Open ended descriptions of computer assisted interpretations of musical performance: An investigation of individual differences.

Emery Schubert
UNSW Australia, E.Schubert@unsw.edu.au;
Gunter Kreutz, Carl von Ossietzky
Universität Oldenburg, Germany, gunter.kreutz@uni-oldenburg.de

Abstract. This paper reports findings of a study in which a number of individual differences were examined in order to try to explain why people describe the same performance of a piece of computer generated/assisted music differently. Now that computers are able to replicate and be confused with human performances, attention is turning to methodological and psychological issues in assessment of the different musical outputs of various algorithms. A study with 103 participants was conducted using the same stimuli in an earlier study, but now with a larger sample. Open ended responses were requested to describe the different performances. Those responses were coded according to whether they described formal/structural aspects of the music versus aspects related to things outside the formal structure of the music, such as emotions, memories, characters and so forth. A semi-automated coding scheme was developed which returned a Cohen's-Kappa of .762. Individual differences in music cognitive style—music empathizing and music systemizing—demonstrated little predictive value. It was concluded that the complex question of individual differences and understanding why we respond differently to music is a large aim, and there is much value in reporting non-significant results to assist with further research.

Keywords: automated music performance; assessment; individual differences; music cognitive style; semi-automated content analysis

1 Introduction

Individual differences is a natural topic in psychology, in general, and in music psychology, in particular [1, 2]. However, despite some progress in this field, there is still paucity of research that examines individual differences in the reception of expressive music performance. For example, little is known about how individuals describe different computer generated performances and whether specific trait variables can explain differences in such responses. It is important for system developers and the robotics community to now begin to understand the more subtle issues affecting the appreciation of the computer performances in order to direct their

Workshop Proceedings of IAS-13
13th Intl. Conf. on Intelligent Autonomous Systems
Padova (Italy) July 15-19, 2014
ISBN 978-88-95872-06-3 pp. 565-573

research effort to the most relevant aspects and to improve the effectiveness of the systems. Furthermore, since we are investigating open ended descriptions, another important aim of this paper is to propose a novel method of content analysis that is simple to apply but exhibits reasonable face validity and reliability.

1.1 Aims and research questions

The aims of this study are: (1) To examine whether descriptions of different computer generated/assisted performances can be related to individual differences. (2) To develop a simple method of content analysis for fairly large data sets. To these ends, we ask to what extent so called expressionist and formalist responses are associated with respective cognitive styles, namely music empathizing (ME) and music systemizing (MS) [3].

2 Method

2.1 Participants and Procedures

Excluding invalid responses, 103 participants completed the study in return for course credit. The participants listened to the 9 stimuli and rated a number of qualities on a scale of 0 to 10 for each piece [for an overview of results of rated items, see 4]. Of interest in this paper, participants were asked to describe the different performances in their own words. Participants also completed the Music Cognitive Style scales [3] and provided additional background information using the KeySurvey interface.

2.2 Materials

The stimuli were different renderings of a short piano piece by Kuhlau, as described in [4] in the sequence Perf1, Perf2, Perf3, Perf4, Perf5 (human), followed by the first four stimuli presented again in the same order (hence '1234h1234'). A second cohort of participants was added in which the order of presentation was changed to 1234h1342. The first four pieces are referred to as 1a, 2a, 3a and 4a respectively. When played the second time they are referred to as 1b, 2b, 3b and 4b respectively. The systems used to generate the version are reported in [5, p. 354].

The human (h) performance is presented once only in the sequence. Participants completed the study via KeySurvey survey software (https://www.worldapp.com/surveys/overview.html), at their own pace on their own computer/sound-system. They were not told that some of the pieces were repeated.

Workshop Proceedings of IAS-13
13th Intl. Conf. on Intelligent Autonomous Systems
Padova (Italy) July 15-19, 2014
ISBN 978-88-95872-06-3 pp. 565-573

3 Results

A grand total of 927 data units were reported, of which 593 had some text, and 334 were left blank. Although the blank responses could reflect an inability of participants to provide a description of the performance, it could also mean that there was no interest in responding on the occasion of the investigation, and consequently, those data units were omitted from further analysis (hence, final N=87). A content analysis approach was taken to quantify the remaining open ended responses in terms of response style. The response styles of two kinds were coded as either expressionist or formalist or both, or neither [after 6]. Formalist responses are characterized by descriptions that refer directly to the music, that is, direct descriptions of the music (its tone, pace, loudness etc.) without references made to anything outside the music itself. Expressionist responses consisted of emotions, characters, memories and images to which the music refers. References to the beauty, expressiveness, enjoyment of the performance etc. were not coded as they were considered outcomes of the music listening process, rather than justifications/descriptions explaining those outcomes [this is a semantic complex process, but see 7, for the justification and approach that we take]. These categories were expected to reflect differences in individuals, particularly as measured by cognitive music styles. In a previous study, my colleagues and I hypothesized that empathisers were more likely to response to music in an expressionist way, while systemisers would respond in a formalist way [4, 8]. Because a reasonably large data set needed coding, a method of coding that was assisted with some level of automation was developed.

3.1 Semi-automated coding of open ended responses

There are numerous approaches to analysis of text. The approach taken is within the family of content analytic techniques. The relatively large amount of text to be coded (927 data units, coded twice each – for expressionist and formalist text, hence up to 1854 codings) encouraged the development of a simpler approach than line by line hand coding. For reasons of experimental economy [see 9, p. 241, for a discussion], and following a rather pragmatic approach as proposed by Grimmer & Stewart [10, p. 6], we aimed at finding a compromise between simplicity and speed of coding versus precision and validity. The latter authors also suggested that small losses are tolerable in larger data sets without loss of viable hypothesis testing.

The method outlined below is in accord with basic principles of text content analysis [in particular, 11, 12, 13] and can be applied to verbal data that are stored in commonly accessible formats such as Microsoft ('MS') Excel and Word.

1. All text was amalgamated and parsed into a word list. This was achieved by replacing spaces with return characters (can be done in MS Word by 'paste special' -> 'unformatted text', then using replace all spaces with return character, code ^p), and 'cleaning' each word by removing non alpha characters from the beginning and end of each word, should they appear – most commonly these were punctuation marks at the end of the word. This clearing was done in MS Excel.

2. A round of unusable text removal was applied. This was achieved by deleting words with three characters of fewer, as well as removal of stop-words. The list of stop-words available in the open source database, MySQL (http://dev.mysql.com/doc/refman/5.5/en/fulltext-stopwords.html) was used, which consists of 543 words.
3. Repetitions of words were then removed with a tally of number of words retained for possible later reference, and the final list is sorted in alphabetical order. Words with incorrect spelling are retained so as to capture responses with incorrect typing, provided the spelling does not introduce ambiguities.
4. This list is then coded, word by word according to the required criteria. In our analysis, 927 open ended responses consisting of 13,185 words (before step 1) were reduced to 1,057 unique words. The words were divided into columns according to the coding category to which they belong. In the present case there were two – on for expressionist responses and the other for formalist responses.
5. A search algorithm was applied to identify whether any of the 1,057 words appeared in each of the open ended data units. If so, the code was applied to that entry, one column per coding category.
6. The coded list was then checked by hand to ensure coding was meaningful, and captured the intended category uniquely, excluding other categories. Unusual and erroneous codings were corrected in the list prepared in step 4 and the search algorithm of step 5 reapplied. After some experimentation, a strict, rather than broad, inclusive coding strategy was adopted. That is, a word from the final list was coded into a category only if fairly certain to be part of that category, and at the same time found evidence that the term could *not* be linked to out-of-category phenomenon. This meant some loss of coded data units, but greater face validity. Such a trade off is considered reasonably when the data set is large. Thus, the process was repeated from step 4 until a satisfactory coding with good face validity was found.
7. A final set of checking was conducted. Data units that received no codes were double checked to make sure this was *not* due to omissions on the part of the overly limited, conservative word list (step 6). In some cases, longer expressions were included to the list of step 4 to correct for incorrectly uncoded data units provided overly general, ambiguous terms were not introduced. For example, one participant described stimulus 4a as "*urgent. not that* ***bright and relaxing***." Since this was an expressionist response (that is, not a description of the musical forms, but to what it refers), 'bright' was not in the list of words to be coded (because it could describe timbre which was considered formalist) or 'relaxing' or 'urgent' (possible felt emotion outcome and off task comment respectively) - all being considered too ambiguous. So, instead, the text string 'bright and relaxing' was added to the list of step four (since the terms together as a string 'bright and relaxing' did evoke a unique and exclusive expressionist response). This step was therefore a check and balance of the more conservative coding approach of step 6. The semi-automated coding search was then again repeated from step 4 until all coding was considered rigorous and complete.

Workshop Proceedings of IAS-13
13th Intl. Conf. on Intelligent Autonomous Systems
Padova (Italy) July 15-19, 2014
ISBN 978-88-95872-06-3 pp. 565-573

To test the validity and reliability of the system, a subset of data were coded manually before the semiautomatic system was developed. 504 lines of data were coded independently of the semi-automated coding. They were compared with the semi-automated codes allocated to the corresponding 504 lines using Cohen's Kappa procedure [14], returning $\kappa = 0.762$ (95 percent CI, 0.700 to 0.824), $p < 0.001$. This was considered sufficiently reliable to retain the semi-automated coding for further analysis.

173 units were coded as both Expressionist and Formalist such as one participant who described stimulus 1a as "*Very **cheerful** at B, a little bit **mysterious** and **playful*** at D as the beat goes up and down, then it's back to **quick fast paced** starting from E. At F, the notes are going down very swiftly, *thus preparing for a **feeling*** for ending." (italics is expressionist text, underline is formalist text, bold indicates the terms used by semi-automated content analyser to identify the respective categories. The capital letters refer to rehearsal marks on a musical score that was presented during listening, to which participants could refer if they wished). 287 units were coded as exclusively Formalist, 66 as exclusively Expressionist, and 58 could not be classified as either. The latter consisted of comments about repetition, fatiguing, inability to differentiate performances and off-task responses.

A general observation of the results is that there were a large number of responses that were coded as Formalist, and this outnumbered the Expressionist responses by about 3:2 (460 versus 293 respectively). This could be explained in a number of ways. First, perhaps there were a large number of musicians in the sample, who were more comfortable using the more technical language associated with formalist/structural language. However, evidence of this was not obvious, and the range of musical backgrounds of the participants was quite broad, as mentioned in the Participants section, above. A more likely explanation lies in the way that formalist coding was conducted. Any description was categorized as formalist that referred to the music, whether in technical language or non-technical, even if naive. For example, in response to stimulus 3a a participant reported "I like at the ending how it goes from **high to low key**. I think throughout this piece there was[sic] very expressive and interesting moments." which was coded as formalist because reference is made to a musical feature (from high to low key), even though the expression does not strictly make sense – the participant probably is referring to pitch rather than key. That is, well educated, scholarly descriptions of music were not required for coding as formalist. This way we reduced the possibility of excluding participants who had a particular processing style, but cold not express it due to a limited music education.

3.2 Individual Differences

Individual Music Cognitive Style was assessed as indicated in Kreutz et al [3], producing scores for each of two subscales, music empathizing (hence ME) and music systemizing (MS). The two scores were divided into terciles, with group 1 being the lowest scoring participants for each of ME and MS, and group 3 being the highest. This allowed us to directly investigate the relation between expressionist responses, which were expected to be related to ME, and formalist responses, which were

expected to be related to MS. A comparison of counts for each pairing is shown in Figure 1. The figure reveals no systematic relationships, with the exception of a trend for the MS levels in terms of the absence of formalist descriptions (left half of Figure 1b). Here we see the MS group with progressively higher scores exhibit a greater *absence* of formalist responses (absence of formalist response is indicated by a coding a 0 for a data unit description). However, none of the comparisons of terciles produced significant χ^2 at $p = 0.05$, and the conclusion, apart from the observed trend, is that music cognitive response style does not predict the kinds of descriptions made of the performances.

To further examine the possibility of a role of music cognitive style on description category, repeated measures ANOVA was conducted using the music cognitive style scores as a dependent variable, and classifying the absence/presence of expressionist and formalist descriptions as the two independent variables. Since no differences were statistically identified in the previous analysis, a liberal approach was taken, of pooling degrees of freedom for each response a participant made. Even so, no significant difference was found for any combination of independent variables against the ME and MS scores (Multivariate tests returned the lowest p value of 0.206 for the main effect ($F(1, 505) = 1.603$, Pillai's Trace).

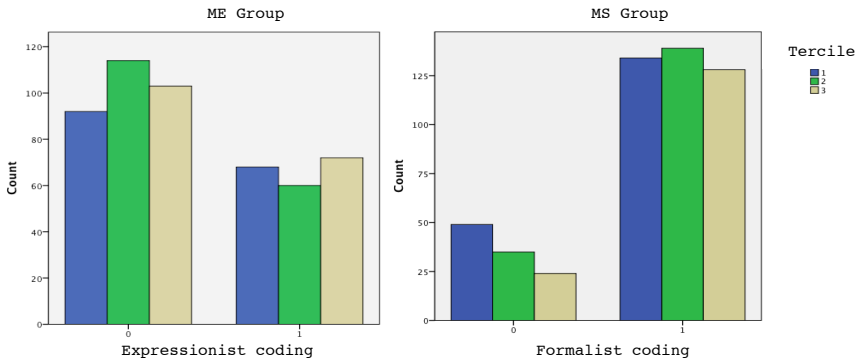

Figure 1. Bar charts showing count of codings of (a) ME score tercile (1 is bottom third of scores and 3 is top third of scores) by expressionist codings (1 present, 0 absent) in the left panel and (b) MS score terciles by formalist codings (1 present, 0 absent) in the right panel. None of the comparisons produce significant χ^2 statistics at $p = 0.05$. However, one expected trend can be observed – data units coded as being *not* formalist (0) decreased in number as Music Systemising scores increase. Note, too, that overall, more formalist responses are made (1 coding in b) than expressionist responses (1 coding in a).

3.3. Description type for human performance

A final analysis examined with description type was consistent with the explanation proposed by De Poli et al [8], that when a listener is aware that a computer is rendering a performance, listeners place less attention on basic technical aspects of the performance (assuming that this is trivial for a computer), and more on the emotive aspects. To support this, we might expect the number of formalist

descriptions to be higher and expressionist comments to be lower for the human performance, compared to the computer assisted/generated performances. Figure 2 shows the counts for each stimulus according to expressionist coding (a) and formalist coding (b). Overall χ^2 tests were significant for both expressionist coding (χ^2 (df = 8) = 18.325, p = 0.019) and formalist coding (χ^2 (df = 8) = 48.837, p < 0.001). Follow-up testing compared the human counts for each of the response styles against the corresponding highest expressionist count stimulus (which was stimulus 1a – see Figure 2a) and against the corresponding lowest formalist count stimulus (which was stimulus 2b – see Figure 2b). The expressionist analysis returned a significant result (χ^2 (df = 1) = 15.77, p < 0.001 with Bonferroni correction) and formalist analysis was non-significant (counts were identical [46] for stimulus 2b and stimulus h). This result provides some support for the explanation proposed by De Poli et al. It is also possible that participants were (a) cognizant at some level that the human performance was a human performance because (b) less attention was paid to it from an expressionist perspective (emotional character etc.) than at least one of the other computer generated performances. However, (c) the other performances were referred to with the same or more amount of formalist description. If the explanation proposed by De Poli et al [8] holds, if could be that the current coding scheme for formalist comments was too crude, and did differentiate between basic technical aspects (correct timing etc.), to which the participants (if aware that the performance was human, as per (a)) may have reported in one form or another (see De Poli et al for further details of this argument).

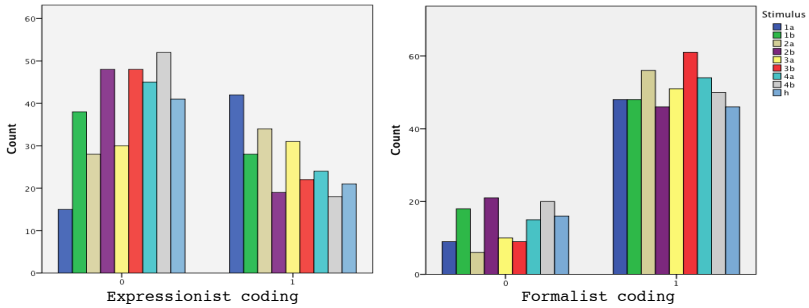

Figure 2. Bar charts showing count of expressionist codings in left panel (a) and formalist codings in right panel (b) by stimulus. Repeated stimuli are adjacent for ease of inspection of test retest reliability. Light blue bar is human performance (h).

4 Conclusion

This study extended recent work examining whether individual differences in music cognitive style could be used to explain individual differences in response to various computer renderings of the same piece of music. A content analysis method was also proposed to better understand how computer assisted music performances can be evaluated using open ended responses. In this study, our overall finding was that music cognitive styles, whether music empathizing or music systemising, is not implicated in differentiating different renderings of the same work. Apart from the

possibility that music cognitive styles do not contribute to differentiation in response justification, the following four possible explanations are offered. First, limitations of the Music Cognitive Styles need to be considered, as they share in common some of the concerns with their progenitor scales developed by Baron-Cohen. That is, either scale (ME or MS) may not be based on a single factor. In other words, the dichotomy of systemizing empathizing may not be sufficient to capture significant differences in attributing meaning to musical performances. Second, the lack of differentiation between ME and MS may be a result of one or more confounds. One confound could be enjoyment of music. ME-MS-theory is based on the notion that aesthetic enjoyment should be similar for music empathizers and systemizers alike. Third, coding of open ended data may have led to some variability. Future studies might put more weight on hand-coding to further test the validity of the semi-automated content analysis approach developed here, although we have evidence of its reliability. Fourth, a coding strategy that was highly specific was chosen to avoid false alarm coding, but allowing very general within-category terms so as not to exclude less educated participants from being correctly coded. Perhaps high music systemisers also tend to have more music education, meaning that a more strict, sophisticated language approach to coding may have produced different results. However, the data presented here, in concert with earlier analysis of quantitative data [4] most likely indicates that computer generated performances are now more commonly able to match the expressive capacity of human performances. There may be only a very small number of audiophiles who are able to distinguish between the source (human/computer) of the interpretation, suggesting that the computer assisted renderings have passed the musical expression Turing test. Indeed, this suggests that work on individual differences in responses needs to take a more central role in understanding how to examine and assess expressiveness in performance, both from an experimental design and psychological perspective. This development is needed to complement the increasing sophistication of computer generated and assisted music performance.

Acknowledgements

This research was supported by the Australian Research Council.

References

[1] A. E. Kemp, *The musical temperament* Oxford: Oxford University Press, 1996.

[2] P. J. Rentfrow, L. R. Goldberg, and D. J. Levitin, "The structure of musical preferences: A five-factor model," *Journal of Personality and Social Psychology,* vol. 100, p. 1139, 2011.

[3] G. Kreutz, E. Schubert, and L. A. Mitchell, "Cognitive styles of music listening," *Music Perception,* vol. 26, pp. 57-73, 2008.

[4] E. Schubert, G. D. Poli, A. Roda, and S. Canazza, "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)*, Athens, Greece, accepted.

[5] S. Canazza, G. De Poli, and A. Roda, "How do people assess computer generated expressive music performances?," in *Proceedings of the Sound and Music Computing Conference 2013 (SMC 2013)*, Stockholm, Sweden, 2013, pp. 353-359.

[6] B. Reimer, "Should there be a universal philosophy of music education?," *International Journal of Music Education,* vol. 29, pp. 4-21, 1997.

[7] E. Schubert, "Loved music can make a listener feel negative emotions," *Musicae Scientiae,* vol. 17, pp. 11-26, 2013.

[8] G. De Poli, S. Canazza, A. Rodà, and E. Schubert, "The role of individual difference in judging expressiveness of computer assisted music performances by experts," *ACM Transactions on Applied Perception,* under review.

[9] D. J. Hopkins and G. King, "A method of automated nonparametric content analysis for social science," *American Journal of Political Science,* vol. 54, pp. 229-247, 2010.

[10] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," *Political Analysis,* vol. 21, pp. 267-297, 2013.

[11] B. Downe−Wamboldt, "Content analysis: method, applications, and issues," *Health care for women international,* vol. 13, pp. 313-321, 1992.

[12] P. Gottschalk, "Descriptions of responsibility for implementation: a content analysis of strategic information systems/technology planning documents," *Technological Forecasting and Social Change,* vol. 68, pp. 207-221, 2001.

[13] K. Krippendorff, *Content analysis: An introduction to its methodology*, 4th ed. Newbury Park, CA: Sage, 1980.

[14] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics,* vol. 33, pp. 159-174, 1977.