonneel et al. [2] compared different auditory Level-Of-Detail (LOD) and visual LOD for falling/bouncing 3D models. Their results show that with a high LOD sound, an object with a low LOD visual can be perceived at a higher visual quality (it is somewhat similar to the results of [4]). Malpica et al. [14] showed that visual motion perception is affected by auditory stimulus in VR, but at a lower degree than reported for 2D screen. They also showed that audiovisual cross-modality (seeing a sphere and hearing a congruent or non-congruent impact sound) can also affect material property perception.

3 Our Textual Sound Effects

We use a total of 6 TEs (c.f., Fig. 1). A TE is a combination of a text (e.g., "Knock" for wood), a text color and an animation. Almost all TEs stay visible for 0.3s (only the TE-CarHorn-Long

stays visible longer, for at most 5s). All the TEs have a short ease-in-out animation (on scale and opacity for [0s,0.1s]) to make them appear seamlessly in the VE. Some TEs also have an additional animation: TE-Wood has a dangling animation on each letter; TE-Metal has a jiggling animation on each letter (and the animations are faster than those on TE-Wood's letters); and TE-SoundSymbol expands with a bit of bounciness.

We present in table 1 the congruent TE - sound associations. When the TE - sound association is non-congruent, then the TE is associated to one of the other sound.

Figure 1: Our TEs, here in Japanese. From left to right: TE-Wood; TE-Metal; TE-Water; TE-CarHorn-Short; TE-CarHorn-Long; and TE-SoundSymbol (the only TE without text).

4 System Configuration

The system consists of a Head Mounted Display (HMD), a stereo sound compatible headset and is implemented in an Unity 2019[2] Steam VR[3] application.

Hardware When the experiment was done in our laboratory, then the participants used a *Valve Index*[4] and the application ran on a VR-able laptop (there was no noticeable latency). When the experiment was done remotely, the participants used the HMD and the application ran on the VR-able PC they had available. The remote participants used the following headsets: *Valve Index*, *HTC Vive*[5], *HTC Vive Pro*[6] and *Oculus Quest*[7] with the Oculus Link. The participants had to wear an over-hear headset with stereo sound capability (in our laboratory we used the Audio-technica ATH-M20x).

Virtual Environment The VE consists of a plane and a grey skybox to not distract the participant and make the colored TEs more apparent. The participants answer the tasks' questions directly in the VE. In step A1 they do it with the controllers trigger, and in steps A2, B3, and B4 by pointing at a UI / 3D object. The number of remaining trials is always displayed. Also, the participants can do the step A1, A2 and B3 while seated.

Materials and sounds The wood and metal materials use realistic high-quality textures from *Poliigon*[8]. For steps A1 and A2, we used the sound of real wood and metal recorded while being knocked on. For steps B3 and B4, we used a short or long car horn sound taken from a license-free sound effects library.

5 Experiment Design

We present here the design, the protocol and the different steps of the experiment.

5.1 Participants

There were 13 participants: 11 males (age: M = 25.4 years, SD = 2.35 years) and 2 females (age: M = 32.0 years, SD = 8.0 years). The experiment took place in HIDDEN UNIVERSITY in Japan (June-July 2020), but 6 participants (≈ 46%) did the experiment remotely. . The participants were all students. Among them, 1 (≈ 8%) had "some experience" with VR, and 12 had "extensive experience" with VR. 11 participants were from Japan (≈ 85%) and 2 from France. All of them frequently read Japanese comics; 3 (≈ 23%) did not read Western comics, and 9 had at least some experience with Western comics. All of them were fluent in Japanese and were able to speak at least conversational English. No participants had an auditory disorder.

[2]https://unity.com/
[3]https://www.steamvr.com/
[4]https://www.valvesoftware.com/en/index
[5]https://www.vive.com/eu
[6]https://www.vive.com/eu/product/vive-pro/
[7]https://www.oculus.com/quest/
[8]https://www.poliigon.com/

5.2 Protocol

We divided the experiment into 2 parts of 2 steps: parts A and B, with their respective steps A1, A2 and B3, B4 (c.f., Fig. 2). The order of the part, their steps, and their conditions were randomized. Part A focuses on the perception of objects, that is, can we alter the visual perception of an object with TEs. Part B focuses on the perception of sounds, that is, can we alter the auditory perception with TEs.

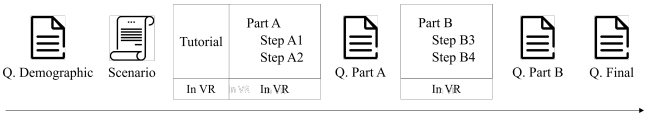

Figure 2: The workflow protocol. Q stands for questionnaire.

Participants are invited one by one to join the experiment. Only the participant and the assistant are present in the experiment room[9]. The participant answers a demographic questionnaire and signs a consent form. Then, they read the experimental instruction (c.f., Tab. 2) and watches videos examples of the tasks. The participant then puts on the HMD and follows a tutorial (c.f., Sec. 5.3) to get used to the controls. They then do the steps in order (e.g., A2, A1, then B4, B3). Each step has a training session. After doing part A or B of the experiment, the participant puts down the HMD and answers the part's questionnaire on a computer (c.f., Tabs. 3) then puts on the HMD and continues to the next part (if any). After doing part A and B, the participant leaves the room[10].

About remote participants. Remote participants were supervised through video conference software. They were asked to share their view and audio for the entirety of the experiment. Participants were also required to wear over-the-hear headphones to take part in steps B3 and B4. They were also required to keep their camera and microphone activated. Any internet problem that would hinder the supervision would have resulted in the early termination of the current step (but we did not encounter such an issue).

5.3 Tutorial

Here the participant gets used to the controls. The participant is in an empty VE in front of a set of balls to interact with (i.e., hitting them, grabbing them, launching them). There is no time limit. When the participant feels used to the controls, they stops the tutorial (tutorial time: M ≈ 59s , SD ≈ 16s).

5.4 The Steps

We present here the 4 steps of the experiment. For each step, we present their concept and conditions, then we present the implementation details. Please note that the steps are independents of each other.

5.4.1 Step A1

Here we try to answer H1 and H2. We combine a ball's material (either a "Wooden" or a "Metallic" material) with an impact sound (either a "Wood" or a "Metal" sound) and an impact TE (either the TE-Wood or the TE-Metal) and ask the participant to categorize the

[9]For remote participants, only the participant is present in their room.
[10]Remote participants have to transmit the recorded data to the assistant.

TE name	Sound	Japanese text	English text	French text	Used in step	Animation
TE-Wood	Wood	コン/kɑn/	Knock	Toc	A1,A2	Ease in/out + dangling
TE-Metal	Metal	ティン/tɯ/	Ching	Ting	A1,A2	Ease in/out + fast jiggling
TE-Water	Water	ピチャ/ˈpɪtʃaɔ/	Splash	Sploch	A2	Ease in/out
TE-CarHorn-Short	Car horn (short)	プゥー/puu/	Doot	Pouet	B3	Ease in/out
TE-CarHorn-Long	Car horn (long)	ファーン/faan/	Honk	Tuut	B3	Ease in/out
TE-SoundSymbol	Car horn (long)	-	-	-	B4	Ease in/out + bounciness

Table 1: The congruent TE - sound association in the participants' language, as well as the step the TE is used in and the TE animation(s).

Step	Instruction
A1	You will see an object fall on a platform. You will then be asked to choose if the object's material is of type A or B. You will have to choose an answer FAST (we will measure your reaction time). There will be an answer on the LEFT and an answer on the RIGHT. To choose the answer on the LEFT press a button on the LEFT controller, to choose the answer on the RIGHT press a button on the RIGHT controller. The answer only depends of your perception. There is not a correct or an incorrect answer. *[Video example]*
A2	You will hit an object with your controller(s) for a few seconds. You will then be asked several questions about the object's properties. You can take your time to answer. *To select the answer X, point your controller toward X, press your controller trigger (left or right) to select X, then press NEXT to validate X and go to the next question. [Video example]*
B3	You will hear 2 sounds. You will then decide if the last sound was QUIETER or LOUDER than the first sound. You can take your time to answer. *To select the answer X, point your controller toward X, press your controller trigger (left or right) to select X, then press NEXT to validate X and go to the next question. [Video example]*
B4	There will be 18 speakers behind you (stay in position until you hear a sound). You will hear a sound coming from one of the speakers behind you, and you will turn to try to locate the speaker from which the sound is coming from. Please remember, THE SOUND IS ONLY COMING FROM 1 SPEAKER. *To select the speaker X, point your controller toward X, press your controller trigger (left or right) to select X, then press NEXT to validate X and go to the next question. [Video example]*

Table 2: The instructions that the participant read before the experiment

ID	Question
QA-1	What do you think the experiment purpose is?
QA-2	In the experiment we use text effects. Did you like them or did you dislike them? And why?
QA-3	In the experiment we use text effects. Did you feel they are useful? And how?
QA-4	Did you feel the chosen text were appropriate (コン; ティン ; ピチャ)? Would you have chosen different texts?
QA-5	Did you focus more on the text effects, the objects' material, the sound or something else? (Order them from the most important to the least important).
QA-6,7,8,9	*nda. Here the participants did the NASA TLX on a web page and reported, about the text effects, the: Mental Demand, Temporal Demand, Effort, Frustration*.
QA-10	Do you have any remarks regarding the experiment?
QB-1	What do you think the experiment purpose is?
QB-2	In the experiment we use text effects. Did you like them or did you dislike them? And why?
QB-3	In the experiment we use text effects. Did you feel they are useful? And how?
QB-4	Did you feel the chosen text were appropriate (honk, doot)? Would you have chosen different texts?
QB-5	Do you have any remarks regarding the experiment?

Table 3: QA-X: Questionnaire part A, regarding steps A1 and A2; and QB-X: Questionnaire part B, regarding steps B3 and B4. "Text effects" refers to the TEs.

ball (wood or metal?). Please refer to Tab. 4 to see the different combinations. The main idea is to know, depending on the congruence of the material, sound, and TE, if the TE can shift the perception of the object category toward the material or sound or TE. Because the combinations are mostly non-congruent, no correct answers are expected from the participants.

Implementation detail A transparent ball falls on a platform in front of the participant. When the ball hits the platform, the ball's material becomes visible, and the sound and TE start/are instantiated at the contact point for 0.3s before stopping/disappearing. The ball's physics impact (e.g., bouncing, rolling) is not implemented to constrain the participant's attention on the material, sound, and TE. The participant quickly answers whether the ball is a wood/metal ball with the left/right controller's trigger. We record the mean Reaction Time (RT; RT = time of the answer from the moment the ball hits the platform) and the answer.

5.4.2 Step A2

Here we try to answer H3. We compare several ball's impact's TE: the TE-Wood or the TE-Water, with or without water droplets[11] and ask the participant to rate the ball properties on a 7 points Likert scale (c.f., Tab 5). Please refer to Tab. 6 to see the different combinations.

[11] i.e., Symbol of water droplets popping out next to the TE's text

Implementation detail A wooden ball appears on a platform in front of the participant. The participant must interact with the ball (e.g., hit it) for 7s. After 7s, the ball disappears, and the questions appear. Each time the participant hits the ball, the sound and the TE start/are instantiated at the hit point for 0.3s before stopping/disappearing. A hit is registered by moving the controller towards the object with adequate speed (approx. 20 cm/s). Unlike step A1 there is no time limit to answer the questions (because there are several ordered questions, the properties are more difficult to identify, and this isn't necessary to answer H3). We record the answers.

5.4.3 Step B3

Here we try to answer H4. We play a sound and display a TE, then wait a bit, then play another sound and display another TE. Both sounds can have a different volume, and both TE can have different sizes. We then ask the participant if the last sound volume was lower or the same or higher than the first. The main idea is to know if the TEs size can bias users' perception of sound volume (i.e., a bigger text would lead to higher volume perception, etc.).

Implementation detail The participant is in front of a speaker. The speaker plays a "Car horn (short)" sound at a given volume (noted $\text{Volume}_{\text{FIRST}}$) and displays a TE-CarHorn-Short at a given size (noted $\text{Size}_{\text{FIRST}}$). After a time interval of [0.5s,2s], the speaker plays another "Car horn (short)" sound at a given volume (noted $\text{Volume}_{\text{LAST}}$) and displays another TE-CarHorn-Short at a given size (noted $\text{Size}_{\text{LAST}}$). Then the participant answers if $\text{Volume}_{\text{LAST}}$

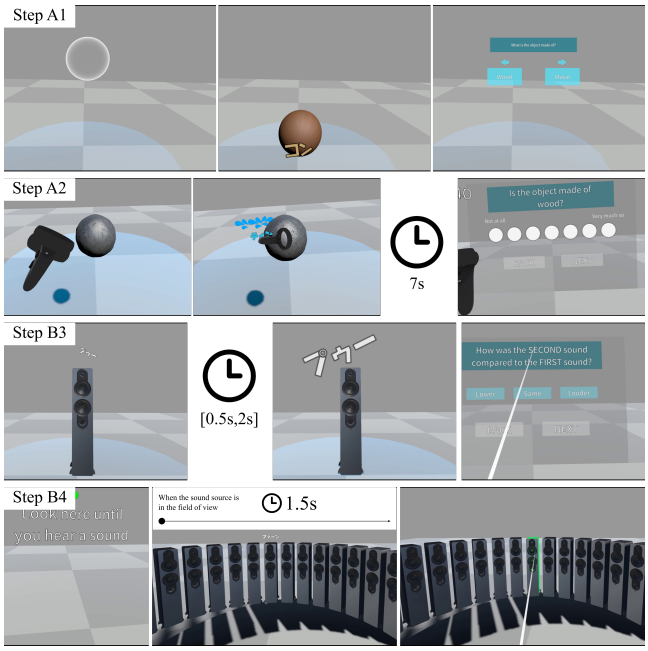

Figure 3: The 4 steps. From top to bottom (each row corresponds to a step): step A1, step A2, step A3, and step A4.

Condition name	Are x and y congruent?			Is TE?
	M and S	M and TE	S and TE	
M-S-~~TE~~	Yes	N/A	N/A	No
M-$\overline{\text{S}}$-~~TE~~	No	N/A	N/A	No
M-S-TE	Yes	Yes	Yes	Yes
M-S-$\overline{\text{TE}}$	Yes	No	No	Yes
M-$\overline{\text{S}}$-TE	No	Yes	No	Yes
$\overline{\text{M}}$-S-TE	No	No	Yes	Yes

Table 4: The congruence conditions for the step A1. M stands for material and S for sound, therefore M-S-TE stands for: Material - Sound - TE. Letter(s) with a bar are non-congruent with the others, therefore M-S-$\overline{\text{TE}}$ stands for: the material and sound are congruent, but not the TE. If a condition is crossed over, then it wasn't used in that condition.

was lower, the same, or higher than $\text{Volume}_{\text{FIRST}}$. We record the answer. 4 volume conditions for the sounds were used: very low volume, low volume, high volume and very high volume. And 3 size conditions for the TEs were used: small size (scaled at 0.3), normal size and big size (scaled at 1.7). Before the experiment, we defined the discrimination of the different volume settings in a small test (N=3). The volume settings were separated proportionally alongside a logarithmic scale: very low volume (62 dbSPL), low volume (65 dbSPL), high volume (70 dbSPL) and very high volume (72 dbSPL). To normalize the sound levels between local and remote participants, only over ear headphones were used and the PC volume level was set at 60 %.

5.4.4 Step B4

Here we try to answer H5. We play a sound at a position and display the TE at a possibly different position. Then we ask the participant to localize the sound position.

Implementation detail The participant stands in front of an empty space. 18 speakers are positioned behind them. After 3s, a given speaker (noted $\text{Speaker}_{\text{SOUND}}$) plays a "Car horn (long)" sound, and a given speaker (noted $\text{Speaker}_{\text{TEXT}}$) displays a TE-CarHorn-Short. The participant then turns back to try to locate

ID	Item	Lowest value	Highest value
M1	Is the object made of wood?	Not at all	Very much so
O1	Was the object...	Wet	Dry
O2	Was the object...	Cold	Warm
O3	Was the object...	Soft	Hard
O4	Was the object...	Smooth	Rough
S1	Was the sound...	Dull	Sharp

Table 5: The questions for the step A2. We ask about the material properties (M), the object properties (O) and the sound properties (S), and all are on a 7 points Likert scale.

Sphere material + sound	TE	Is water droplet
Wood	TE-Wood	No
Wood	TE-Wood	Yes
Wood	TE-Water	No
Wood	TE-Water	Yes

Table 6: The different conditions for the step A2.

$\text{Speaker}_{\text{SOUND}}$. $\text{Speaker}_{\text{SOUND}}$ continues playing the sound while the user is turning, but stops playing it 1.5s after $\text{Speaker}_{\text{TEXT}}$ appears in the participant's field of view. Just like in step B3, the TE is displayed on top of the speaker. Once the sound stops playing, the participant guesses which speaker is $\text{Speaker}_{\text{SOUND}}$, by selecting one of the speakers. We record the answer. Placing the speaker behind the speaker allows us to potentially reduce the strength of the TE's bias. By first having only access to auditory information, we can alleviate the case where users would just follow the TE blindly and select $\text{Speaker}_{\text{TEXT}}$ without paying attention to the sound. The speakers are arranged equidistantly on a 170° slice of a 1.65m radius circle. The speaker offset (i.e., the number of speaker from $\text{Speaker}_{\text{TEXT}}$ to $\text{Speaker}_{\text{SOUND}}$) can be 0 (same speaker), 1 (directly to the left or right), 3 (separated by 2 speakers) and 4 (separated by 3 speakers). We compare the TE-CarHorn-Long, the TE-SymbolSound and absence of TE at the different speaker offsets.

6 Results

We present the results of each step, as well as the results of the NASA-TLX.

6.1 NASA-TLX

The scores of the questions QA-6,7,8,9 are as follow: Mental demand $M = 51.1$, $SD = 29.9$; temporal demand $M = 50.36$, $SD = 32.9$; effort $M = 47.5$, $SD = 31.1$; frustration $M = 43.2$, $SD = 21.8$. This is a somewhat high workload for those items, but not a very high workload (> 70). The participants were therefore not overwhelmed by the TEs.

6.2 Step A1 - Perception of Material Category

We removed data with a RT < 0.10ms and > 3s, since they were deemed respectively too fast or too slow for a fast classification experiment.

About the Reaction time. We grouped the reaction time by participants and conditions (i.e., $\overline{\text{M}}$-S-~~TE~~, M-S-~~TE~~, $\overline{\text{M}}$-$\overline{\text{S}}$-$\overline{\text{TE}}$, $\overline{\text{M}}$-S-TE, M-$\overline{\text{S}}$-$\overline{\text{TE}}$, $\overline{\text{M}}$-S-$\overline{\text{TE}}$, M-$\overline{\text{S}}$-$\overline{\text{TE}}$; as described in Tab. 4), and calculated the mean of the RTs in each group. We compared the RT of each participants for each conditions with a repeated measure ANOVA. There was a significant difference between the conditions: $F(5,70) \approx 7.62$, $p < 0.00$. We then ran post-hoc test for pairwise t-test. There were several significant differences between the conditions as seen in Tab. 7.

The RT is therefore significantly faster when a TE is congruent with the material / sound than when it is not (c.f., Fig. 4) or than

	M-S-~~TE~~	M-~~S~~-~~TE~~	M-S-TE	M-S-$\overline{\text{TE}}$	M-$\overline{\text{S}}$-TE
			$T(14) = ?; p = ?$		
M-S-~~TE~~	$-2.80; \mathbf{0.014}$	-	-	-	-
M-S-TE	$-3.10; \mathbf{0.008}$	$-3.98; \mathbf{0.001}$	-	-	-
M-S-$\overline{\text{TE}}$	$-1.86; 0.084$	$-1.47; 0.163$	$-3.97; \mathbf{0.001}$	-	-
M-$\overline{\text{S}}$-TE	$-3.20; \mathbf{0.006}$	$-0.459; 0.654$	$-4.98; < \mathbf{0.001}$	$-1.63; 0.125$	-
$\overline{\text{M}}$-S-TE	$-2.66; \mathbf{0.019}$	$+0.367; 0.719$	$-3.67; \mathbf{0.003}$	$-1.28; 0.222$	$-0.122; 0.904$

Table 7: The p-value for the pairwise comparison of the reaction time. M = material; S = sound; M-S-TE = presence of Material-Sound-TE. Overlined letters are non-congruent with the others; for example M-S-$\overline{\text{TE}}$ = material and sound are congruent, but not the TE. Crossed letters = absence of the property.

when there no TE at all. This faster RT means that participants are faster to categorize objects when there is a congruent TE.

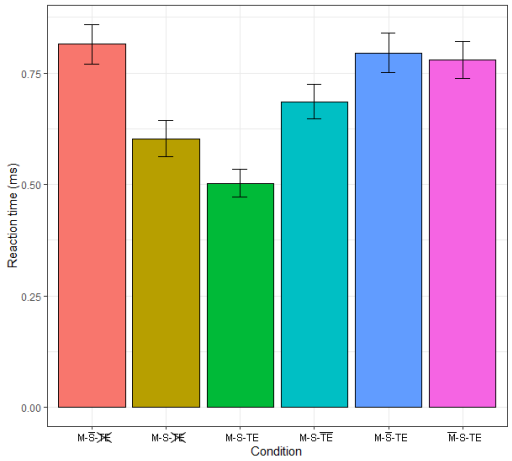

Figure 4: Reaction time by congruence condition.

About the answers. We grouped the answers by participants and conditions, then counted the number of congruent and non-congruent answers with the material (e.g. an answer "Wood" when there is a wooden looking sphere is a congruent answer, else if the answer is "Metal" then it is a non-congruent answer).

We compared the number of congruent answers with the material (c.f., Fig. 5) with a non parametric Friedman test of difference among repeated measures (the number of congruent answer with the material of each participants for each condition). There was a significant difference between the conditions ($\chi^2(5) \approx 50$, $p < 0.001$). We then ran post-hoc test for pairwise comparison Conover test with a Bonferroni correction. There was no significant differences between the 3 combinations of pair between: M-S-~~TE~~, M-S-TE and M-S-$\overline{\text{TE}}$, nor between the 3 combinations of pair between: M-$\overline{\text{S}}$-~~TE~~, M-$\overline{\text{S}}$-TE and $\overline{\text{M}}$-S-TE. The other combinations were all significantly different ($p < 0.001$). When the material and sound were congruent, there was no significant difference in the categorization of an object, whether the TE was congruent with the material, non-congruent, or not at all. Those results mean that the "strength" of the material/sound on the perception makes the TE impact non significant when categorizing an object. Moreover, when the sound was non-congruent with the material, there still was no significant difference. This lack of difference means the TE did not significantly "shift" the answer toward the material or the sound it was congruent with.

6.3 Step A2 - Perception of Material Properties

The results of each question performed in the step A2 were plotted on individual interaction plots, where we compared the relationship

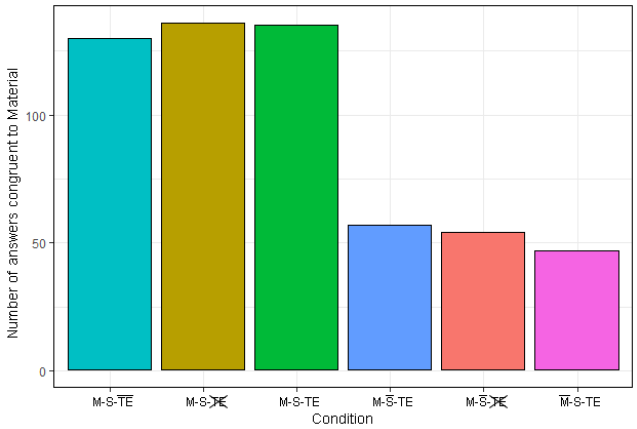

Figure 5: The number of answers congruent with the Material for each conditions

between the TE and the water droplets. We present the interaction plots in Fig. 6.

We ran a non parametric 2-way ANOVA repeated measure using ANOVA of Aligned Rank Transform (ART) [30] with the ARTool toolkit [11] on every question. We followed with a planned contrast with a Tukey adjustment for post-hoc comparisons. There was significant differences for M1, T1, T2 and T3, we show the post-hoc results in Tab. 8.

Table 8: Post Hoc test for A2

Question	Comparison	t ratio	p-value
M1 (wood?)	TE	-2.56	**0.0111**
	droplet symbol	-0.44	0.6603
	TE:droplet symbol	-1.38	0.1697
T1 (wet/dry)	TE	-7.00	< **0.0001**
	droplet symbol	16.8	< **0.0001**
	TE:droplet symbol	-3.25	**0.0013**
T2 (cold/warm)	TE	-3.33	**0.0010**
	droplet symbol	6.13	< **0.0001**
	TE:droplet symbol	-1.13	0.2579
T3 (soft/hard)	TE	-3.25	**0.0013**
	droplet symbol	1.59	0.1130
	TE:droplet symbol	-1.91	0.0566

T1 presents some interesting results. On T1 (wet/dry) the spheres with the TE-Water was perceived as significantly wetter than the ones with the TE-Wood. Also the spheres with the water droplets were perceived as significantly wetter than the ones without water droplets. Moreover there was a small significant interaction between the TE and the water droplets.

6.4 Step B3 - Perception of Sound Volume

The results of step B3 were grouped by the volume change between $\text{Volume}_{\text{FIRST}}$ and $\text{Volume}_{\text{LAST}}$ ("Higher", "Same" and "Lower"). Also, the TE were grouped by the size change between the $\text{Size}_{\text{FIRST}}$ and the $\text{Size}_{\text{LAST}}$ ("Grow", "No size difference" and "Shrink"). The average interaction plot between the volume change and the user response to the perceived loudness is shown in Fig. 7.

We ran a 1-way between subjects ANOVA to know if the TE size change had an effect on the perceived loudness. There was a significant difference: $F = 15.0, df = 2, p < 0.001$. Then we ran a post-hoc tests for pairwise comparison Tukey HSD test. The data was separated by volume condition and the test performed on each volume condition. The alpha value was adjusted to $\alpha = 0.016$ $(0.05/3)$ to consider the repeated tests. In the Lower volume condition, there was no significant difference. However, there was a

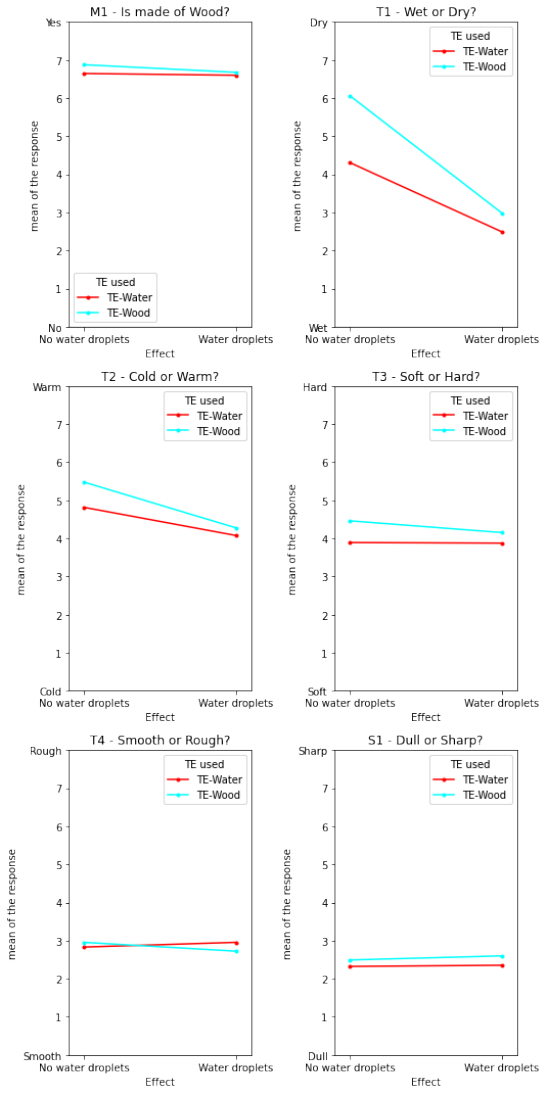

Figure 6: The interaction plots for the step A2. The vertical axis of the plots indicates the mean of the user response over the Likert scales. The contrasting adjectives are indicated on the vertical axis of the plots (e.g., dry/wet).

significant difference in the Same and in the Higher volume conditions when comparing Growing and Shrinking TEs, meaning that users were generally more inclined to rate louder sound associated with bigger TEs as louder than those associated with smaller TEs, even at equal volume. This also means that the relative increase in the TE size also plays a role, as comparison with similar sized TEs did not yield significant results. Overall, the size change is consistent with the perceived loudness (when there is grow, the perceived loudness is higher than when there is a shrink).

6.5 Step B4 - Sound localization

The results of step B4 were transformed along the distance between the selected speaker and $\text{Speaker}_{\text{TEXT}}$ (the distance text). The average interaction plot between the speaker offset is shown in Fig. 7.

We ran a non parametric 2-way ANOVA repeated measure using ANOVA of Aligned Rank Transform (ART) [30] with the ARTool toolkit [11] to know if the TE had an effect on the distance text or distance sound. We present the results in Tab. 10.

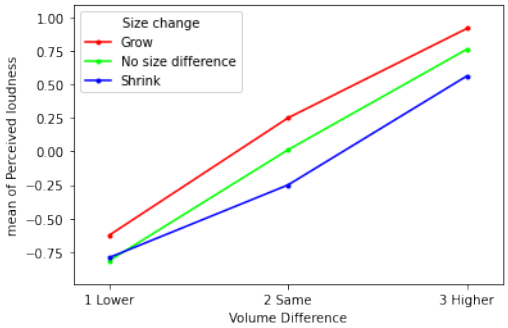

Figure 7: Interaction plot of the step B3. On the Y axis, 1 means $\text{Volume}_{\text{LAST}}$ was perceived in average as higher than $\text{Volume}_{\text{FIRST}}$, 0 as the same, and -1 as lower.

Table 9: Post Hoc test for B3

Volume condition	Comparison	M	p-value
Lower	Grow vs No size diff	-0.19	0.03
	Grow vs Shrink	-0.16	0.176
	No size diff vs Shrink	0.02	0.9
Same	Grow vs No size diff	-0.24	0.018
	Grow vs Shrink	-0.5	**0.001**
	No size diff vs Shrink	-0.26	0.082
Higher	Grow vs No size diff	-0.15	0.096
	Grow vs Shrink	-0.35	**0.001**
	No size diff vs Shrink	-0.2	0.021

Then we ran a post-hoc analysis for multiple comparisons using planned contrast with a Tukey adjustment. There are significant differences between TE-CarHorn-Long and None as well as between TE-SoundSymbol and None (resp. $p < 0.0025$ and $p < 0.001$). The distance offsets are statistically significant between every pair except 0-1. Therefore when there is a TE, the selected speaker is significantly closer from the $\text{Speaker}_{\text{TEXT}}$ although the offset appears to be somewhat constant (as there is no interaction between the speaker offset and the TE).

7 Discussion

In step A1, we focused on answering H1 and H2. We showed that the participants: (i) are affected in their RT by the TE when classifying the object; (ii) considered whether the TE was congruent with the material/sound when classifying an object; and (iii) are not biased by the TE in their classification. Regarding the (ii) above, the participants are significantly faster when classifying an object with a congruent TE than when classifying an object with a non-congruent TE or even classifying an object with no TE visible. Therefore, the TE is taken into account when classifying an object. Regarding the (iii) above, the TE did not significantly change the number of answers for a congruent material/sound. We can therefore assume that it is likely that the participants do not prioritize a TE when classifying an object, especially if they already have enough "cues" from a congruent material/sound. QA-5 supports this possibility ("Did you focus more on the text effects, the objects' material, the sound or something else? [...]"): the participants mostly answered that they focused first on the material, then on the sound, then on the TE. Also, the TE did not significantly change the number of answers for a non-congruent material/sound. Therefore, we can assume that this is not only a matter of prioritizing the material/sound over the TE but that the TE additional data is very weak next to the ones from the material/sound when it comes to the categorization. Even when in doubt in front of a non-congruent object, the participants will still not rely on the TE. A possibility is that the material and the sound are not strictly equal in term of influence on the perception, hence the participants were not necessarily in doubt. 2 contradictory results

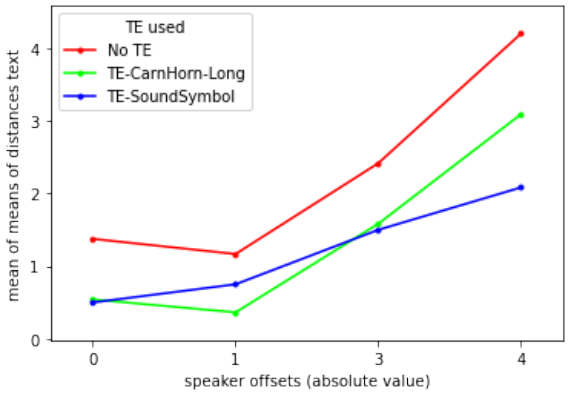

Figure 8: Interaction plot of the step B4.

Table 10: ANOVA of Aligned Rank Transform for B4

Distance X	Factor	F	p-value
Text	Speaker offset	F(3,10) = 23.1	$<$ **0.001**
Text	TE	F(2,11) = 13.6	$<$ **0.001**
Text	Speaker offset:TE	F(6,7) = 0.932	0.473

support this: the answers from QA-5 were participants said they focused first on the material; and the number of answers congruent with the material, when the material/sound were not congruent, were less than the number of answers congruent with the sound (about 50 against 100).

In step A2, we focused on answering H3. There were numerous significant differences in the perception of an object's properties. For example, the TE-Water conveyed a perception of wetness to the object. Moreover, the addition of water droplets to an otherwise "dry" TE conveyed a similar perception of wetness, or even of cold. It hints that the participants consider the TE in its entirety, and focus on the textual characters and the symbolic characters. However, not every property is affected by the TE or by its water droplets. The results hint that the TE has to be somewhat related to the property we want to influence. Indeed, roughness or sharpness, that are not typically associated with water, were not affected at all. This shows that the properties of a TE are perceived as transferable to the object's properties. It also hints that a TE and the object evolve in the same "perceptual space", meaning the participants might perceive both of them as a whole, and not as separate objects. The significant results of M1 and T3 likely follows the same logic, the TE-Wood feels more like the object is "Made of wood" (M1 might arguably be more related to the category than the property); and the TE-Wood feels more like the object is hard (since wood is harder), yet more work is needed on such "non-obvious" properties.

In step B3, we focused on answering H4. The statistical results indicate that the TE size can bias an audio cue's sound level perception of the participants. When the participants are comparing volumes, they tend to be biased by the changes in visual cue, if the change is significant enough (here, the TE was almost x6 bigger/smaller).

In step B4, we focused on answering H5. We showed that TE affects the audio localization (except when there was very little distance difference between the audio and the TE). This is the excepted consequence of the ventriloquism illusion. However, there was no significant effect between the 2 TEs (one was a text and the other a symbol). The text of the sound, therefore, had no significant incidence on the bias.

Onomatopeias have been mostly confined to comic books, with the occasional application to other mediums. We raise the possibility of using them in a 3D setting and use their compact information and signification to help with scene understanding for a limited cost of development. While this study mainly focuses on VR, we plan on furthering our research in AR. More in-depth analysis of each step should also be pursued to understand the possibilities and limitations of TEs fully. TEs can be useful for entertainment purposes (games, movies, ...) and for the hearing impaired part of the population. Thanks to the linguistic properties of onomatopoeia, if placed and used correctly, TEs can increase the situational awareness of hearing-impaired people. Further research is therefore also needed to fully grasp the potential of virtual text effects for hearing impairment.

8 Limitations

The participants guessed for the most part what we were trying to accomplish, as we noted when reviewing QA-1 and QB-1. This is a relatively typical and expected limitation. Despite our effort to select the most accurate texts for the VSE, about half of the Japanese participants suggested to use キン /kiːn/ or チン /tʃiːn/ or カキン /kakiːn/ instead of ティン /tɪn/ for the SFX-Metal. The rest of the texts were mostly well received. Moreover, because of the pandemic at the time of the experiment, we allowed the participants to do the experiment remotely. This is a cause for concern, mainly because the audio level and audio quality are somewhat uncertain for the remote participants, and this was important for B3 and B4. We did ensure that the remote conditions were as close as possible as the local ones, but this is nonetheless a limitation. Finally, we did not implement a Head Related Transfer Function (HRTF) for the sound in B4. While we believe our setup was sufficient for our need (because we only use a horizontal panning for B4, and because upon testing we were able to locate sound precisely), HRTF is the more suitable candidate for sound positioning, and would have been very suited for the step B4.

9 Conclusion

This research aimed to explore the impact of Textual Sound Effects on the VE perception in VR. We highlighted the potential of TEs as a visual tool to make full use of audiovisual cross-modal bias in a VR scene. While the influence of TEs isn't strong enough to influence our perception of the category of an object (H2), we show that it is still being registered as an added stimuli and being processed by people, as demonstrated by the overall increase in their reaction time (H1). While the observed difference in mean reaction time between the presence and absence of a congruent text to a congruent audiovisual stimulus is significant, there is a matter of prioritization of material and sound over TE.Furthermore, the added information that a TE may bring can change the perception of object properties (H3). While we mainly tested for the "Wetness" of a material, different TEs could change the perception of other properties yet to be determined. TEs can also change the perceived loudness of a sound (H4) and its perceived spatial location (H5). Our findings show that TEs can be used not only for their design but also for their perception impact to, for example, recognize more easily surrounding objects and their properties. Here, we only focused on onomatopeia-like words and did not investigate each part of the TE (like the word, color or animation). Such investigations are part of our future work. Moreover, we believe the participants integrated the TE as part of the object's property (rather than a separated element). It might be interesting to see whether the TE is indeed perceived as part of the object or not, and then whether the TE is associated with the object receiving the impact, or the one giving the impact (even if it is our own body).

Acknowledgments

This project was supported by JST ERATO Grant Number JPMJER1701, Japan.

REFERENCES

[1] S. Boer. The Sounds of Violence : Textualized sound in Frank Miller ' s Sin City and Batman : The Dark Knight Returns. 2017.

[2] N. Bonneel, C. Suied, I. Viaud-Delmon, and G. Drettakis. Bimodal perception of audio-visual material properties for virtual environments. *ACM Transactions on Applied Perception*, 7(1):1–16, jan 2010. doi: 10.1145/1658349.1658350

[3] A. Brown, J. Turner, J. Patterson, A. Schmitz, M. Armstrong, and M. Glancy. Exploring Subtitle Behaviour for 360°Video. Technical report, WHP 330, 2018.

[4] H. Choi, J. Oh, M. Chang, and G. J. Kim. Effect of accompanying onomatopoeia to interaction sound for altering user perception in virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, pp. 1–2. ACM, New York, NY, USA, nov 2018. doi: 10.1145/3281505.3281614

[5] W. Fujisaki, N. Goda, I. Motoyoshi, H. Komatsu, and S. Nishida. Audiovisual integration in the human perception of materials. *Journal of Vision*, 14(4):12–12, apr 2014. doi: 10.1167/14.4.12

[6] B. L. Giordano and S. McAdams. 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, 2006. doi: 10.1121/1.2149839

[7] M. Grassi, M. Pastore, and G. Lemaitre. 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, May 2013. doi: 10.1016/j.actpsy.2013.02.005

[8] S. A. Guynes. Four-Color Sound: A Peircean Semiotics of Comic Book Onomatopoeia. *The Public Journal of Semiotics*, 6(1):58–72, 2013.

[9] S. G. Hart and L. E. Staveland. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In *Advances in Psychology*, pp. 139–183. Elsevier, 1988. doi: 10.1016/s0166-4115(08)62386-9

[10] P. Hauck and H. Hecht. The Louder, the Longer: Object Length Perception Is Influenced by Loudness, but Not by Pitch. *Vision*, 3(4):57, oct 2019. doi: 10.3390/vision3040057

[11] M. Kay. Package 'artool'. `hcran.r-project.org/web/packages/ARTool/ARTool.pdf`. Accessed: 2020-09-15.

[12] R. L. Klatzky, D. K. Pai, and E. P. Krotkov. Perception of material from contact sounds. *Presence: Teleoperators and Virtual Environments*, 9(4):399–410, Aug. 2000. doi: 10.1162/105474600566907

[13] K. Knoeferle, J. Li, E. Maggioni, and C. Spence. What drives sound symbolism? Different acoustic cues underlie sound-size and sound-shape mappings. *Scientific Reports*, 7(1):5562, dec 2017. doi: 10.1038/s41598-017-05965-y

[14] S. Malpica, A. Serrano, M. Allue, M. G. Bedia, and B. Masia. Crossmodal perception in virtual reality. *Multimedia Tools and Applications*, 79(5-6):3311–3331, feb 2020. doi: 10.1007/s11042-019-7331-z

[15] J. Oh and G. J. Kim. Effect of accompanying onomatopoeia with sound feedback toward presence and user experience in virtual reality. In *Proceedings of the 24th ACM Symposium on Virtual Reality Software and Technology*, number 1, pp. 1–2. ACM, New York, NY, USA, nov 2018. doi: 10.1145/3281505.3283401

[16] N. K. Pratha, N. Avunjian, and N. Cohn. Pow, Punch, Pika, and Chu: The Structure of Sound Effects in Genres of American Comics and Japanese Manga. *Multimodal Communication*, 5(2):93–109, jan 2016. doi: 10.1515/mc-2016-0017

[17] V. S. Ramachandran and E. M. Hubbard. Synaesthesia – a window into perception, thought and language, 2001.

[18] G. H. Recanzone. Rapidly induced auditory plasticity: The ventriloquism aftereffect. *Proceedings of the National Academy of Sciences*, 95(3):869–875, Feb. 1998. doi: 10.1073/pnas.95.3.869

[19] S. Rothe, D. Buschek, and H. Hußmann. Guidance in Cinematic Virtual Reality-Taxonomy, Research Status and Challenges. *Multimodal Technologies and Interaction*, 3(1):19, mar 2019. doi: 10.3390/mti3010019

[20] S. Rothe, K. Tran, and H. Hußmann. Dynamic Subtitles in Cinematic Virtual Reality. In *Proceedings of the 2018 ACM International Conference on Interactive Experiences for TV and Online Video - TVX '18*, pp. 209–214. ACM Press, New York, New York, USA, jun 2018. doi: 10.1145/3210825.3213556

[21] R. Rzayev, S. Mayer, C. Krauter, and N. Henze. Notification in VR. In *Proceedings of the Annual Symposium on Computer-Human Interaction in Play*, pp. 199–211. ACM, New York, NY, USA, oct 2019. doi: 10.1145/3311350.3347190

[22] E. Sapir. A study in phonetic symbolism. *Journal of Experimental Psychology*, 12(3):225–239, 1929. doi: 10.1037/h0070931

[23] scopatgames. vrkshop. `https://store.steampowered.com/app/1344530/vrkshop/`. Accessed: 2021-05-26.

[24] M. Shibatani. *The Languages of Japan*. 05 1990.

[25] L. Sidenmark, N. Kiefer, and H. Gellersen. Subtitles in Interactive Virtual Reality: Using Gaze to Address Depth Conflicts. Technical report.

[26] D. M. Sidhu and P. M. Pexman. Five mechanisms of sound symbolic association. *Psychonomic Bulletin & Review*, 25(5):1619–1643, oct 2018. doi: 10.3758/s13423-017-1361-1

[27] M. Takashima. Perceived weight is affected by auditory pitch not loudness. *Perception*, 47(12):1196–1199, Oct. 2018. doi: 10.1177/0301006618808937

[28] R. Varnum. *The language of comics : word and image*. University Press of Mississippi, Jackson, 2001.

[29] Wikipedia. Xiii (video game), 2020.

[30] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI '11*. ACM Press, 2011. doi: 10.1145/1978942.1978963