Audio Engineering Society
Convention Express Paper 203

Presented at the 156th Convention
2024 June 15-17, Madrid, Spain

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections, or consideration by the Review Board. The AES takes no responsibility for the contents. 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.

Humanizing AI-Generated Music: Do Listeners Hear a Difference?

KC Collins[1] and Adel Manji[1]

[1] *School of Information Technology, Carleton University, 1125 Colonel By Drive, Ottawa, ON K1S 5B6 Canada*

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

ABSTRACT

The rise of large language model AI generative tools has resulted in multiple AI music generators being made available to the public. But how good are they at tricking us into believing that they may be human-created? And can we manipulate the music produced to sound more human using production techniques like adding small human sound effects or introducing convolution reverbs? We ran four internet-based listening studies to try to answer these questions. AI-generated music was able to fool approximately half of the listeners into believing that it could be human generated. However, our interventions did not increase the likelihood of the music being labelled human-generated

1 Introduction

One of the purposes of music in media is to emotionally connect with the audience, and express the feelings and emotions of the on-screen characters. If we acknowledge that, currently at least, machines do not have emotions, then there is a barrier between human and AI-created music. Research has shown that we can recognize and feel the emotion conveyed by a performer when we listen to music, and that these changes in emotion result in biophysical changes (blood pressure, heart rate, galvanic skin response, etc.)[1]. One hypothesis into why this response occurs is that we understand human-made sounds in terms of our own experience of making similar sounds and movements. In particular, we have a relationship to music through the mental mimicry of the sentiment of the performers, a process that has long been referred to in music literature as the ideomotor principle [2], kinematic empathy [3], bodily hearing [4], anthropomorphic projection [5], kinaesthetic sympathy [6], vicarious performance [7] or corporeal signification [8].

The role of the body in understanding emotional content has been shown to be critical to empathic engagement. In performing, we attempt to imbue sound with an expressiveness through our instrument. Our mental re-creation of the music as a listener causes a neuronal and motor-sensory response that mimics the performer, and thus we can interpret the emotional inflections through our mental re-creation of the action. We therefore give meaning to sound in terms of emulated actions, or corporeal articulations. In other words, we mentally imitate the expressiveness of the action behind the sound, based on our prior experiences of producing similar sounds [9].

Our body knows the difference between live and recorded music. One study has shown how head movements of audiences differed in response to live music versus recorded music, with live music having greater engagement [10]. Another study focused on paediatric critical care patients, examining the effects of live versus recorded music interventions on physiological parameters like heart rate, blood pressure, and respiratory rate. The study found that

Internal use

live music, particularly delivered by trained music therapists, might have a more pronounced effect in reducing physiological markers of stress and anxiety compared to recorded music[11]. Another study of the reactions of infants to live and recorded music showed differences in the engagement levels, with them more engaged and overall more sustained attention and physiological synchrony with live instruments [12]. These studies suggest that even when we do not know what we are listening to is live or recorded, our body understands, on some level, and responds differently to human-played performance. In other words, this relationship between what we hear coming out of the speakers and the body of the performer is broken when it comes to virtual instruments [13].

With the explosion of machine learning/generative artificial intelligence (AI) models to create music and sound, there are many questions that arise about the reception and perception of this music. AI-generated music is proliferating across streaming services and making news headlines as artists fight for removal of copy-cat songs, and streaming services struggle to keep up with number of songs being produced. As of May 2023, Mashable reported that Boomy, launched in 2021, had generated over 14.5 million songs [14]. This number will only increase with the ease of use of not just Boomy, but other AI music generators on the web, such as SoundDraw, Beethoven, and AIVA. A musical Turing test has long been the goal in AI-focused music research [15]. With the explosion in machine learning approaches, there is no longer a question that AI can produce music that can fool listeners into thinking it was written by humans. The humanness of performance and production aspects, however, is still a question that remains to be answered.

Recent surveys have found participants reported they liked music less when they thought it was composed by AI, and in general there are negative perceptions towards AI-generated music [16][17]. In 2021, Moura and Maw for instance conducted an online survey which indicated "negative perception, low purchase intention for AI music and a negative credibility perception."[18] They point to the social nature of music, and the innate emotional intentionality behind music as key reasons for the negative perception. Writing on AI music, Huang and Sturm described that some viewed the AI creation of music as "insensitive, blasphemous, and disrespectful" [19]. Other more recent surveys have found there is currently a lack of consensus on the issue of AI taking on creative pursuits [20].

Notions of "authenticity" in music studies have long been debated, and the subject frequently arises in discussing AI-generated content. However, the technology has moved—and continues to move—quickly, and questions about authenticity are perhaps moot in the face of the current generation of AI tools. There are certainly styles of music that do not require a sense of authenticity—indeed, that are not even possible for humans to play ("Black MIDI" for instance, is a genre that intentionally produces music that could only be created and played back artificially[21]). The widespread use of Autotune also indicates that going beyond human capabilities is an approved aesthetic in itself. Moreover, the popularity of known holographic performers like Hatsune Miku, Ashnikko and so on indicates many people don't care if music is created or even performed by human beings. Nevertheless, many sample libraries are often designed with built-in human performance limitations, where notes are only playable within the correct corresponding human range [22], suggesting that their main use is to *imitate* human performance, not extend it.

For many music genres—particularly of the sort that are likely to rely on AI, such as music for screen media (video games, film and television), there remains a strong preference for live orchestral recording [23], in part due to the perception that virtual instruments are less realistic and authentic. While once digital music production relied on synthesized approximations of instruments, now they can use high quality acoustic instrument sampling (recording). These are sometimes referred to as "deep-sampled" instruments. When it comes to recent studies of sampled versus live (acoustic/electric) instruments, evidence suggests that even music students sometimes had difficulty discriminating the sounds [24]. Another study, from 2016, showed 72.5% of overall listeners could tell the difference between sampled and live orchestral music playing: Listeners with musical expertise scored slightly higher than others, but more than half of all people could discern the difference [25].

There are many advantages to the use of virtual instruments. Virtual instruments offer (someone's version of) perfection: Changes can be made quickly and cheaply, and as noted above, the instruments can go beyond human abilities to offer additional musical possibilities. Nevertheless, even though many people

Internal use

cannot distinguish between virtual and live orchestras, it is still commonplace to incorporate acoustic instruments where budgets allow. There is a perception that there is something "missing" from an entirely sampled mix. It is commonplace where budgets allow to incorporate at least one or more acoustic instruments overtop of a virtual instrument mix to achieve a sense of liveness and humanness [26]. There are many tutorials online about how to blend a real instrument in with samples in order to create the right mix and the illusion of humans playing. [27] Another approach is to add convolution reverbs to increase the sense of physical space in the mix [28].

One of the "problems" with human players, aside from their cost, is their imperfection. However, it is this imperfection (at least in part) that brings the sense of humanness to the mix. Sample libraries are designed to perfection, but in this perfection, they miss the human aspects of playing. There are "Humanizer" plug-ins in most sequencers to stochastically de-quantize the beat timings, accents and dynamics, so that it sounds more human. However, there are other nuances to human playing that have not been captured by most sample libraries, including the small articulations and expressions musicians may make, human sounds behind playing (such as breathing) and the sound of movement around/with the instrument and any accompanying objects (such as turning the page of the sheet music). Explained by one viola player who posted samples on the free sound library Freesound.org, "There's this 'grabby' sound that a well-rosined bow makes just is it is first being drawn across the string. Listen carefully to any of the pros and you'll hear it. In your laptop sequences, if used subtly, right at the point where the first note of a phrase is initiated, this sound can give the string part a marked sense of realism"[29]

While sample libraries have come a long way in recent decades, when it comes to virtual instruments, these small details are still often lost. Each instrument has nuanced "non-musical" sounds that are a part of human playing, from the clunk of a piano foot pedal being released, the string squeak from fingers sliding on a close-miked guitar, the subtle creak of the stool they sit on, and their breath as they play. When a human plays, a careful listen reveals these sounds, such as in for instance, Alexander Malter's performance of Arvo Pärt's "Für Alina", in which we can hear his body shift on the stool, his foot on the pedals, and his breathing in several places. These aspects of the performance help to define it, and add depth and feeling that likely would not be present using entirely synthesized sounds.

Game composer Troels Folmann described at the Game Developers Conference in 2006 that he introduced this "random chaos" into the virtual instruments he was using for the soundtrack to *Tomb Raider: Legend* (Eidos 2006). In the mix, he incorporated sounds of page turns, chair creaks, hall noise, human breaths/coughs, and instrument noises[30]. The soundtrack won Music of the Year award at the 2007 Game Audio Network Guild Awards, and Best Original Score at the 2006 BAFTAs. Were some of the accolades a response to the imperceptible inclusion of sound effects perhaps? With the rise of AI music creators, the human is even further removed from the process of music creation and production. But, given the addition of some of these types of sounds, will humans be fooled into thinking that they are listening to human-created music? Will a human intervention in the production of AI-generated music alter the perception? These questions are behind this current study. Our purpose was to see if the addition of sound effects and/or convolution reverbs to an AI-generated piece of music would increase the illusion of humanness as a baseline to then further explore the issue. We conducted four studies which are presented below, before we conclude with some general thoughts.

2 General Survey Methods

Surveys were constructed in Qualtrics and participants were recruited from Prolific, a paid research survey service similar to, but of reputably higher quality responses than, Amazon's Mechanical Turk (for a comparison of the judgement of quality of response between MTurk and Prolific, see [31]). Participants took about five minutes to complete the surveys, which were each conducted approximately two months apart, and were paid between £0.75 and £1.50 (GBP) depending on the length of the survey. Different participants responded to each survey. Participants were first asked some basic demographic information. They were also asked about the listening device used for the study in the form of multiple choice. They were then given a calibration sample, normalized to -3dB and instructed to set their volume to a "comfortable level". Then, the block of study questions were presented. The AI-generated music was created in AIVA, a service which allows you to generate music and separately download the stems,

Internal use

MIDI, MP3 or wave.[1] Files were all created using the "Epic Orchestra" source, with "GP Ensemble" instrumentation. Files were all approximately 30 seconds long. All files used in the survey were wave format at 44.1kHz sample rate.

3 Study One: Synthesizer Sounds and Addition of Sound Effects

For the first study, we used three pairs of tracks. Two AI-generated orchestral songs were created; one 60BPM in F Minor and one at 120 BPM in D♯ Minor. One track was human-created but using virtual instruments, known as "Sunlit Rain" from one of the co-author's own music pieces as performing artist Redshft, and was 70 BPM and was in C Major. None of these tracks were performed by humans—that is, they all used virtual instruments.

One of the tracks in each pair had additional sound effects mixed into it. These effects were all normalized to -24dB, and consisted of breathing, coughing, and chair creaks. These effects were downloaded from Freesound.org.[2] This amplitude was selected because the sounds could easily be heard on our studio monitors (a pair of Adam A7Xs) without being obvious in the mix. Each of these sound effects were about 10 to 15 seconds long (that is, containing multiple coughs, multiple creaks, and multiple breaths were heard), and they were overlapped and edited in places, then randomly placed on the timeline.

All tracks were presented with and without the additional sounds, in block random order, for a total of six sounds in three pairs. Participants were asked to indicate if either track A or B, both tracks, or neither tracks, in each pair were *played* by humans.

221 responses were collected in total. 95 participants reported they identified as women, 6 as other and the rest (119) were men. 101 were from the European Union & UK, 26 from North America, and 6 from Australia and Asia. 193 participants reported no known hearing deficits, 24 said they had known deficits, and 3 said they may have deficits. 120 participants used headphones, 51 used a built-in laptop speaker, 31 used external loudspeaker, 10 used the built-in desktop speaker and 6 used a mobile phone speaker. There was no correlation between hearing ability and response, between equipment and response, between gender and response or between country and response.

There were 89 participants in the 18-24 age group, 79 in the 25-34 age group, 34 in the 35-44 group, 12 in the 45-54 group, 4 in the 55-64 group, and 1 aged 65+ (Figure One). There *was* a correlation between age group and response, and this was somewhat consistent across the three questions, whereby the older age groups tended to be more correct in their responses. In particular, 35-44 year olds were the most skeptical, most likely to indicate "neither" of the tracks were human created.

3.1 Question 1

For Question 1, sample B contained the additional effects. The piece, by Redshft, was written by a human, rather than AI generated, although played on synthesizers/sampler; the correct answer to if the piece was played by humans was therefore technically "neither", but given that there was human intervention in this track, "human" could also be considered correct (Figure 1).

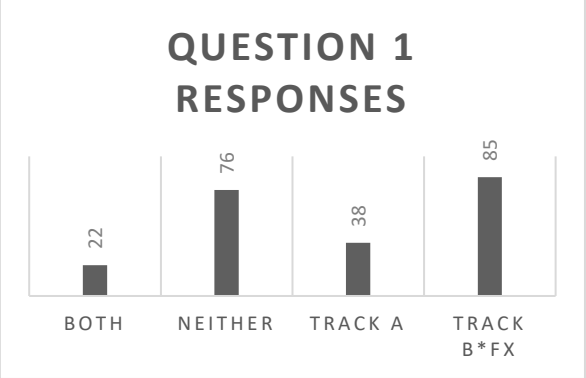

Figure 1. Responses to first pair of tracks

Across all age groups, sample B was the most popular choice, with the exception of 35-44 age group, which had a 61.76% response for the correct answer (Figure 2).

[1] https://www.aiva.ai/ Files were created between January 2024 (Study One) and March 2024 (Study Four). As far as we are aware, there were no updates to the models during that time.

[2] Chair creak, https://freesound.org/people/unkleceeg/sounds/276473/ light coughing, https://freesound.org/people/14GPanskaMuzatko_Matej/sounds/419781/ and breathing https://freesound.org/people/pekena_larva/sounds/667286/

Internal use

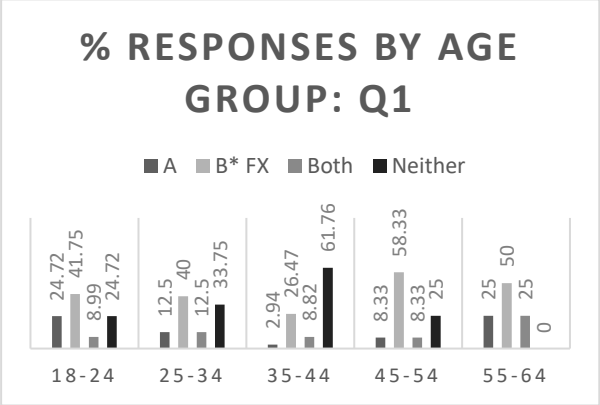

Figure 2. Percent of responses to first pair by age group

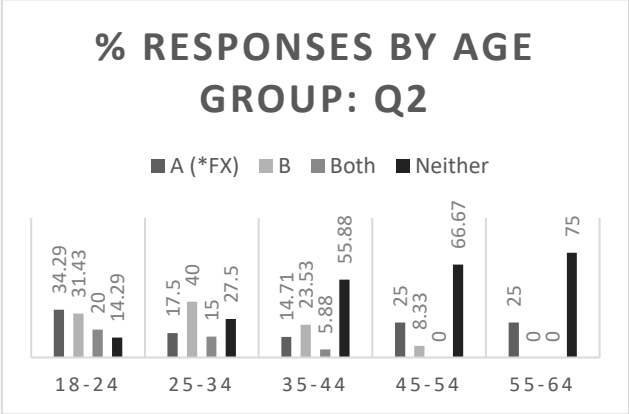

Figure 4. Percent of responses to second pair by age group

3.2 Question 2

For Question 2, Sample A contained additional effects, although the correct answer was again "neither". The piece was written by AI and performed on synthesizers. Here, the participants did not select Sample A, our effects sample, as the human version (Figure 3). Again, responses tended to be more accurate with older age groups.

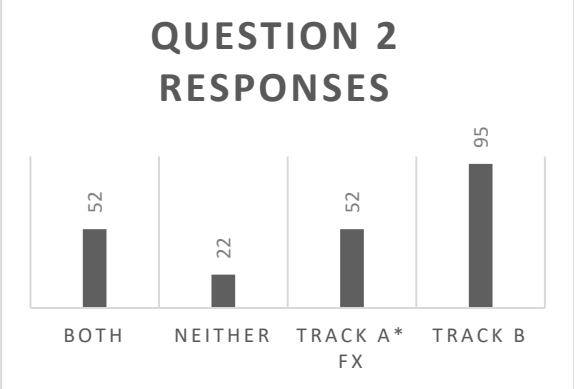

Figure 3. Responses to second pair of tracks

3.3 Question 3

For Question 3, Sample A contained additional effects, although the correct answer was, again, "neither". The piece was written by AI and performed on synthesizers. Here, the participants again did not select Sample A as the human version (Figure 5). Again, older age groups tended to be more accurate in responses (Figure 6).

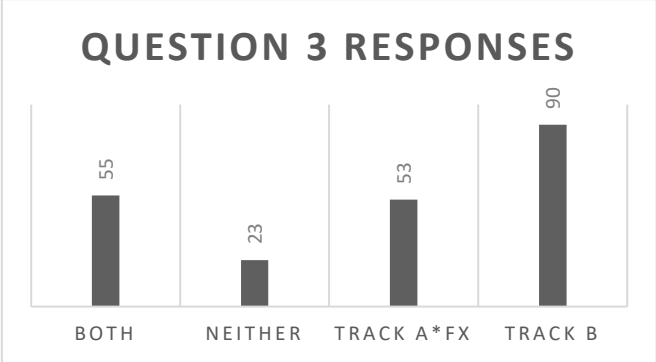

Figure 5. Responses to third pair of tracks

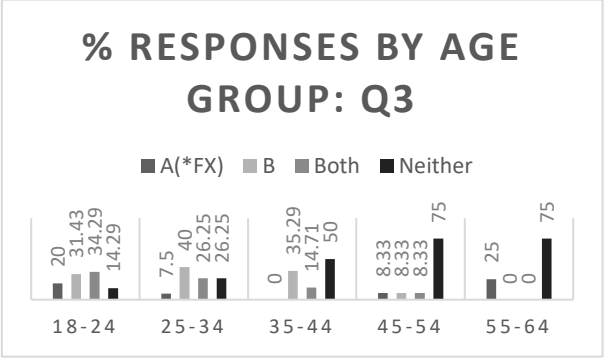

Figure 6. Percent of responses to second pair by age group

The results of this study showed that the addition of sound effects did *not* increase the likelihood that our respondents would hear the samples as human-created. However, across the samples the correct answer "neither" was significantly more likely to be selected with the human-composed file.

Internal use

4 Study Two: Live Orchestra and Synthesized Music, with and Without Effects

For the second study, we wished to reduce confusion of human-performance, since a human-composed piece using virtual instruments could be argued to be human-played. For this study, we used comparisons of real orchestral performance created by humans, and synthesized music created by AI, with and without effects. This time, sound effects were mixed to -30dB, which over headphones and our loudspeakers were barely discernible. We also changed some of the breath sound out to be less repetitive, and added chair creaks at less regular intervals. The human-generated track was "Revenge of the Underworld (No Choir)", by Iouri Sazonov, purchased from the ShockwaveSound.com library. The song featured a full orchestra, recorded live.[3]

For this study, we employed a Check All That Apply approach, asking the participants not to compare samples to each other, but to listen to one sample and then answer if they thought the sample was: written by humans, performed by humans, written by AI, or performed by AI. There are some challenges with this approach—namely, we are still leaving it up to the respondent to determine what is meant by "performed by humans": for instance, if a piece is using samples that were originally performed by humans, does that make the piece performed by humans? However, we left this up to the respondents to determine.

We received 120 responses to this study. 78 identified as men, 39 as women, and 3 other. 112 reported no hearing impairments, and 8 reported known hearing deficiency. 100 were from the European Union and twenty were from North America. 81 participants used headphones, 15 used external speakers, 12 used laptop built-in speakers, 4 desktop speakers, and 7 mobile phone speakers. Respondents using headphones had a slightly higher percentage of correct answers compared to other types of equipment, but there was no significant correlation between the technology used and the results. Once again, there was no correlation between gender, hearing impairments, or location of participants and the responses.

There were 50 participants in the 18-24 age group, 43 in the 25-34 age group, 19 in the 35-44 age group, 6 in the 45-54 age group, and 1 each in the 55-64 and 65+ age groups. There was a small correlation between ages and response (Pearson correlation coefficient of 0.187, p-value 0.041). As the age group increases, the correct answers tended to increase as well (Figure 7).

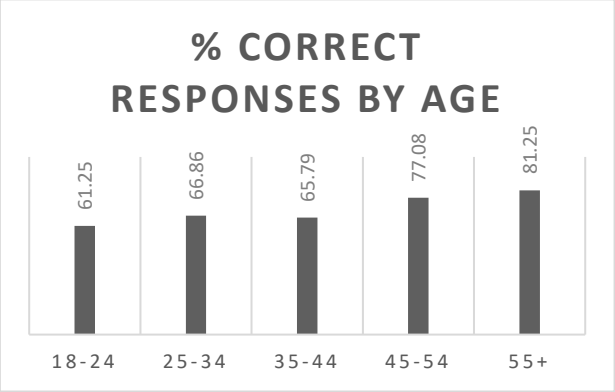

Figure 7. Percent of correct guesses by age group

A chi-square test indicated that there was no statistically significant association between the presence of effects and participants' responses ($\chi 2$ = 4.000, df = 3, p = .261, phi coefficient ~0.1825). This suggests, once again, that the inclusion of effects in the AI music samples did not significantly affect participants' ability to identify whether the music was AI or human written or performed. However, as can be clearly seen in the graphs, while most people were able to determine the human-written and human-performed samples, when it came to the AI written and performed samples, results were far more mixed, with roughly only half of participants able to determine if the songs were written and/or performed by humans or AI (Figures 8-11).

[3] https://www.shockwave-sound.com/search/Revenge%20of%20the%20Underworld We used the FNo Choir mix, 30 seconds.

Internal use

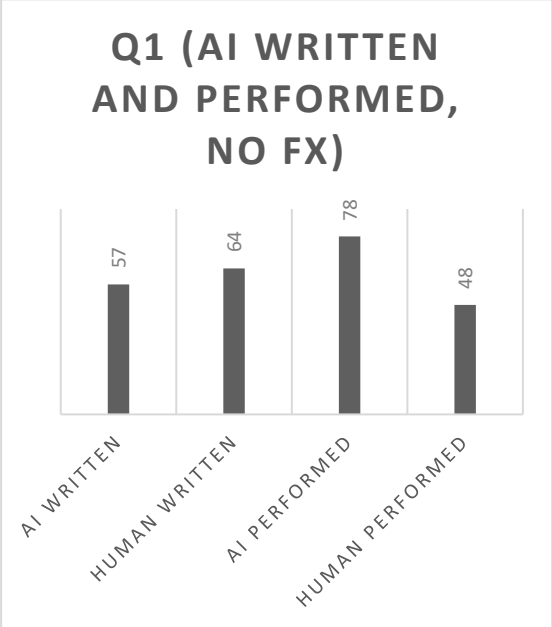

Figure 8. Q1 responses

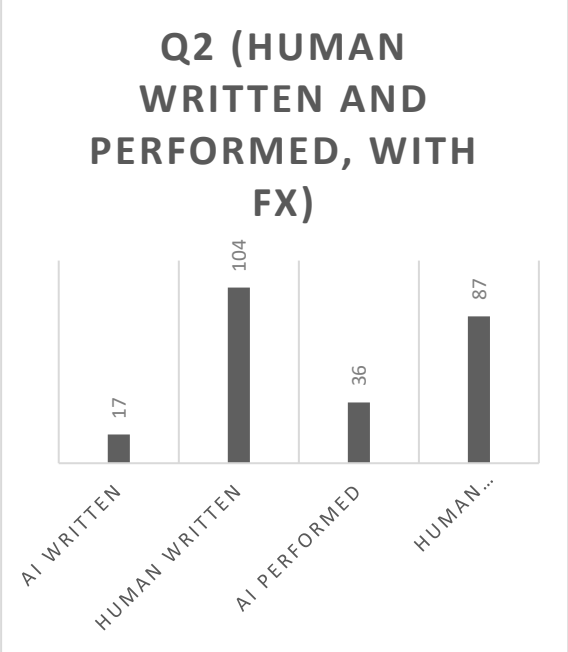

Figure 9. Q2 responses

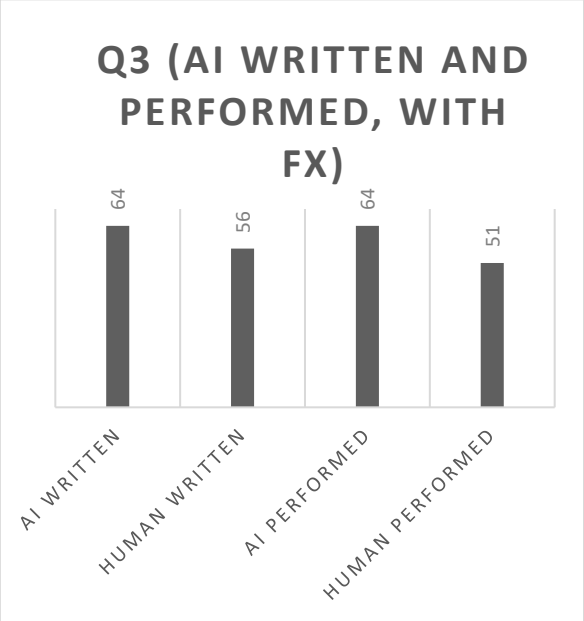

Figure 10. Q3 responses

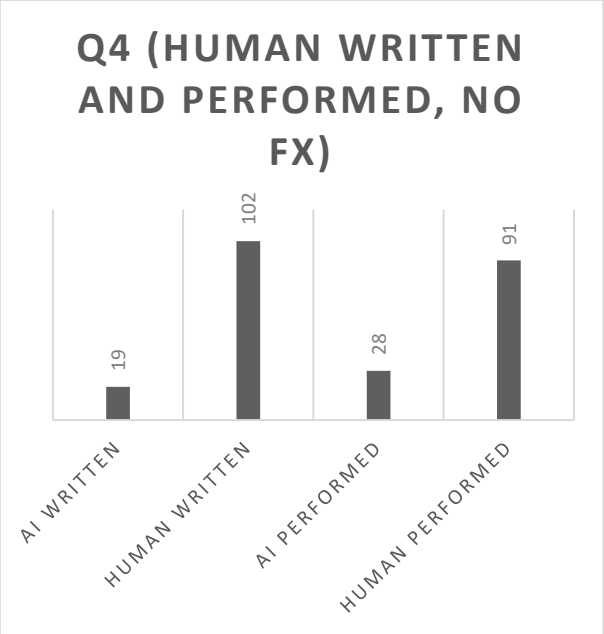

Figure 11: Q4 responses

5 Study Three: Live Orchestra compared to AI/Synthesis with Production Manipulations

Having seen no statistically significant difference with the addition of sound effects to the samples in the first two studies, we decided to switch focus to using different production methods. For this study, we manipulated both the AI and human created files.

Internal use

Two AI files used the same raw samples as above, without the addition of effects. This time the files were given convolution reverbs, the Central Hall at University of York (UK) impulse responses.[4] The human-generated file, with real orchestra, the same sample outlined above in Study Two, was treated to some EQ to bring down the brightness a little. We gradually rolled off the volume as the frequencies increased above about 1.5KHz, giving the song a "duller" sound that was closer to the mix of the AI generated tracks.

We received 109 responses. There were 49 women respondents and 60 men. 98 reported no hearing impairment, and 11 reported some impairment. 87 were from Europe or the UK, 14 from North America, 4 from Australia and Asia, and 4 from Africa. 88 used headphones, 1 an external speaker, 14 a laptop speaker, and 6 their phone speaker. We ran bivariate and Chi-square tests but found no statistically significant correlations between gender, country, technology used or hearing impairment. There were 54 in the 18-24 age group, 39 in the 25-34 group, 9 in the 35-44 group, 5 in the 45-54 group, and 2 in the 55-64 age group. This time, there was also no significant correlation between age and accuracy of response.

As can be seen in Figure 13, overall, it appears that participants were generally able to distinguish between music written by humans and AI *when the song was written by humans* (Q2). However, when songs were written by AI, there was less likelihood of guessing correctly. It's interesting, then, that people are noticing subtle nuances in the human generated music, even when we have reduced the clarity of the piece through EQ to more closely resemble the AI-generated music. The AI pieces were able to fool, in Q1, nearly half of the respondents. Our addition of real impulse responses to the AI-generated pieces may have had a small impact, however, we did not test for the inclusion or exclusion of impulse responses in this study, and even with the changes made, it seems most listeners can still hear the difference when they are played a piece generated by humans.

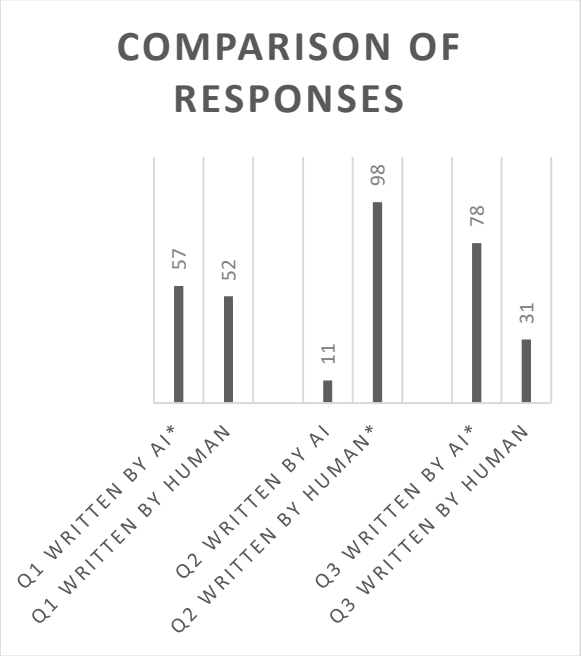

Figure 13. Comparison of raw numbers of responses to AI-written (Q1, Q3 and human written tracks with production manipulation)

6 Study Four: AI-Generated Music with and without Impulse Responses

For the final study, we created new AI-generated music, to test with and without impulse response. All four pieces were created using AIVA and we used 44.1KHz wave format files. All files were rendered in the same "Epic Orchestra" style, with two in a major key (E, A♭), and two in minor key (G♯, D♯). Again, they were all approximately 30 seconds long. 2 of 4 pieces played had impulse responses added: the same York impulse response as described in Study 3. Playback of the selections were randomized.

We received 109 responses. 76 were from the EU & UK, 15 were from North America, 2 were from the Middle East and 16 were from Africa. 101 reported no known hearing impairment and 8 reported they may or definitely did have some hearing loss. 44 were women, 62 were men, and three identified as other. 40 were aged 18-24, 48 were 25-34, 10 were 35-44, 8 were aged 45-54, and 3 were 55-64. The majority (83) used headphones, 14 used laptop speaker, 4 used external speakers, 5 used desktop speaker and 3 used

[4] We used the position centre row ir_row_1l_sl_centre1L. The impulse responses can be downloaded free of charge from the link and were recorded by Alexander Vilkaitis, Ilias Antonopoulos, Joska De Langen and Xaun Liu in 2015. https://www.openair.hosted.york.ac.uk/?page_id=435

Internal use

phone speaker. Once again, we ran Pearson bivariate correlations but failed to find any significant correlation between gender, equipment, location or hearing ability and the responses. There was also no correlation this time with age.

Responses for both the impulse response-containing files and non-impulse response containing files were roughly evenly divided between believing them to be human or AI created. We ran a two-sample proportion z-test (Track 1 p-value is 0.9856. Track 2 p-value is 0.1695. Track 3, p-value 0.1835. Track 4, p-value 0.0105). There were no significant differences in the proportion of correct answers for files with or without the addition of the convolution reverb (Figure 14). In other words, the addition of convolution reverb made no discernible difference to the likelihood to believe that the songs were human created.

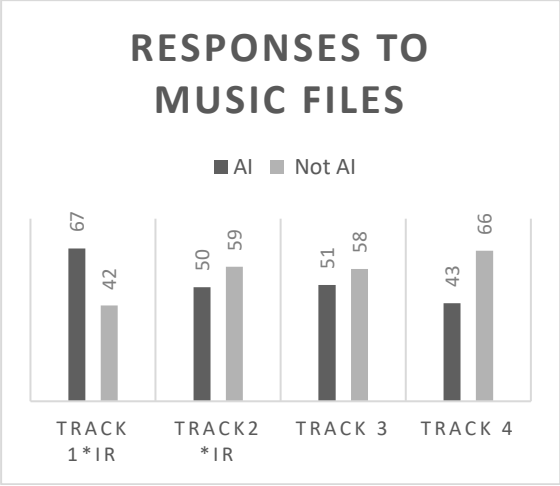

Figure 14. Guesses as to if music was AI or not AI generated, with (tracks 1 and 2) and without impulse responses. All tracks were AI generated.

7 Discussion

The results of the four studies presented here show that, at least in the songs chosen and the effects we used, there was no difference in the likelihood of hearing an AI-generated song as human with the addition of "human" sound effects or by attempting to manipulate the tracks with impulse responses or equalization to sound more similar to each other in production value.

The studies show that a surprising number of people are unable to differentiate human versus AI-composed or performed music, suggesting that the quality of AI generated music is improving. However, when human-generated and performed music was compared, results overwhelmingly suggest that these can be detected as human. When presented with AI music, roughly half of respondents can tell a difference. Our attempted interventions did not improve the likelihood that people would guess that it was created by humans. This could mean that AI-generated music is still being perceived as less authentic due to unnatural transitions, repetitive patterns or inconsistencies in tone and expression rather than production values. Or, simply, our production interventions were not enough to make a difference.

We consistently found no statistically significant differences between responses by gender, audio technology, country of origin, or even hearing ability limitations. However, in two of the four studies age did play a factor: the youngest respondents were least likely to be able to differentiate between AI and human generation, which is interesting. The reasons for this disparity could be many, and are interesting in that one might assume that as one ages they are more likely to be fooled by attempts to humanize sound. It could be that the older participants were more experienced listening to music, were more likely to take their time answering the questions and listen carefully, or had higher quality audio equipment, even if they were technically using the same method (e.g. headphones can vary widely of course, and we did not factor in type or model of headphone worn). It could also be that younger people growing up on lower quality MP3 and streamed audio are less able to hear details in productions that older people are able to hear. We initially thought perhaps younger people were hearing something in the very high frequencies of the files that could not be heard by older age groups, but spectrogram analyses revealed nothing of interest.

The disparity in ages is interesting in light of other recent research that suggests that when it comes to AI imagery, younger participants were more likely to answer correctly [32]. While we all commonly have experience reading faces, perhaps we have vastly different listening experiences. It would be interesting to run a future study where we gauge the listening and music performance experience of listeners. Recent work has suggested that those with musical experience may be better at discerning AI-generated music: A University of York study of fifty respondents with some musical knowledge found that AI music was rated as inferior to human-created music[33].

Internal use

We see this paper as one piece of research that should be repeated as AI continues to improve. While many listeners could not differentiate between AI and human composed or performed songs, one of our limitations was that we used 30 seconds of songs, rather than entire, lengthy songs, and there may perhaps be more distinction with longer songs where structure may come into play.

We used default settings in the AI generated tracks, used orchestral music, and left the mixing up to AIVA, apart from the addition of effects and convolution reverbs. Our next study will create simplified non-orchestral songs using the more detailed ways AIVA allows you to compose with it (selecting specific chord progressions, removing instruments, separating stems and mixing ourselves). We plan to then compare the AIVA tracks with our own recorded performance of these AI-written songs. Future studies could also explore a combination of effects and treatments on the AI files to see if a variety of production methods used together might increase the likelihood of people guessing that these files were human generated.

8 Conclusions

We presented here four studies that attempted to explore how various interventions into AI-generated music might influence the likelihood of that music passing a musical Turing test. Despite multiple attempts at humanizing AI-generated music through augmenting the music with human effects and convolution reverbs, we failed to improve the likelihood of these songs as being heard as human generated. We were surprised by this result, as we felt that such subtle differences in production values would fool a greater number of people. An interesting finding was that the technology used to listen to the sound files, which ranged from headphones to mobile phone speakers, did not influence the results at all. This lack of statistical significance of technology used suggests that it is not production values that influence people to hear songs as AI or human generated, but likely something in the structure, harmony or melody, and/or instrumentation of the music being created. Further research should explore these potential factors to attempt to determine why the pieces of music that were human composed were more easily identifiable as human-made, whereas those created by AI fooled people only roughly half the time. It would be interesting to also ask participants to rate their confidence in response.

References

[1] D. Rocchesso, R. Bresin, N. Streitz, A. Kameas, and I. Mavrommati (Eds.). “Emerging Sounds for Disappearing Computers.” *The Disappearing Computer*, LNCS4500, pp. 233–254 (2007).

[2] C. Stöcker and J. Hoffmann. “The Ideomotor Principle and Motor Sequence Acquisition: Tone Effects Facilitate Movement Chunking.” *Psychological Research* vol. 68, no 2, pp 125–137 (2004).

[3] N. Todd and P. McAngus. “The Kinematics of Musical Expression.” *Journal of the Acoustical Society of America* vol 97, no. 33, pp. 1940–1949 (1995).

[4] A. Mead. “Bodily Hearing: Physiological Metaphors and Musical Understanding”. *Journal of Music Therapy* vol. 43, no. 1, pp. 1–19 (2003).

[5] J. Chadabe. Electronic Music and Life. *Organised Sound* vol. 9, no. (2), pp. 3–6 (2004).

[6] J. Cage. *Silence: Lectures and Writings. Middletown*, CT: Wesleyan University Press (1988).

[7] E. Cone. *Musical Form and Musical Performance*. New York: Norton (1968).

[8] M. Leman. *Embodied Music Cognition and Mediation Technology*. Cambridge: MIT Press (2008).

[9] A. Cox. “The Mimetic Hypothesis and Embodied Musical Meaning.” *Music Sci*, vol. 5 no. 2, pp. 195–212 (2001).

[10] D. Swarbrick, D. Bosnyak, S.R. Livingstone, J. Bansal, S. Marsh-Rollo, M.H. Woolhouse and L.J. Trainor. “How Live Music Moves Us: Head Movement Differences in Audiences to Live Versus Recorded Music.” *Frontiers in Psychology* vol. 9. pp. 2682 (2019).

[11] H.I. Bush, A.B. LaGasse, E.H. Collier, M.A. Gettis, K. Walson. “Effect of Live Versus Recorded Music on Children Receiving

Internal use

Mechanical Ventilation and Sedation." *American Journal of Critical Care*, vol. 30, no. 5. pp. 343-349 (2021).

[12] H. E. Kragness, M. J. Eitel, F. A. Anantharajan, B. Gaudette-Leblanc, B. Berezowska, and L.K. Cirelli. "An itsy bitsy audience: Live performance facilitates infants' attention and heart rate synchronization." *Psychology of Aesthetics, Creativity, and the Arts.* Advance online publication. (2023). https://doi.org/10.1037/aca0000597

[13] J. Croft. "Theses on liveness." *Organized Sound*, vol. 12, no.1, pp. 59–66 (2007).

[14] E. Cavender. "Spotify removes thousands of AI-generated songs". *Mashable*. https://mashable.com/article/spotify-ai-crackdown., (May 9, 2023).

[15] C. Ariza, "The Interrogator as Critic: The Turing Test and the Evaluation of Generative Music Systems," *Computer Music Journal* vol. 33, no. 2, pp. 48–70 (2009).

[16] D.B. Shank, C. Stefanik, C. Stuhlsatz, K. Kacirek, and A.M. Belfi. AI composer bias: Listeners like music less when they think it was composed by an AI. *Journal of Experimental Psychology: Applied*, vol. 29, no. 3, pp. 676–692 (2023).

[17] S. Knotts and N. Collins. **"**A survey on the uptake of Music AI Software", *NIME'20, July 21-25, 2020,* Birmingham, United Kingdom

[18] F.T. Moura and C. Maw. "Artificial intelligence became Beethoven: how do listeners and music professionals perceive artificially composed music?" *Journal of Consumer Marketing,* vol. 38 no. 2, pp. 137–146 (2021).

[19] R. Huang, and B. Sturm, B. "Reframing Aura: Authenticity in the application of AI to Irish traditional music" In *Proceedings of AI Music Creativity Conference,* online, July 2021 (2021).

[20] K. Déguernel and B. Sturm. "Bias in Favour or Against Computational Creativity: A Survey and Reflection on the Importance of Socio-cultural Context in its Evaluation." *Proc. International Conference on Computational Creativity* (2023)

[21] J. Mullen, S. Gilbert, and J. Simpson. "Insider Interviews: Underground Music in the UK in the 2020s." *French Journal of British Studies* vol. 26 no. XXVI-3 (2021).

[22] E. Klein. "Feigning Humanity: Virtual Instruments, Simulation and Performativity." *IASPM Journal* vol. 6, no. 2, pp. 22-48 (2016)

[23] G. Wilson. *An Autoethnography on Film Composition Practice: Integrating Virtual Instruments and the Live Performance Realisation within the DAW*. Doctoral dissertation, University of Huddersfield (2018).

[24] J. V. Ruiz, P.K. Cooper. and J. N. Muhammed. "Can they hear a difference? Professional digital composition and the ability of music students to discriminate deep-sampled vs. acoustic instrumental performance recordings." *Journal of Popular Music Education*, vol. 4 no. 1, pp. 81–99, (2020).

[25] R. Kopiez, A. Wolf, F. Platz, J. Mons. "Replacing the Orchestra? – The Discernibility of Sample Library and Live Orchestra Sounds." *PLoS ONE"* vol. 11, no. 7, pp. e0158324. (2016).

[26] J. Doyle. "Mixing and Recording a Small Orchestral Ensemble to Create a Large Orchestral Sound". In *Innovation in Music: Performance, Production, Technology and Business*. R. Hepwroth-Sawyer, J. Hodgson, J. Paterson and R. Toulson, eds. Routledge. pp. 226-247 (2019).

[27] "Combining Virtual and Acoustic Instruments" *Puremix*, https://v2.puremix.com/blog/combining-virtual-and-acoustic-instruments.html (2017)

[28] S. Andersson. "Can a Dry Recording Treated with a Convolution Reverb in a Surround Sound Speaker System be Perceived as Maintaining the Same Degree of Naturalness as Real Recorded Reverberation?" Thesis, Luleå University of Technology, (2023).

[29] B. Burbank, "grabby bow original",

Internal use

Freesound, https://freesound.org/people/Bruce%20Burbank/sounds/220917/ (2014)

[30] T. B. Folmann. "Tomb Raider Legend: Scoring a Next-Generation Soundtrack." *Game Developers' Conference,* https://slideplayer.com/slide/4124878/ (2006).

[31] D.A. Albert and D. Smilek. "Comparing attentional disengagement between Prolific and MTurk samples". *Nature Scientific Reports* vol. 13 pp. 20574 (2023)

[32] A. Pocol, L. Istead, S. Siu, S. Mokhtari and S. Kodeiri. "Seeing is No Longer Believing: A Survey on the State of Deepfakes, AI-Generated Humans, and Other Nonveridical Media." *Advances in Computer Graphics. CGI 2023. Lecture Notes in Computer Science*, vol 14496 (2024).

[33] Z. Yin, F. Reuben, S. Stepney et al. "Deep learning's shallow gains: a comparative evaluation of algorithms for automatic music generation." *Machine Learning* 112, 1785–1822 (2023).