Toward a musical Turing test for automatic music performance

Antonio Rodà[1], Emery Schubert[2], Giovanni De Poli[1], and Sergio Canazza[1] *

[1] Dept. of Information Engineering, University of Padova, Italy
[2] School of the Arts and Media, University of New South Wales, Australia
roda@dei.unipd.it

Abstract. This paper reports a "musical Turing test" conducted at a live concert of algorithm-generated performances, where one group of participants were invited to rank the most human-like performance while knowing that one of the performances was by a human, and another group of participant were asked to do the same, but without knowing that there was a human performer on the program. The program consisted of five pieces from the classical/romantic period, played on a Disklavier. High quality music-expression algorithms were used to generate the algorithmic renditions. Regardless of the group, musical experience and a number of other factors, the subjects were unable to identify the human performer out of the five performances. The group that did not know there was a human performer had a wider range of votes compared to the group that did know. Furthermore, subjects were less confident of their answers when they knew that they were comparing human and computer-generated performances . On the contrary, if subjects believed they were only comparing computer-generated performances the task may have been less demanding. Findings suggest that computer algorithms are able to substitute for human performance, but the role of the physical presence of the performer (who was absent in this study) could be an area for further investigation.

Keywords: algorithm-generated performance, expressive music performance, turing test

1 Introduction

Computer software has been able to send messages to acoustic pianos since the 1980s allowing the development of automated, live performances of piano music. With recent improvements in algorithmic generation of standard (mostly piano) classical/romantic repertoire, a new research question has been emerging: Will there ever be a time when a listener cannot distinguish between an algorithm performing a piece (for example via a Disklavier and the Bösendorfer SE reproducing piano [1]) versus an expert human performer? The ability of an algorithm or robot to be human-like has been a matter of fascination since the possibilities of automation have and robotics arose (in the case of music, see [2]).

* This research was partly supported by an Australian Research Council Fellowship (FT120100053).

Since 1950, when Alan Turing published his first paper on the issue [3], the "thinking capacities" of a computer have also been measured in terms of how well the computer can do in the "Imitation game", i.e. a party game in which a man (player A) and a woman (player B) go into separate rooms and aim to convince the guests (player C) that they are the other. Although the so-called Turing test has been largely used in the last decades, controversy has arisen over which of the alternative formulations of the test Turing intended [4]. Turing never makes clear whether the interrogator in his tests is aware that one of the participants is a computer: he states only that player A is to be replaced with a machine, not that player C is to be made aware of this replacement. As many researchers have highlighted, (e.g.,[5]), this makes a big difference to the implementation of the test and experimental studies have revealed significant differences between the responses of participants who knew versus those who did not know that a computer was involved [6]. In contrast to this result, Bishop et al. [7] showed that knowing/not knowing did not make a significant difference in the evaluation of judges asked to distinguish human from machine both in a real human-machine comparison and with control pairs of two humans and two machines . Another variation of the test is described as the subject matter expert Turing test, where a machine's response cannot be distinguished from an expert in a given field [8].

The Turing test in its original formulation is based on the natural language: it assumes that humans have minds and that natural language is sufficient to represent their mind. In other words, the test uses human-like conversation as the sole determinate of thinking [9]. In the last decades, this limitation was overcome and the Turing test was applied in a large set of fields, including the evaluation of creativity in computer systems [10]. In several contexts, e.g. computers beating world chess champions [11], it is now accepted that machines can do some things better than humans, at least in those fields, such as chess, where there are explicit rules that provide a framework within which the protagonists seek a positive outcome, such as winning. However, in terms of artistic endeavour the situation is far from ended or even waning. We suspect that there are many views about what an algorithm can do, and people's beliefs about the ability of the algorithm to do it. Music provides an interesting case and more than one computational system for making music has been formerly evaluated using the Turing test (e.g, see [12] for an application to free jazz improvvisation). For at least twenty years researchers have been trying to identify the 'rules' of music performance to see if there are alternatives to the straightjacket MIDI style playback of programmed pitches with literal interonset timings that are associated with 1980s technology. Without covering too much history (but see, e.g., [13]; [14]; [2]), one of the landmarks for progress in the area of music performance has been the RENCON series, an international music performance rendering competition. Here a number of researchers present algorithms for playing established compositions, and listeners are asked to rate the playback from which a winner is determined (for further details, see [15]; [16]; [17]). However, few studies that we have cited have reported explicit examination of the question: Have the algorithms used for automated music playback reached a point where the listener can no longer distinguish between human and robot performer? Some recent, but limited evidence suggests that the typical listener is no longer able to distinguish between an algorithm and a human performance

[18,19]. In this paper we aim to continue addressing this question by asking listeners at a live performance to judge the source of Disklavier renditions.

2 Method

2.1 Participants

The experiment was carried out during a public concert held July 18, 2014 at the concert hall of the Music Conservatory in Venice. The concert was organized as an evening event at the 13th International Conference on Intelligent Autonomous System and was open both to participants of the conference and to a general audience. In total, it was attended by more than one hundred spectators.

The audience was divided into two groups, depending on the seat position: people seated on the left half (i.e. stage left, house right) of the hall were assigned to group A, people at the right half to group B. Each spectator was asked to fill a survey on a paper form. Group A and B receives a different version of the survey, differing for the incipit. Participants were not aware about the subdivision into two groups and, consequently, about the existence of two versions of the survey. 51 participants completed the survey. Table 1 report some information about the overall background characteristics of the two groups, in terms of genre, age, musical background, musical instrument played, and years of musical studies. The two groups did not significantly differ by gender (χ^2 = 2.225, df = 1, p-value = .136), musical background (χ^2 = 3.719, df = 5, p-value = .591), music instrument (χ^2 = .765, df = 2, p-value = .682), and years of musical studies (t = .291, df = 32.696, p-value = .773). A significant difference was found for the age (t = -2.488, df = 40.572, p-value < .05): on average, the group B (mean=47.9 years) was older than group A (mean=37.0). Group A also had proportionally more males than did group B.

Table 1. Summary of participant data by group membership.

Group	Gender		Age		Musical background					Musical instrument			Years of study	
	F	M	Mean	Std	a	b	c	d	e	Piano	Other	None	Mean	SD
A	2	18	36.9	13.7	2	5	4	6	1	8	6	5	7.2	11.1
B	10	21	47.9	16.1	5	7	9	5	0	9	11	10	6.3	9.9
Tot	12	39	43.7	16.0	7	12	13	11	1	17	17	15	6.6	10.2

2.2 Stimuli

The developers of the best classified systems at the last two Rencon contests (held on July 2011 and August 2013, see http://renconmusic.org for more information) were asked to provide the MIDI recording of a music performance of a short (five minutes at most) piano piece obtained with their algorithms. The music piece was freely chosen by each

developer from the Classical/Romantic period of Western music repertoire. Moreover, the same request was made to a professional pianist, with a long experience as a concert performer. The human performance was recorded as a MIDI file on a Yamaha Disklavier. The final list of stimuli were recorded as MIDI files: (1) Clementi's Piano Sonata Op. 12 No. 1 (1st mov.), played by CaRo 2.0 [20]; (2) Mozart's Piano Sonata KV 545 (2nd mov.), played by Director Musices [21]; (3) Mozart's Piano Sonata no. 4 KV 282 (3rd mov.), played by a human pianist; (4) Beethoven's Piano Sonata Op. 2 No. 1 (3rd mov.), played by Mixer Basis [22]; (5) Kuhlau's Allegro Burlesco Op. 88 No. 3, played by Virtual Philarmony [23].

2.3 Procedure

At the opening of the concert a compere announced that all the piano pieces were played by automatic systems. Before the start of the concert, the subjects of each group received a survey that differed in the introductory text only. Specifically, the survey of group A started with "Please listen carefully to the first five pieces of the concert. One of these pieces is actually the recording of a human performance and not an automatic performance as announced". Group B received a survey with "Please listen carefully to the first five pieces of the concert. These pieces are played by means of different algorithms for automatic music performance". After each performance, during the clapping, a slide with the name of the system/algorithm which played the piece and a photo of its designer/developer's was projected to the public. For the human performance, a fake name and photo was presented.

Participants were asked to select the performance out of the five that sounded most human-like, specifically answering the question: "Which is the most human-like performance?" with a checkbox provided for each performance listed in the program.

After this, the participants were asked to rate their confidence in making the choice: "How confident are you with the previous answer?" with responses made along a scale of 0 (labelled 'no confidence') to 3 ('very confident').

3 Results

Tab. 2 shows the results of the responses to question 1 ("Which is the most human-like performance?"). The Human performance ranked second least human-like for group A (3 votes) and equal lowest for group B (5 votes).

Differences can be observed between the judgments of the two groups (e.g., the CaRo performance was judged as the most human like by five subjects in Group B and none in Group A). Moreover, the scores for Group B appear to be more uniformly distributed (in a range from 5 to 8) than Group A (range from 0 to 8). However, these differences are not statistically significant following a Pearson's Chi-squared test (χ = 4.55, df = 4, p-value = .337). A possible explanation of this result is that the number of subjects is not big enough to render significant these differences and the effect size (Cramer's V = .296, i.e. medium) partially supports this interpretation.

No significant difference can also be found in within-group scores: the judgments do not differ significantly by chance both for Group B (χ = 1.125, df = 4, p-value = .89)

Table 2. Participants' responses to question 1 by group.

System	Observed		Expected		Tot
	A	B	A	B	
CaRo	0	5	1.92	3.08	5
DM	4	7	4.23	6.77	11
Human	3	5	3.08	4.92	8
BM	8	7	5.77	9.23	15
VP	5	8	5	8	13

Table 3. Participants' responses to question 1 by musician (M) versus non-musician (NM) subjects. The total numbers of answers is different from Table 2 because one subject did not report his musical experience.

System	Observed		Expected		Tot
	M	NM	M	NM	
CaRo	1	4	1.86	3.14	5
DM	4	7	4.10	6.90	11
Human	3	5	2.98	5.02	8
BM	6	9	5.59	9.41	15
VP	5	7	4.47	7.53	12

and Group A (χ = 8.5, df = 4, p-value = .075), even if in the latter case the p-value is close to the significance threshold.

Quite surprisingly, almost no difference can be found between musicians and non-musicians (χ = 0.789, df = 4, p-value = .94; Cramer's V = .124): Table 3 shows that observed and expected values are very similar. And even more unexpected is that none of the 17 piano players, some with many years of instrumental practice (mean = 9.3 years, SD = 10.5 years), selected the human performance as the most human-like (see Table 4). Even in this case the differences are not statistically significant (χ = 6.362, df = 4, p-value = .174), but the effect size is medium (Cramer's V = 0.36).

Regarding question 2 ("How confident are you with the previous answer?"), the level of confidence of Group A (mean = 1.46) is less than Group B (mean = 1.93) and this difference is statistically significant (t = -2.304, df = 47.5, p-value < .05). This result shows that subjects are less confident of their answers when they know that they are comparing human and computer-generated performances and they have to find *the* human one. On the contrary, if they thought they were comparing only computer-generated performances, the task may have been less demanding. No significant difference in the level of confidence was found for Musician versus Non-Musician subjects (t = 1.088, df = 32.4, p-value = .285) and Piano versus Non-Piano players (t = .54, df = 35.4, p-value = .59), confirming that the musical expertise seems not to play a role in this context. Finally, no significant difference was found in relation to gender (t = .97, df = 16.9, p-value = .344) and age (t = -1.081, df = 42.2, p-value = .289).

Table 4. Participants' responses to question 1 by piano players (P) versus non-piano players (NP). The total numbers of answers is different from Table 2 because some subjects did not complete the item indicating which musical instrument they played.

System	Observed		Expected		Tot
	P	NP	P	NP	
CaRo	3	2	3.27	1.73	5
DM	8	3	7.18	3.82	11
Human	7	0	4.57	2.43	7
BM	7	8	9.80	5.20	15
VP	7	4	7.18	3.82	11

4 Discussion and conclusions

This experiment showed that state-of-the-art algorithms for the automatic performance of piano music are sufficiently advanced from the perspective that those performances are evaluated as being human-like to the same extent, statistically, as an actual human performance. Then, we examined whether the subjects' judgements were influenced by knowing or not knowing that a human performer was included in the program, even though the actual human performance was not revealed. We also examined the influence, if any, of the listeners' musical expertise, because presumably a highly experienced musician, and pianist in particular, may be better able to identify a human performance should it be possible to do so. Results show that the subjects who were aware (Group A) that a human performance was present responded slightly differently to those who were led to believe that none of the performances were played by a human (Group B), even though these differences are not statistically significant. Importantly, the level of confidence of the two groups is significantly different, showing that knowing or not knowing changes the subjects' perception of the task. The preconceived knowledge of the participant appears to have critical effect on how the musical experience is interpreted [24,25].

With regard to the musical expertise, this factor seems not to play a role in the subjects' ability to distinguish between human and computer-generated performances. In concert with this finding, none of the piano players participating in the experiment rated the human performance as the most human like.

This study has some limitations. First, we operated in the context of a live concert. The naturalistic setting of a live concert, with public in attendance created an ecologically plausible environment, but it meant that we had less experimental control. Other setups (e.g. a listening test by means of an online music player) could give different results, especially if the subject could re-listen and compare the performances more times, which is not possible in a live context. Furthermore, having different pieces with different levels of familiarity, complexity and so forth may have influenced results as much as did the algorithms themselves. Using the same piece repeatedly, however, may lead to fatigue effects [26] and were not considered appropriate for the naturalistic, live concert setting we were seeking. Another issue is more subtle: In our experiment we asked the subjects to evaluate a human performance (or more correctly, the playback by means of a real grand piano of a human performance recorded earlier) without a human pianist

on the stage. But to what extent does a performance without the physical presence of the human performer influence the perception of the performance? In other words, to what extent are listeners influenced by the visual spectacle of seeing an acoustic grand piano on the stage that was playing alone? Despite our attempts to generate a naturalistic setting, there may well be a rather dis-embodied aspect to such performances that is contrary to the more typical, traditional concert environment (for more discussion on the role of the visual perception of the performer in a musical context see e.g. [27]). And, so the presence/absence of the human player may, too, be an interesting variable to investigate in trying to understand psychological aspects of judging a music performance [28,29]. Finally, it is possible that the absence of statistically significant effects is due to the limited number of listeners, although the effect sizes did support the no-difference conclusions. Anyway, future research is addressing this problem involving a larger group of subjects [30].

Despite these limitations, we believe that this study provides modest evidence suggesting that expressive algorithms are able to 'pass the Turing test'. Just as with Turing's scenario described in the introduction, we set up a situation where one performer was a human, but the listener did not know 'which one'. While for another group, none of the performers were believed to be human, acting as a kind of control group. The poor ability of all participants, regardless of condition, to identify the real human player suggests that the complex, sensitive area of whether a machine can outdo a human in music performance must now be faced as a serious possibility.

References

1. Goebl, W., Bresin, R.: Measurement and reproduction accuracy of computer-controlled grand pianos. The Journal of the Acoustical Society of America **114** (2003) 2273–2283
2. Kapur, A.: A history of robotic musical instruments. In: Proceedings of the International Computer Music Conference. (2005) 21–28
3. Turing, A.M.: Computing machinery and intelligence. Mind **59** (1950) pp. 433–460
4. Moor, J.: The Turing test: the elusive standard of artificial intelligence. Volume 30. Springer Science & Business Media (2003)
5. Saygin, A.P., Cicekli, I., Akman, V.: Turing test: 50 years later. In: The Turing Test. Springer (2003) 23–78
6. Saygin, A.P., Cicekli, I.: Pragmatics in human-computer conversations. Journal of Pragmatics **34** (2002) 227–258
7. Bishop, M., Shah, H., Warwick, K.: Testing Turing's five minutes, parallel-paired imitation game. Kybernetes **39** (2010) 449–465
8. Feigenbaum, E.A.: Some challenges and grand challenges for computational intelligence. Journal of the ACM (JACM) **50** (2003) 32–40
9. Ariza, C.: The interrogator as critic: The turing test and the evaluation of generative music systems. Computer Music Journal **33** (2009) 48–70
10. Pease, A., Colton, S.: On impact and evaluation in computational creativity: A discussion of the turing test and an alternative proposal. In: Proceedings of the AISB symposium on AI and Philosophy. (2011)
11. Kasparov, G.: The chess master and the computer. The New York Review of Books **57** (2010) 16–19
12. Pachet, F.: The future of content is in ourselves. Computers in Entertainment (CIE) **6** (2008) 31

13. Kirke, A., Miranda, E.R.: An overview of computer systems for expressive music performance. In: Guide to computing for expressive music performance. Springer (2013) 1–47
14. Canazza, S., De Poli, G., Rodà, A., Vidolin, A.: Expressiveness in music performance: analysis, models, mapping, encoding. In Steyn, J., ed.: Structuring Music through Markup Language: Designs and Architectures. IGI Global (2012) 156–186
15. Hiraga, R., Bresin, R., Hirata, K., Katayose, H.: Rencon 2004: Turing test for musical expression. In: Proceedings of the 2004 conference on New interfaces for musical expression. (2004) 120–123
16. Hiraga, R., Bresin, R., Katayose, H.: Rencon 2005. In: Proceeding of the 20th annual conference of the Japanese Society for Artificial Intelligence (1D2–1). (2006)
17. Katayose, H., Hashida, M., De Poli, G., Hirata, K.: On evaluating systems for generating expressive music performance: The Rencon experience. Journal of New Music Research **41** (2012) 299–310
18. De Poli, G., Canazza, S., Rodà, A., Schubert, E.: The role of individual difference in judging expressiveness of computer-assisted music performances by experts. ACM Trans. Appl. Percept. **11** (2014) 22:1–22:20
19. Canazza, S., De Poli, G., Rodà, A.: How do people assess computer generated expressive music performances? In: 10th Sound and Music Computing Conference, Stockholm, Sweden (2013) 353–359
20. Canazza, S., De Poli, G., Rodà, A.: Caro 2.0. an interactive system for expressive music rendering. Advances in Human-Computer Interaction **2015** (2015) 13 pages
21. Friberg, A., Colombo, V., Frydén, L., Sundberg, J.: Generating musical performances with director musices. Computer Music Journal **24** (2000) 23–29
22. Grachten, M., Widmer, G.: Linear basis models for prediction and analysis of musical expression. Journal of New Music Research **41** (2012) 311–322
23. Baba, T., Hashida, M., Katayose, H.: "VirtualPhilharmony": A conducting system with heuristics of conducting an orchestra. In: Proc. of the Conf. on New Interfaces for Musical Expression (NIME 2010). (2010) 263–270
24. Rao, A.R., Monroe, K.B.: The moderating effect of prior knowledge on cue utilization in product evaluations. Journal of Consumer Research (1988) 253–264
25. Bransford, J., Johnson, M.: Consideration of some problems in comprehension. In Chase, W., ed.: Visual information processing. Academic Press, New York (1973)
26. Schubert, E., Poli, G., Roda, A., Canazza, S.: Music systemisers and music empathisers–do they rate expressiveness of computer generated performances the same? In: The International Computer Music Conference (ICMC2014) jointly with the Sound and Music Computing (SMC2014). (2014)
27. Davidson, J.W.: Visual perception of performance manner in the movements of solo musicians. Psychology of music **21** (1993) 103–113
28. Leman, M.: Music, gesture, and the formation of embodied meaning. In Godøy, R., Leman, M., eds.: Musical gestures: sound, movement, and meaning. Routledge, Oxon, UK (2010)
29. Stevens, C., Dean, R., Vincs, K., Schubert, E.: ln the heat of the moment: Audience real-time response to music and dance performance. In Burland, K., Pitts, S., eds.: Coughing and Clapping: Investigating Audience Experience. Ashgate Publishing, Ltd., Surrey, England (2014) 69–88
30. Schubert, E., Rodà, A., Canazza, S., De Poli, G.: Algorithms can mimic human piano performance: The deep blues of music. In: Frontiers in Performance Science, Kyoto, Japan (abstract accepted March 18, 2015)