The Role of Musical Features in the Perception of Initial Emotion

David Taylor[1], Emery Schubert[1], Sam Ferguson[2], Gary McPherson[3]

[1] Empirical Musicology Group, University of New South Wales, Sydney, Australia
[2] University of Technology, Sydney, Australia
[3] University of Melbourne, Melbourne, Australia
david.anthony.taylor@gmail.com

Abstract. 170 participants were played short excerpts of orchestral music and instructed to move a mouse cursor as quickly as possible to one of six faces that best corresponded to the emotion they thought the music expressed. Excerpts were analysed and the musical cues coded. Relationships between the number of cues and participants' response times were investigated and reported. No relationship between the number of cues available to the listener and the speed of response was found. Findings suggest that the initial response to ecologically plausible musical excerpts is quite complex, and requires further investigation to provide emotion-retrieval models of music with psychologically driven data.

Keywords: Music, emotion, modelling, initial response, cue utilisation, response speed

1 Introduction

Modelling emotional responses to music has developed considerably in the last ten years [1]. We now have models that can predict a typical emotional response to a piece of music expected from an individual in a Western culture reasonably accurately, simply by processing certain features extracted from the same piece [2]. Furthermore, use of continuous response methods has allowed understanding and modelling of moment-to-moment changes in musical features and subsequent emotional response [3-6]. Of the many dilemmas these approaches have left, however, continuous responses have highlighted a curious problem regarding the initial response to a piece of music. Recent research suggests that it takes some eight seconds after the start of a piece of music before a listeners emotional response 'settles' or becomes reliable [7, 8].

Since emotional responses appear not to be immediate according to continuous response studies, but can nevertheless be performed quite quickly according to post-performance response data, the question of how quickly one can decide on emotion expressed by music becomes an inviting and relevant question.

Peretz et al. [9] reported that participants were able to differentiate correctly between 'happy' and 'sad' emotions based on mode and tempo as quickly as 0.25 of a second. In a similar study, Bigand et al. [10] compared participants' grouping of one-second excerpts into groups of similar emotional character with the grouping of

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

twenty-five-second excerpts. Strong correlations between the groupings for each excerpt duration confirmed the Peretz et al. [9] findings that only a very small amount of time is needed to induce strong emotional responses. Although these studies demonstrate the extreme rapidity with which emotional responses can be made, the area of reaction time, or 'response' time is one that has remained relatively unstudied and calls for further insight. Bachorik et al. [7] looked at the response time (or what the authors refer to as 'integration time') for participants whom were expressly instructed to 'move [a] joystick as soon as they began feeling an emotional response to the music' whilst their movements were plotted on a two-dimensional grid of arousal and valence. The 'integration time' was measured as the time taken for participants to make a movement of the mouse beyond a pre-determined 'jitter' of 15 pixels. However the study only reports typical integration times (between 8.31s and 11s) and again fails to scrutinise individual results that are faster than this, or indeed, faster than the one-second reported in Bigand et al. [10].

On further scrutiny of the musical content (or 'musical and psychoacoustical structures') of their excerpts, Bigand et al. [10] suggested that the responses were governed by highly cultural compositional and performance-related features or cues. They discovered that many of the one-second excerpts (half of all cases) contained only a single chord or interval, and some only a single pitch and concluded after a 'cautious analysis' that performance cues within the music are enough to induce emotions in Western listeners at this very quick speed. Similar studies that yield comparable response times in making non-emotional evaluations of music would tend to support this [11]. By examining both performance-rated and non-performance-related cues from the position of the fastest possible emotional responses it may be possible to build a clear picture of a) which cues are utilised b) the number necessary in order to make a decision and c) how cues are utilised (i.e. based on their 'usefulness' in any hierarchical structures). By looking at fastest plausible response times it is possible to say with some degree of confidence 'this much music was required before the participant was able to make an emotional response'. By then carefully analysing the musical content and coding the cues with it, it may be possible to start to see more clearly what type or number of cue listeners most rely on.

Substantial work has been undertaken to identify the factors within musical structures that allow us to perceive emotional expression. For example, fast tempo has been associated with the expression of emotions such as 'exciting', 'happy' and 'glad' whereas slow tempo has been associated emotions such as 'serenity' and 'sadness' (for a summary of musical features from reviewed studies see: [12]). Juslin asserts that the number of cues impinges directly on the music's effectiveness to communicate emotion [13], however little research has attempted to reconcile these findings with the initial moments of an emotional response in the listener. This paper will, therefore, ask a number of questions: How quickly is it feasibly possible to recognise emotions in music? Does the number of musical features or 'cues' have any effect on the response time? How many cues are needed before a listener can make an informed decision? The overarching hypothesis is that there exists a cue accumulation effect whereby ambiguous cues (i.e. cues that provide information that conflicts with that from the majority of others) delay the evaluation of cues already made available by requiring further cues in order to confirm an assessment resulting in longer response times.

2 Method

2.1 Participants

A total of 170 participants took part in the experiment, 101 female and 69 male. All were tertiary level students either undergraduate or postgraduate between the ages of 17 to 50 (median age = 21) with a mean average of 6.99 years musical experience (range = 0 – 30 years).

2.2 Materials

The experiment used a pool of 19 short excerpts (duration = 7 s – 27 s) taken from the soundtracks of Disney Pixar animated films. It was deemed that music from this genre had high potential in best conveying the target emotions due to their programmatic and narrative nature. A pilot study was conducted to ascertain which emotion each excerpt was deemed to express the best. In this pilot study, participants were asked to select from a list of emotions (excited, happy, calm, sad, scared or angry) the one they thought was most applicable to each excerpt used in the current trial. The results were consolidated to arrive at a putative emotion for each excerpt (this became the 'target emotion'. See: [14]). All excerpts were purely instrumental. Each target emotion had 3 excerpts from which the software (written by author SF using Max 5) used in the trial could select, with the exception of the excited target emotion, which contained an extra excerpt in order to avoid better the possibility of participants guessing the last target emotion to be played. The software would randomly select one excerpt at a time from each of the target emotions whilst displaying a question at the top of the screen asking participants to identify the emotion they think the music best expresses (as opposed to the emotion the music makes them feel).

2.3 Procedure

Participants were presented with an on-screen display of six 'target faces' arranged in a circle. Each face was a simple representation of a target emotion (either excited, happy, calm, sad, scared or angry) and ordered according to valence and arousal (level of valence arranged horizontally and arousal vertically so that target emotions of similar valence and arousal are adjacent). To further aid quick reference and elicit the fastest possible response, the faces were also coloured (red for angry; yellow for excited, happy and calm; blue for sad, and darker blue for scared). Participants were explicitly instructed to move the mouse cursor to their decision as quickly as possible. Through headphones, they were then played a series of seven excerpts randomly drawn by the software from the pool. The participant initiated playback of each excerpt by clicking a green quaver symbol in the centre of the circle. The computer

then logged each participant's time for their 'out of box' time: the time it took them to move the cursor out of a hidden centre box, and their 'first face' time: the time it took them to move the cursor to the face they first selected. The software also logged all faces 'visited' and in which order together with all other cursor movements. It is to these tasks and responses that the present study refers.

3 Results and Discussion

Data were separated into two groups: data from those participants whose first face was that of the target emotion, plus or minus one (one face either side of the target face); and data from those whose first face was not (but instead was in another location on the screen, including one of the three remaining faces). The second group was eliminated from further analysis. The reason for this was because we are interested, in this study, in the fastest plausible response time after the music begins.

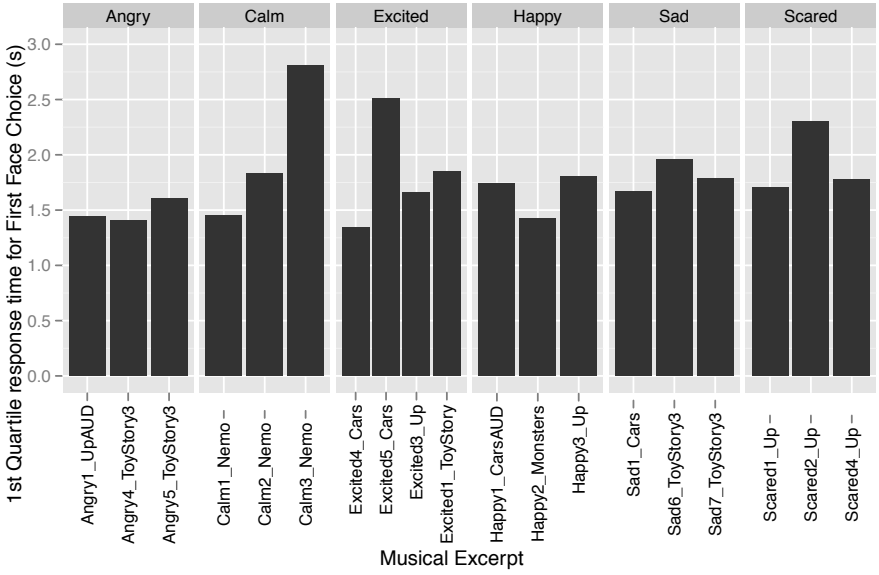

Fig. 1. Comparison of first quartile 'first face' times between individual excerpts

Fig. 1 shows the response times for each individual excerpt by target emotion (the emotion we supposed that the participant would select). First quartile response times (the first face time of the participant who was 25% of participants behind the fastest participant's first face time) were examined because we were interested in the fastest plausible time that participants could make emotional responses. Median time, for example, gives an estimate of typical response time, and minimum response time is

susceptible to error, for example due to prevarication, random error or guessing. First quartile time was therefore deemed a conservative and reasonable comprise (what we term 'plausibly fastest'). From a null-hypothesis perspective, each response time within a target emotion would be identical, which seemed to be the case for 'angry' excerpts and for 'sad' excerpts. Yet some excerpts in other emotions, such as excerpt 3 for 'calm' and excerpt 2 for 'excited' and 'scared' stand out.

Therefore, the question arose, considering that both calm excerpts 1 and 3 were putatively considered to be good examples of music that expressed that emotion in the pilot study, why was there such a difference in response times? The difference suggested that something in the music of the first excerpt made it easier to judge quickly the intended expression. These two excerpts were therefore selected for closer analysis.

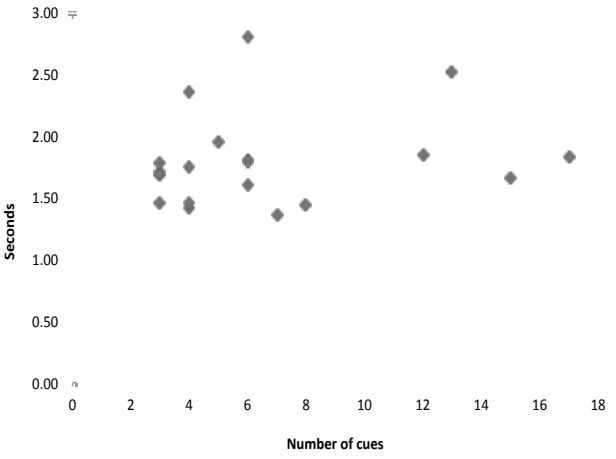

Fig. 2. Scatter chart depicting relationship between 'first face' first quartile response times for each excerpt and number of cues available to the listener (total number of cues counted in that excerpt before the first quartile time)

We were interested to see whether the number of cues available to the listener was in any way related to the speed at which responses were made, i.e. the more cues there were available to the participant, the quicker their response. The first level of cue analysis involved a very basic level of coding. Using 'Audacity', each excerpt was analysed in spectrographic form. A time label was entered at each occurrence of an 'event' (e.g. a note, chord, cymbal clash) between the start of the excerpt and first quartile time. At this level of coding, vertical events (i.e. simultaneous notes from different instruments or chords) were treated as one event. It was not possible to situate cues that unfolded in time in one location, i.e. with regard to a change in loudness where the actual cue started or ended, or indeed where the information from the cue was gleaned. Therefore, for simplicity, cues of variance were omitted at this

stage. .txt files of the cue labels and times were then exported to facilitate counting. Once the cues for each excerpt had been counted they were plotted on a graph along with the first quartile time (Fig. 2).

The main hypothesis was that there would be a correlation between the number of non-ambiguous cues available to the participants and the speed of their response. However, Fig. 2 shows that, at least with a basic count of cues, this did not appear to be the case. If it were, we might expect a trend to be visible where as the number of cues increases, the response time decreases. The first quartile time for Angry1_Up was one of the fastest (at 1.45s), yet participants only received three cues before that time, whereas the fastest 25% of participants responding to Excited5_Cars required thirteen unambiguous cues before feeling able to make a decision (managing only a first quartile time of 2.51s).

This assumes however, that the spacing of cues was consistent and regular, but it may be that in some cases, for example, very few cues were available in the first second followed by a many number in quick succession. Also, clearly the slower the response time, the more cues there will be counted – simply due to the duration. If Angry1_Up had, for example all three cues in the first second, and Excited5_Cars had ten of its thirteen in the first second yet still yielded the slow first quartile time of 2.51s, it would be much stronger evidence against the number of cues being important. Therefore, we counted the number of cues available in each excerpt within the first second only.

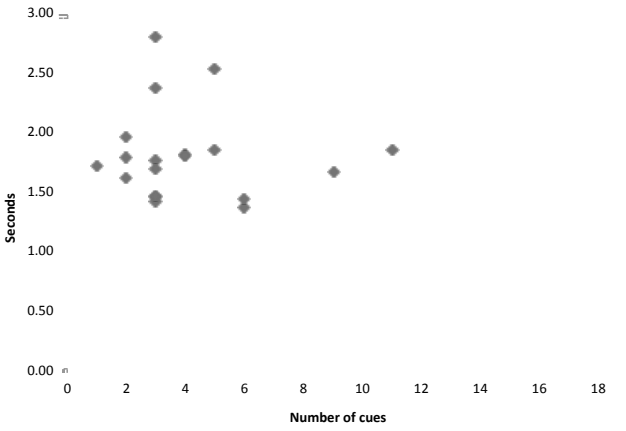

Fig. 3. Scatter chart depicting relationship between 'first face' first quartile response times for each excerpt and number of cues available to participants within the first second

Again, there was no link between the number of cues available to the participant in the first second and the response times. Furthermore, Fig. 3 shows that six of the excerpts all contained three cues in the first second of music but yielded response times ranging from 1.40s to 2.79s.

4 Conclusions and Further Research

The present study examined timing and number of cues and suggests that the initial emotional response to a piece of music is quite complex, and requires considerable further research. Further research is required to examine whether the quality of the cues are relevant in determining response speed. For example, some cues might have a greater weighting than others, giving rise to a hierarchical cue structure, as is the case in some theories of music cognition [15]. While other studies described above have identified rapid identification of emotions in the order of one second, our study was conducted with ecologically typical musical extracts, allowing the underlying complexity of emotion response time to the start of a piece of music to emerge. This has important implications for automated modelling of emotion in music systems because until we can retrieve the underlying nature of human emotional response to music at this transitional, initial orienting period, it will be difficult for time varying models to produce psychologically plausible representations of emotions at the start of a piece of music.

Acknowledgments. This research was funded by the Australian Research Council (DP1094998).

References

1 Schubert, E.: Continuous self-report methods. In: Juslin, P.N., & J. A. Sloboda (eds.) Handbook of music and emotion, pp. 223-253. Oxford University Press (2010),
2 Hug, A., Bello, J.P., and Rowe, R.: Automated Music Emotion Recognition: A Systematic Evaluation. Journal of New Music Research, 39, (3), pp. 227-224 (2010)
3 Schubert, E.: Modeling percieved emotion with continuous musical features. Music Perception, 21, pp. 561-585 (2004)
4 Schubert, E.: Continuous measurement of self-report emotional response to music. In: Juslin, P.N., & Sloboda, J. A. (eds.) Music and Emotion: Theory and research, pp. 393-414. Oxford University Press (2001)
5 Schubert, E.: Measuring emotion continuously: Validity and reliability of the two-dimensioned emotion-space. Australian Journal of Psychology, 51, pp. 154-165 (1999)
6 Schubert, E.: Measurement and time series analysis of emotion in music. Unpublished doctoral dissertation, University of New South Wales (1999)
7 Bachorik, J.P., Bangert, M., Loui, P., Larke, K., Berger, J., Rowe, R., and Schlaug, G.: Emotion in Motion: Investigating the Time-Course of Emotional Judgments of Musical Stimuli. Music Perception, 26, (4), pp. 355-364 (2009)
8 Schubert, E.: Reliability issues regarding the beginning, middle and end of continuous emotion ratings to music. Psychology of Music, In Press
9 Peretz, I., Gagnon, L., and Bouchard, B.: Music and emotion: Perceptual determinants, immediacy, and isolation after brain damage. Cognition, 68, pp. 111-141 (1998)
10 Bigand, E., Vieillard, S., Madurell, F., Marozeau, J., and Dacquet, A.: Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts. Cognition & Emotion, 19, pp. 1113-1139 (2005)
11 Gjerdingen, R.O., and Perrot, D.: Scanning the dial: The rapid recognition of music genres. Journal of New Music Research, 37, (2), pp. 93-100 (2008)

12 Gabrielsson, A., and Lindstrom, E.: The role of structure in the musical expression of emotions. In: Juslin, P.N., & Sloboda, J. A. (eds.) Handbook of Music and Emotion, pp. 367-400. Oxford University Press (2010)
13 Juslin, P.N.: A Brunswikian approach to emotional communication in music performance. In: Hammond, K.R., and Stewart, T.R. (eds.) The Essential Brunswik: Beginnings, Explications, Applications, pp. 426-430. Oxford University Press (2001)
14 Schubert, E., Ferguson, S., Farrar, N., and McPherson, G.E.: Sonification of Emotion 1: Film Music. In: 17th International Conference on Auditory Display (ICAD 2011). Budapest, Hungary: International Community for Auditory Display (ICAD) (2011)
15 Lerdahl, F., and Jackendoff, R.: A Generative theory of tonal music. MIT Press (1983)