Internal use

Audio Engineering Society

Convention Paper 10693

Presented at the 156th Convention

2024 June 15-17, Madrid, Spain

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Presentation Mode and its Impact on Sentiment in Free Verbalization Responses to Sounds

KC Collins[1], Hannah Johnston[1], Adel Manji[1] and Bill Kapralos[2]

[1] *School of Information Technology, Carleton University, 1125 Colonel By Dr. Ottawa, ON Canada*

[2] *Faculty of Business and Information Technology, Ontario Tech University, 2000 Simcoe St. N. Oshawa, ON, Canada*

Correspondence should be addressed to KC Collins (kccollins@cunet.carleton.ca)

ABSTRACT

With the increase in online testing being used for sound perception research, it is necessary to explore whether changing the presentation mode impacts the responses to sound. We ran three different experiments, audio-only (online), audio with video (online), and in person with a sound-making object (a robot). We tested four different versions of five sounds (human and robot emotes). Participants used free verbalization to respond to the question of what the sound was communicating to them. Results showed that, when it comes to sentiment analysis, there was no statistical difference between presentation modes. However, there were differences in responses, including significantly more onomatopoeia used in the audio-only condition.

1 Introduction

More and more research is being conducted online as technological enhancements have led to an increased use of online tools across a variety of disciplines. When it comes to sound, however, there are unique problems and questions that arise in conducting research online, since audio playback devices can vary greatly depending on the end user's equipment. While there has been a growing body of research exploring solutions to accommodate this variation in playback devices for online audio research, there are many questions yet to be explored specific to conducting online sound-based studies. Moreover, when audio studies are conducted, online or in person, will the accompaniment of imagery, or a live object associated with a sound, alter the perception of that sound, when compared to sound without any image? In other words, when testing a product sound, should an image or video be developed and/or should the test be conducted in person, or is the sound alone enough for some research questions?

We present here three experiments that sought to explore these questions with a series of sounds for robots. The experiments used a free verbalization method to gather responses to sounds presented to the listeners. The first experiment was conducted entirely online and consisted of only the sounds. The second experiment was also conducted online, and included video context to the sounds. The third experiment was conducted in person with a live robot making the sounds.

We used natural language processing tools to analyze the qualitative data, comparing the results to human analysis. We previously presented part of our first, audio-only experiment [1]. The aim in our second and third experiments, compared here with the results from the first, was to determine whether varying responses to the sound under three different conditions would result. We hypothesized that the presence of a robot making the sound via video would alter the responses, and we also hypothesized that the responses to the live robot would differ from the responses to the online robot.

We first present the background in terms of the differences found in prior studies of online versus live presentation, image or sound alone, free verbalization

Internal use

approaches, and sentiment analysis. Then, we present the study design and our analysis procedure, followed by our results. In our first paper we noticed some sounds gave rise to more onomatopoeia keywords, more use of emotional keywords, and/or more anthropomorphic keywords: As such, we also compared if there was greater use of these keywords under the different presentation conditions.

2 Background

There are several difficulties with running studies online, particularly when it comes to sound. With respect to hardware, there is no consistency and thus quality and loudness levels can vary greatly, potentially influencing results [2]. More generally when running studies online, experimenters might exclude certain contextual cues that can be observed during in-person sessions, such as body language, non-verbal cues, and environmental factors. There are also higher participant drop-out rates, online studies are open to fraud [3], and, since they are usually anonymous, online participants may be less motivated to perform to the best of their abilities [4]. However, online testing is generally more cost-effective, since it eliminates the need for physical space, travel expenses, and equipment setup and supervision, also saving time. It is also much easier to recruit participants for online studies, leading to a larger number of volunteers, and thus usually higher statistical power.

Research into online versus offline product usability testing has found that online testing is effective for testing usability and functionality, but less useful for appearance [5][6]. One study found that physical models received more positive feedback than virtual models, and that the evaluation methods significantly influenced the result [7]. In music-based perception testing, studies have shown the validity of remote testing [8][9]. When it comes to auditory psychophysics, however, online/remote testing has been mostly avoided [10]. There are lower levels of internal validity in online audio studies, since there is so much out of experimenter control. For instance, it is difficult to know the hearing capabilities of the end user, and especially difficult controlling for their external environment and hardware. Although there are measures that can be employed to select participants for hearing status [11], they may not even be aware of how external noise might interfere with the results. Accurate auditory functioning screening requires considerable equipment and expertise, and therefore, at best we can only obtain approximate capabilities [12]. There is also much more technical variance between users in the quality of their equipment, their playback devices, as well as the volume levels [13][14], which may lead to significant differences in what they hear [10]. The volume level in particular has been shown to influence test results, depending on what is being tested [15][16]. Different set-ups and incorrect calibrations between participants can impact some test results, depending on the stimuli and the question being researched [16]. Remote testing should therefore take some precautions, including ensuring the user will not be exposed to harmful volume levels, and verifying the hardware meets the required stimulus quality [10].

When it comes to sound, there are some recently developed testing protocols for internet experiments designed to account for variations in the user's playback devices for internet-based experiments, to assist in standardizing volume levels, detect playback formats, and provide basic equipment tests and calibration [13][14][17][18][19]. However, when it comes to perceptual, connotative/semiotic responses to sounds, rather than sound quality, there is no available research on the validity of online research conducted by comparing online responses with live, "real-world" responses.

Decades of research into multimodality suggest that sound when presented with an image will have different perceptual responses than sound presented alone: Well-known cross-modal impacts and perceptual illusions such as the McGurk effect [20], the ventriloquist effect [21], and more recently the auditory-induced flash [22] suggest that the accompaniment of a moving image to a sound can impact both our auditory and visual perception. Studies have shown that sound can impact visual fidelity perception [23], movement perception [24], task response time [25], and preferences [26][27][28][29].

Despite the abundance of research into image and sound's interaction, to our knowledge, there has been little research into how the presence of a moving image, compared to the live object itself, may impact user preferences or feelings towards an object, or emotions evoked by a sound. There is also little research on how that object's presence may impact the perception of the sound in terms of communicated meanings, emotional response or mental imagery evoked.

Most of the research into sound's meanings has developed from either marketing or musicological

Internal use

research (for a recent review, see [30]). There are no well-established methods to answer, "what does this sound mean to you," or, "what is this sound communicating to you?" Free verbalization is one method that may be useful in answering these questions. The user is presented with a task or object, and then asked to spontaneously respond to that object, in their own words. The method has been used, for instance, in analyzing consumer responses to food [31] and music [32]. Free verbalization has the benefit of telling us what sound characteristics users are focusing on—the semantic meaning, the acoustic properties, etc. and as such is commonly used as a first step in determining attributes for semantic differential scales [33]. There are challenges with the free verbalization approach, particularly with respect to analyzing the responses [34]. Automated approaches promise faster analysis of large bodies of text, and the recent rise of tools to assist in the process makes this method more feasible than in the past. One of the most common of these approaches is sentiment analysis.

Sentiment analysis, sometimes called opinion mining, is a natural language processing approach commonly used on text data to analyze blocks of text for their overall sentiment, attitudes and emotions. Sentiment analysis is typically used for large-scale text to categorize an overall response into positive, negative or neutral. It is used, for instance, to take social media chatter about a topic and determine how the overall sentiment may be shifting. There are different techniques used to process the text [35], but most recently a variety of machine learning approaches have been used [36].

In our prior, audio-only study of 68 sounds [1], we ran a sentiment analysis on a block of text through several different software packages, including SpeakAI, Textrics, and IBM Watson and compared the resulting output to human results.[1] Those results showed relative consistency across the tools and human evaluators: That is, they were consistent enough for our purposes, but there were some small differences and some confusing results, showing there is still a need for human intervention in the analysis.

3 Methods

For this study, we used the sound of robot *emotes*, non-verbal vocal sounds, sometimes referred to as non-linguistic utterances, as we were simultaneously testing robot sounds for another study. These sounds also allowed us to explore the different responses to human and synthesized sounds, as well as entirely abstract sounds (beeps, etc.).

Initially, seventeen sound files were recorded in four formats; raw human voiceover ("raw"), and three types of robot voice: vocoded voice 1 ("robo"), vocoded voice 2 with musical keyboard ("kbd"), and a "beep" voice. The "robo" voice approximately resembles the sound of Pixar's robot Wall-E who communicates with a semi-vocal synthesized "language", and the "beep" voice referenced the sound of *Star Wars'* R2-D2 robot (who communicates with beeps and chirps).

Each sound was recorded as 44.1kHz, 24-bit wav files in a professional recording studio. The raw voiceover was recorded by professional voice actor DB Cooper, who has been the robot/AI voice for several video games, and the voice of the BMW DEE internal car AI voice system. Cooper recorded the emotes using a Sennheiser MKH-416. Professional sound designer pdx Drescher, an expert in interface sound design, created three sets of robot voices from the original voice files. pdx was tasked with trying different approaches to turning the samples into three different types of robot voice. The first set, "robo", used clips from one of pdx's prior robot voice projects, modifying them to mimic the prosody and emotional intention of the voiceover. For the "kbd" takes, Vocoder patches modified the signal by original voice formants, and the pitch was determined by MIDI notes and pitch-bend controllers. The output of the synthesizer was then edited with additional synth patches and effects. He made particular use of a plugin called Envy by Cargo Cult,[2] which takes the volume, pitch, and EQ envelopes of one sound (the original voiceover) and applies them to another sound: This helped make the synth resemble the prosody of the original sound. The "beep" sounds underwent a similar development process as the robot takes, but with interface "bleeps and bloops" derived from various sound effects libraries, including the *Star Trek LCARS* sound set.

[1] SpeakAI: https://speakai.co/ Textrics: https://www.textrics.ai/ Watson: https://www.ibm.com/products/natural-language-understanding

[2] Envy: https://www.thecargocult.nz/products/envy

Internal use

These sound groups increase in abstraction, where the raw sounds are the original untreated sound, (the least abstracted), and the beep voice is the most abstracted. We then selected five sounds (in all four types—raw, robo, kbd and beep) from the first study, to use for the experiments presented here:

a) Be careful/warning calm
b) Wrong/ error
c) Confused
d) Calculating
e) Giggle

In other words, with four types, a total of 20 sounds were presented to our participants in the experiments, and all listeners heard all sounds in Experiments 2 and 3.

The robot employed in our study was a Lego Mindstorms Robot Inventor 51515 in the configuration called "Blast". This particular robot presented what we felt was a relatively neutral presence—that is to say, it was neither particularly cute nor threatening. It stands approximately 36 cm tall. The in-built audio was not suitably loud nor of a high enough quality for our purposes, so we custom-built a backpack to hold an external Bluetooth loudspeaker (EWA A106 Mini). The robot's movements were preprogrammed to play the same sound-action pair to each participant, but in randomized blocks. The movements consisted of a combination of moving forward, moving backward, raising arms, moving head, and turning around.

Participants were all over the age of 18. We explored demographics in our first study, but found no correlation between demographic information and the results, and therefore did not collect demographic information in subsequent experiments.

In our original study, participants over 18 years old were solicited via social media (Twitter, Facebook, LinkedIn, and Instagram) and not compensated. The experiment took place online using a survey built in Qualtrics. Participants used their own audio hardware. Participants were told that they were taking part in a robot sound study and asked to complete an answer box with the text, "Please describe what the following sound is communicating to you." Answers were skippable if no suitable answer could be thought of. Each participant heard 20 randomly selected sounds of the 68 total sounds, in a randomized order. Participants could listen to each sound as many times as they wished before responding. We received 81 viable results. Due to the randomization, different numbers of participants answered the questions, and we received between 17 and 25 respondents for each sound. As noted above, we randomly selected five of the 68 sounds for the second and third experiments.

Participants were again recruited from the same social media for the online tests, five months later. Participants were not compensated for this study. The video experiment used HD video, paired with the original audio files dubbed overtop. A series of movement patterns (e.g., "move forward 30 cm and turn") was created, in randomized blocks. The sound was synchronized to start at the beginning of the movements. Playback order was randomized. We also had four of the same sounds paired with a different movement, to determine whether movement would impact the answers: The results showed no impact of movement on results, and therefore, the extra videos were removed from the subsequent analysis. Participants freely responded to the question "In the box provided below, please describe what the robot is communicating to you", a slight variation since this time they could obviously see that it was a robot. As with the first experiment, participants could repeat the sound as many times as they wished before responding. Some responses were left blank and, as a result, between 29 and 40 responses for each sound were obtained.

For the live test, participants were recruited from a Human Computer Interaction on-campus study recruitment Facebook page and through personal networks. Participants were brought to a plain, quiet, windowless room on the Carleton University campus. There was an approximately 2.5 × 2.0 m table in the center of the room on which the robot could move about. The same movement patterns and sound-action pairs as the video condition were used. Participants only observed the sound-action pair once before responding. Participants were asked, "What do you think the robot is communicating?" and were given a paper and pen to freely respond. Participants were compensated for their time with a $25 gift card. There were 24 participants who heard all sounds, but who left the response blank if they could not find an appropriate response. We received between 21 and 24 responses for each sound-action pair.

All text was pre-processed manually for noise removal, lemmatization, and stop-word removal. We used several methods to compare the data. First, all data was analysed using the sentiment analysis tools in SpeakAI. The tool provides a line-by-line analysis of sentiment, with a rating for positive, neutral and

Internal use

negative for both the overall text and for each entry. Then we have a response-by-response breakdown of the analysis which operates on a scale of -1 to +1. For comparison across tools, we used the total, overall scores for the sound. We also ran all responses through IBM Watson's sentiment analysis tool. Finally, we completed a manual analysis, coding each response as positive, negative, or neutral, counting each of these and arriving at a percent score based on the number of responses. These were coded by two human coders. Totals for each sound were then tabulated. Given that the results provided were in slightly different formats, we then normalized the responses for consistency, to end with a score from -1 to +1 (with neutral at 0) for all three analysis tools.

4 Results

We wanted to explore whether the sentiment changed by presentation mode (audio, video, live). Responses, as addressed above, were in free verbalization text format. Some examples of responses are for instance: "Talking bird", "charging up", "robotic meow", "sad. unsuccessful. could be searching for something but didn't find it", "lil homie came to tell me something went wrong! is it him that messed up, or me? not entirely sure. either way, feels like an oopsie!" and the researchers' favourite, "hello accomplice".

To investigate whether sentiment varies by condition (audio, video, live), we analyzed the sentiment scores across these conditions. The mean sentiment scores for each condition suggested small differences in sentiment scores (Audio: -0.275, Video: -0.258, Live: -0.226), with live seeing a slight boost in sentiment (Figures 1, 2, 3).

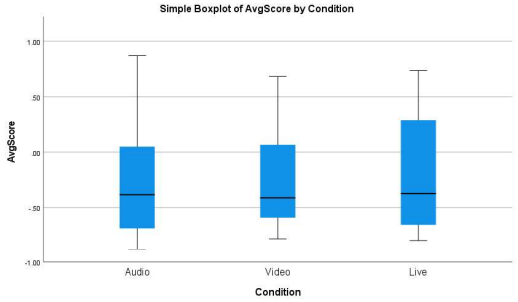

Figure 1. Sentiment Scores by Presentation Mode

In examining the effects of sound type and presentation mode on average sentiment scores, a two-way ANOVA was conducted. The results indicated that while there was a significant main effect of sound on sentiment scores, $F(4, 45) = 11.714$, $p < .001$), suggesting that different sounds significantly affected sentiment scores and accounted for 51% of the variance in the score. In contrast, there was no significant main effect of presentation mode on sentiment scores, $F(2, 45) = .097$, $p = .908$), indicating that the presentation mode did not significantly influence sentiment scores. Moreover, the interaction between sound and presentation mode was not significant ($F(8, 45) = 1.245$, $p = .296$), suggesting that the impact of sound on sentiment scores was consistent across different conditions. The model explained a significant proportion of variance in sentiment scores (R Squared = .559, Adjusted R Squared = .422).

Our prior study suggested that some sounds were more prone to the use of onomatopoeia, more prone to the use of emotion words ("angry" "sad", etc.), and/or more prone to the use of anthropomorphic/zoomorphic references (that is, the responses included insects, robots, aliens, humans, etc.). We wanted to compare these results across the three experiments by presentation mode. In other words, was there a correlation between sounds that give rise to onomatopoeia and those that also give rise to more emotional keywords used, or those that reference living creatures (what we referred to as anthropomorphic references)? Furthermore, are these responses impacted by presentation mode? Two researchers coded the data for onomatopoeia, emotional words, and anthropomorphic words. Keywords were counted per participant, and then divided by the number of overall participants for that response.

By type of sound, as can be seen in Figure 2, the "raw" type of sound had significantly more onomatopoeia (mean 18.52% of responses), and the musical type, keyboard "kbd" (mean 4.5%), had the least onomatopoeia.

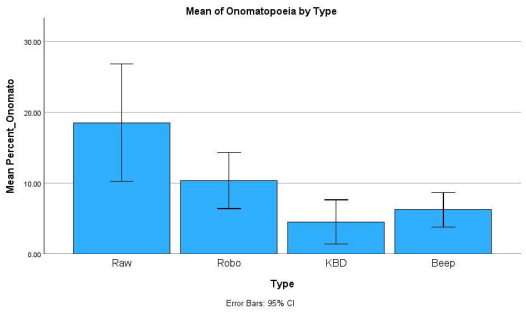

Figure 2. Mean Onomatopoeia by Type

Internal use

A one-way ANOVA on onomatopoeia by type confirmed a significant effect of sound type on the usage of onomatopoeic words ($F(3, 56) = 7.183$, $p < .001$). The effect size for this analysis (Eta-squared) was .278, indicating that approximately 27.8% of the variance in onomatopoeic word usage was associated with the sound type. This large effect size was further corroborated by Epsilon-squared (.239) and Omega-squared (.236) for fixed effects, which similarly suggest that between 23.6% and 23.9% of the variance in the onomatopoeic word usage can be explained by differences in sound types.

We then examined the emotional keywords used to describe the sounds (Figure 3). A one-way ANOVA was conducted to determine whether there were differences in the percentage of emotion word usage associated with different sound types. The analysis did not reveal a statistically significant effect of sound type on emotion word usage, $F(3, 56) = 1.841$, $p = .150$. Although not statistically significant, the effect size (Eta-squared) was .090, suggesting that sound type accounted for approximately 9% of the variance in emotion word usage among the conditions studied. The effect sizes indicated by Epsilon-squared (.041) and Omega-squared (.040) for fixed effects were also small, and the lower bounds of the confidence intervals were negative, indicating that the true effect size could be zero. The small Omega-squared random effect (.014) further suggests a minimal contribution of sound type to the variance in emotion word usage, with a large portion of variability likely attributable to individual differences or other unmeasured variables. It is interesting to note, however, that the musical samples (keyboard, "kbd") again elicited the lowest means (8.27%), roughly half those of the raw voice (17.5%).

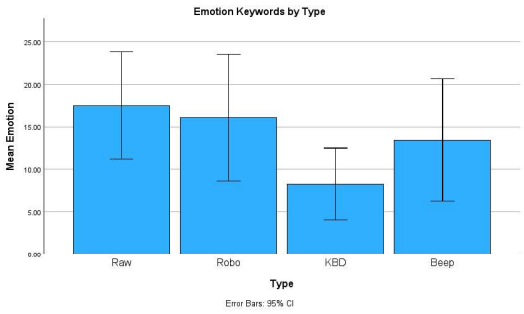

Figure 3. Mean Emotion Keywords by Type.

Anthropomorphic references—that is, references to living creatures (insects, robots, people, and so on) were also subjected to an ANOVA. The results indicated that there was no statistically significant difference in the percentage of anthropomorphic keywords among the sound types ($F(3,56)=1.376$, $p=.260$). The effect size, as measured by Eta-squared, was small ($\eta 2=.069$), suggesting that sound type explained only a small proportion of the variance in the percentage of anthropomorphic keywords. Overall, the results suggest that the sound type did not have a significant impact on the use of anthropomorphic keywords in the experiments (Figure 4).

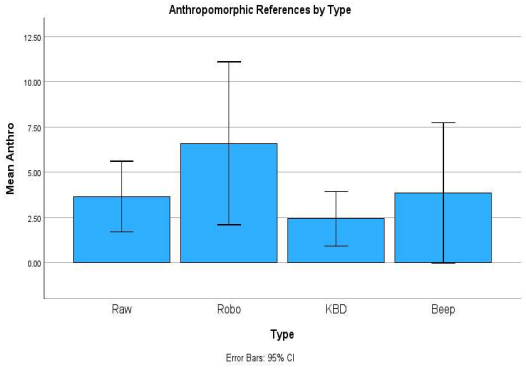

Figure 4. Mean of Anthropomorphic References by Sound Type.

Further analyzing the keywords by presentation mode, based on the means, it was clear that the audio-only condition elicited more onomatopoeia than other conditions, with the live condition significantly lagging (Figure 5). A one-way ANOVA was conducted to examine the influence of presentation mode on the percentage of onomatopoeia keywords used in the experiments. The results indicated that there was a statistically significant difference in the percentage of onomatopoeia keywords among the experimental conditions ($F(2,57) =4.569$, $p=.014$). These results suggest that the choice of presentation mode had a statistically significant impact on the use of onomatopoeia keywords in the experiments. A Tukey HSD post-hoc analysis revealed: no significant difference between audio and video (mean difference = 2.67, $p=.666$), and no significant difference between live and video (mean 6.43, $p=.103$). A significant difference was found between video and live (Mean Difference = 9.09641, SE = 3.09382, $p = .013$. The 95% confidence interval for the mean difference is from 1.6514 to 16.5414.) and between audio and live (Mean Difference = -9.09641, SE = 3.09382, $p = .013$). The 95% confidence interval for the mean difference is from -16.5414 to -1.6514.)

Internal use

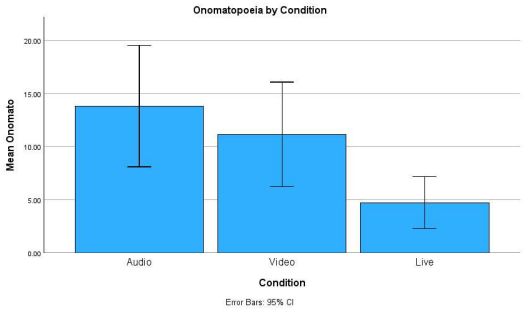

Figure 5. Means of Onomatopoeia by presentation mode.

We measured the emotion and anthropomorphic references the same way. A one-way ANOVA was conducted to examine the influence of presentation mode on the percentage of emotion keywords. The results indicated that there was no statistically significant difference in the percentage of emotion keywords among the experimental conditions, (F(2,57)=0.248 p=.781) (Figure 6).

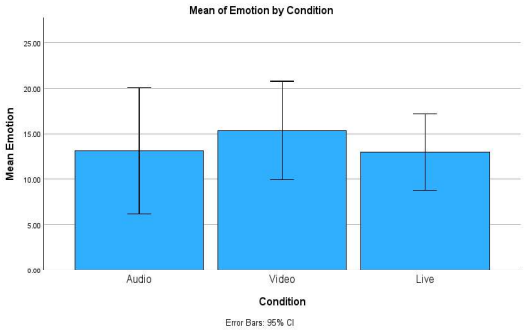

Figure 6. Emotion Keyword Means by Presentation Mode

A one-way ANOVA indicated that there was no statistically significant difference in the percentage of anthropomorphism keywords among the experimental conditions, (F(2,57)=2.689, p=.077). These results suggest that the choice of presentation mode did not have a statistically significant impact on the use of anthropomorphism keywords in the experiments, but it approached significance (Figure 7).

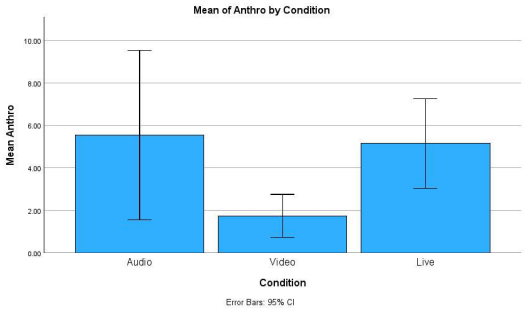

Figure 7. Anthropomorphic Keyword Means by Presentation Mode.

To summarize, the presentation mode under which the sounds are experienced affects how likely respondents are to use onomatopoeic words in their responses. However, the presentation mode does not significantly influence the likelihood of respondents using emotion keywords and shows only a marginal trend for influencing the use of anthropomorphic keywords. This suggests that the context or setting in which sounds are heard may influence the way people linguistically respond to them, particularly regarding the use of onomatopoeic language.

5 Discussion

We presented three experiments on the same set of twenty emote sounds, under three different conditions (audio only online, video online, and live in person). Respondents used a free verbalization to determine what they felt the sound was communicating to them. We then compared the responses across the type of sound manipulation (what we called the raw voice, robotic, keyboard, and beep versions), along with the presentation mode condition. There were no statistically significant differences in presentation mode when it came to overall sentiment analysis, suggesting that testing of a sound's communication/meaning could be conducted online without a degradation in the results.

The results here suggest that the inclusion of image to the interpretation of the sentiment of these particular sounds is negligible. Again, further research is needed to determine whether the same results would carry to other types of sounds. Moreover, would the same sounds with different types of images alter the results? We were surprised to find a lack of significant difference in the sentiment scores between different presentation methods. The results suggest that, at least with respect to testing for sonic meaning, it may not be necessary to conduct studies in person,

Internal use

and it likewise may not be necessary to conduct studies with image or video included, at least for some kinds of sounds. Nevertheless, the differences we did observe, such as increased onomatopoeia in audio-only testing, suggests that there may be differences in cognitive processing for audio-only tests.

There was a higher mean in sentiment scores for the live presentation mode, although the ANOVA showed a lack of difference. This could be due to a limited number of participants and therefore, a greater number of participants might reveal some statistical significance: However, whether having the presence of the experimenters, or the presence of the robot in person made a difference will have to be teased in future experiments. The particular robot and the movements chosen, may have had some influence on the results, and further studies could tease out these variables. The movement of our robot was limited, and the robot itself was physically small (~ 30cm). It would be interesting to test on a much larger robot, or one with a less toy-like nature. The slightly higher rating of sentiment in person may be due to the participant seeing the robot in action and obtaining a sense of its physical size: The video we used, shot against a plain wall, would not have allowed for a sense of scale. It would be interesting to see if results remained consistent with a different robot.

We anecdotally noticed more was written by the respondents in the live condition, which is interesting since these responses were written by hand. We averaged the words per respondent for each condition and found in Experiment 1, respondents wrote approximately 1.68 words per response. For Experiment 2, approximately 3.37 words per cell, and for Experiment 3, this increased to approximately 7.29 words per cell. There could be several reasons for this result: perhaps under the discomfort of having an experimenter in the room, participants felt greater pressure to respond, or since they were being compensated for their participation, participants felt compelled to write more. It could also be because participants preferred writing by hand rather than typing responses. Given that the video condition also doubled the words per response, the presence of the object, in video form and even more so in live form, could lead to more time spent considering the response. It could be, then, that a richer sensory experience prompted longer responses. Regardless, writing more in the words per response suggests that, generally, the live respondents answered differently, and that this could account for the small increase in observed positivity. Future research must tease out the reasons why participants wrote more, and should strive for more consistency across pay rates, writing method, and experimenter presence.

While the results are promising for those who wish to study the sentiment of sounds, there are limitations to this study. We tested a small number of sounds, all vocal emotes/robot sounds, and the results may not carry over to other types of sounds. Moreover, we tested on a small number of participants, with each test conducted on a maximum of 40 people. We plan on conducting further, much larger-scale studies of sounds and meaning with a greater variety of sounds. There is always a challenge of using quantitative methods on qualitative data [37]. While keyword counts and sentiment analysis can offer some insight into the data, and text analytics tools have come a long way in recent years, (particularly with the rise of generative AI), we have taken just one approach to qualitative data analysis. We may achieve different results with other approaches to free verbalization data (see, for instance [38]).

Perhaps the most significant finding was that there was notably more use of onomatopoeia by respondents when the test was audio only. The prevalence of onomatopoeia is interesting because evidence suggests onomatopoeia causes activation in different parts of the brain than other words [39][40]. This is perhaps unsurprising when we consider that sound alone may lead to increased mental imagery, since there is no visual imagery presented. We also know that mental imagery influences perceptual processing [41], and changes in mental imagery evoked by different modes may account for different responses. While there may be some advantages to using an audio-only presentation mode when it comes to attempting to tease out the communicated meaning of sounds, when the final sounds are part of a product design, for instance, including a video or live presentation may lead to minor differences in results. Further research is necessary on larger numbers of participants, with different types of sounds, to explore how the presentation method may influence perceptual results.

6 Acknowledgements

This work was funded by the Social Sciences and Humanities Research Council of Canada in the form of an Insight Development Grant to K. Collins, and the Natural Sciences and Engineering Research Council of Canda through a Vanier Scholarship to H. Johnston.

Internal use

7 References

[1] K. Collins and H. Johnston. "A Free Verbalization Method of Evaluating Sound Design: The Effectiveness of Artificially Intelligent Natural Language Processing Methods and Tools**."** *ACM AudioMostly*. Edinburgh, UK, August 30-Sept 1, 2023

[2] G. Ning, Q. Daggett, A. Perivolaris, B. Kapralos, A. Quevedo, KC Collins, K. Kanev, and A. Dubrowski. "Rethinking audio-haptic perceptual immersion from in-person to remote testing during COVID-19," In *Proceedings of the 14th International Conference on Interactive Mobile Communication Technologies and Learning* (IMCL 2021), November 4-5, 2021, Thessaloniki, Greece, pp. 1-9, 2021.

[3] Z. Zhang, S. Zhu, J. Mink, A. Xiong, L. Song, and G. Wang. "Beyond Bot Detection: Combating Fraudulent Online Survey Takers." In *Proceedings of the ACM Web Conference 2022*, April 25–29, Lyon, France, 2022.

[4] S. Zhao, C.A. Brown, L. L. Holt and F. Dick. "Robust and Efficient Online Auditory Psychophysics." *Trends in Hearing* Volume 26: pp. 1–24. 2022

[5] T.N. Reid, E.F. MacDonald, and P. Du. "Impact of product design representation on customer judgment." *J. Mech. Des*. Vol.135, pp. 091008, 2013.

[6] C. Kim, C. Lee, M.R. Lehto, and M.H. Yun. "Affective evaluation of user impressions using virtual product prototyping." *Hum. Factor Ergon. Man*. Vol 21, pp. 1–13, 2011.

[7] A. Voit, S. Mayer, V. Schwind, and N. Henze. "Online, VR, AR, Lab, and In-Situ: Comparison of Research Methods to Evaluate Smart Artifacts." In *Proceedings of the 2019 ACM CHI Conference on Human Factors in Computing Systems,* Glasgow, 4–9 May 2019.

[8] H. Honing and O. Ladinig. "The Potential of the Internet for Music Perception Research: A Comment on Lab-Based Versus Web-Based Studies." *Empirical Musicology Review* Vol. 3, No. 1, 2008.

[9] Z. E. Peng, S. Waz, E. Buss, Y. Shen, V. Richards, H. Bharadwaj, G. C. Stecker, J.A. Beim, A.K. Bosen, M.D. Braza, A.C. Diedesch, C.M. Dorey, A.R. Dykstra, F.J. Gallun, R.L. Goldsworthy, L.Gray, E.C. Hoover, A. Ihlefeld, T. Koelewijn, J.G. Kopun, J. Mesik, D.E. Shub, J. Venezia. "FORUM: Remote testing for psychological and physiological acoustics." *J Acoust Soc Am* vol. 151, pp.3116–3128, 2022.

[10] T. Eerola, J. Armitage, N. Lavan & S. Knight. "Online Data Collection in Auditory Perception and Cognition Research: Recruitment, Testing, Data Quality and Ethical Considerations," *Auditory Perception & Cognition*, vol. 4 no. 3-4, pp. 251-280, 2021.

[11] B.A. Mok, V. Viswanathan, A. Borjigin, R. Singh, H. Kafi, Hari M. Bharadwaj. "Web-based psychoacoustics: Hearing screening, infrastructure, and validation." *Behavior Research Methods,* vol. 56, pp. 1433-1448, 2023.

[12] N. Otto, S. Amman, C. Eaton, S. Lake. "Guidelines for Jury Evaluations of Automotive Sounds." *Sound and Vibration.* April 2001

[13] S. Zhao, C.A. Brown, L.L. Holt, and F. Dick. "Robust and Efficient Online Auditory Psychophysics," *Trends in Hearing,* vol. 26, pp. 1–24, 2022.

[14] Y. Wycisk, R. Kopiez, J. Bergner, K. Sander, S. Preihs, J. Peissig, and F. Platz. "The Headphone and Loudspeaker Test - Part I: Suggestions for controlling characteristics of playback devices in internet experiments." *Behav Res Methods,* vol 55, no. 3, pp.1094-1107, 2023.

[15] W. Yang and J. Kang. "Acoustic comfort evaluation in urban open public spaces." *Applied Acoustics,* vol. 66 no. 2, pp. 211–229, 2005.

[16] Y. Wycisk, R. Kopiez, J. Bergner, K. Sander, S. Preihs, J. Peissig, F. Platz. "The Headphone and Loudspeaker Test – Part I: Suggestions for controlling characteristics of playback devices in internet experiments" *Behavior Research Methods,* vol 55, pp.1094–1107, 2023.

[17] H. Honing, and O. Ladinig. "The potential of the Internet for music perception research: A comment on lab-based versus Web-based studies." *Empirical Musicology Review*," vol 3 no. 1, pp. 4–7, 2008.

[18] A.E. Milne, R. Bianco, K.C. Poole, S. Zhao, A.J. Oxenham, A. J. Billig and M. Chait. "An online headphone screening test based on dichotic pitch." *Behavior Research Methods,* vol. 53, no. 4. pp. 1551-1562, 2021.

[19] K.J.P. Woods, M.H. Siegel, J. Traer, and J. H. McDermott. "Headphone screening to

Internal use

facilitate web-based auditory experiments." *Atten Percept Psychophys*" vol.79, pp.2064–2072, 2017.

[20] A. Alsius, M. Paré, and K.G. Munhall. "Forty years after hearing lips and seeing voices: The McGurk effect revisited*." Multisensory Research*, vol. 31 no. 1-2, pp. 111-144, 2018.

[21] C. S. Choe, R. B. Welch, R. M. Gilford, and J. F. Juola. "The 'Ventriloquist Effect': Visual Dominance or Response Bias?" *Perception & Psychophysics* vol 18, no. 1, pp. 55–60, 1975.

[22] L. Shams, Y. Kamitani, and S. Shimojo. "Illusions. What You See Is What You Hear." *Nature*" vol. 408 no. 6814, pp. 788, 2000.

[23] N. Bonneel, C. Suied, I. Viaud-Delmon, G. Drettakis. "Bimodal perception of audio-visual material properties for virtual environments." *ACM Trans. Appl. Percept*." vol 7, pp. 1–16, 2010.

[24] G. Mastoropoulou, K. Debattista, A. Chalmers, and T. Troscianco. "The Influence of Sound Effects on the Perceived Smoothness of Rendered Animations." In *Proceedings of the 2nd Symposium on Applied Perception in Graphics and Visualization*, 2005.

[25] Y.C. Chen, and C. Spence. "The crossmodal facilitation of visual object representations by sound: Evidence from the backward masking paradigm." *Journal of Experimental Psychology: Human Perception and Performance,*" vol. 37 no. 6, pp. 1784–1802, 2011.

[26] T. Chandrasekera, and N.D. Souza. "Virtual environments with soundscapes: a study on immersion and effects of spatial abilities." *Environ. Plan. B Plan. Design*" vol. 42, pp. 1003–1019, 2015.

[27] K. Abe, K. Ozawa, Y. Suzuki, T. Sone. "Comparison of the Effects of Verbal Versus Visual Information About Sound Sources on the Perception of Environmental Sounds." *Acta Acustica United with Acustica,* vol. 92, pp. 51 – 60, 2006.

[28] J.R. Williams, Y. A. Markov, N. A. Tiurina, and V.S. Störmer. "What you see is what you hear: Sounds alter the contents of visual perception." *Psychological Science*, vol. 32, no. 12, 2022.

[29] T. J. Cox. "The effect of visual stimuli on the horribleness of awful sounds." *Applied Acoustics* vol. 69, pp. 691–703, 2008.

[30] M. Filimowicz, ed. *Doing Research in Sound Design*. New York: Focal Press/Routledge, 2022.

[31] C. Sester, C. Dacremont, O. Deroy and D. Valentin. "Investigating consumers' representations of beers through a free association task: A comparison between packaging and blind conditions." *Food Quality and Preference*" vol. 28, pp. 475–483, 2013.

[32] P. Tagg. *Music's Meanings*. New York, New York: Mass Media Music Scholars' Press, 2013.

[33] M. Olko, D. Dembeck, Y-H. Wu, A. F. Genovese, and A. Roginska. "Identification of perceived sound quality attributes of 360° audiovisual recordings in VR using a Free Verbalization Method." Presented at the *143rd Audio Engineering Society Convention* October 18–21, New York, NY, 2017.

[34] E. Samoylenko. "Systematic Analysis of Verbalizations Produced in Comparing Musical Timbres." *International Journal of Psychology,* vol. 31, no. 6, pp. 255–278, 1996.

[35] W. Medhat, A. Hassan, and H. Korashy. "Sentiment analysis algorithms and applications: A survey." *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093-1113, 2014.

[36] M. Usama, B. Ahmad, E. Song, M.S. Hossain, M. Alrashoud, and G. Muhammad. "Attention-based sentiment analysis using convolutional and recurrent neural network." *Future Generation Computer Systems*, vol. 113, pp. 571-578, 2020.

[37] S. Almalki. "Integrating Quantitative and Qualitative Data in Mixed Methods Research--Challenges and Benefits." *Journal of Education and Learning*, vol. 5 no. 3, pp. 288-296, 2016.

[38] S. I. Nikolenko, S. Koltcov and O. Koltsova. "Topic modelling for qualitative studies." *Journal of Information Science,* vol. 43 no. 1, pp. 88-102, 2017.

[39] T. Hashimoto, N. Usui, M. Taira, I. Nose, T. Haji and S. Kojima. "The neural mechanism associated with the processing of onomatopoeic sounds". *NeuroImage* vol. 31, pp. 1762–1770, 2006.

[40] T. Hashimoto, N. Usui, M. Taira, I. Nose, T. Haji and S. Kojima. "The neural mechanism associated with the processing of onomatopoeic sounds," *NeuroImage* vol. 31 pp. 1762–1770, 2006.

[41] D. Röders, A. Klepp, A. Schnitzler, K. Biermann-Ruben, and V. Niccolai. "Induced and Evoked Brain Activation Related to the

Internal use

Processing of Onomatopoetic Verbs." *Brain Science*, vol. 12, pp. 481, 2022.

[42] C.C. Berger, and H.H. Ehrsson. "Mental imagery changes multisensory perception." *Current Biology,* vol. 23, no. 14, pp. 1367-1372, 2013.