In A State: Live Emotion Detection and Visualisation for Music Performance

Adinda van 't Klooster
adindavantklooster.com
adinda_18@hotmail.com

Nick Collins
Durham University
nick.collins@durham.ac.uk

ABSTRACT

In A State is a live performance system that detects emotion from live audio (in this case a piano performance) and generates visuals and electroacoustic music in response. This paper discusses the continuing development of the system and the backdrop of emotion in music and arts practice.

Keywords

Emotion models and detection, live performance, art, visualisation, live electroacoustic agent

1. INTRODUCTION

This paper explores the basis for a new live performance system incorporating elements of visualisation and electroacoustic music, driven by computer-perceived emotion. Emotion is a complex topic much studied in music and arguably equally central to the visual arts [12] where this is usually referred to with the overarching label of aesthetics. We begin by surveying existing research and practice in the arts and emotion, with a particular view to the assumptions underlying such systems and research. Models of emotion are central to the ensuing performance system, and this critical approach is part of the aesthetic goal of the project.

The first challenge in making work related to emotion is to define emotion, and which underlying emotion model to choose. There are two main models of emotion that are widely used: the discrete and the dimensional model. The discrete model suggests that there are certain basic universal emotions, where the list typically includes anger, fear, enjoyment, disgust, surprise and sadness [10]. The dimensional model is also called the circumplex model of affect and suggests that all emotions derive from two neurophysiological systems, one related to arousal and the other to valence [24].

The term valence in this context relates to being attracted (positive valence) or repulsed (negative valence) by a stimulus. Arousal denotes intensity and can be more directly read with physiological sensing: a basic GSR (Galvanic Skin response) sensor can provide this information [18]. Whether an emotion is positive or negative is much harder to establish with technology and generally needs to be self-reported. As the circumplex model of affect uses a sliding scale of emotion it leaves more space for different shades and intensities of emotions than the discrete model.

When studying emotional response to music or art both models have their problems as the emotional response in these specific domains includes emotions of an aesthetic nature, over and above the basic emotions of enjoyment, sadness and surprise. Art historian Roger Fry contemplated in 1909 that aesthetic emotion is different from other emotions in its lack of calling us to action: "Morality, then, appreciates emotion by the standard of resultant action. Art appreciates emotion in and for itself. " [11]

NIME'14, June 30- July 3, 2014, Goldsmiths, London, UK.

More recently Scherer and Zentner [25] proposed to differentiate between utilitarian and aesthetic emotions. They describe utilitarian emotions as high-intensity emotions that tend to make the individual prepare for action with their well-being or even survival at stake, and denote aesthetic emotions to be more subtle in nature and triggered in situations that make no immediate material difference to the individual's well-being. Through a series of interrelated studies they compiled a list of music-specific emotions that participants most frequently claimed to experience in response to music: wonder, transcendence, tenderness, peacefulness, nostalgia, power, joyful entrainment, tension, and sadness. Scherer and Zentner's model is based on the discrete model of emotion with its inherent flaws of leaving little space for different shades of emotion, and remains vulnerable to lack of agreement on which emotions ought to be included. The circumplex model of emotion can offer space for many different emotions as (theoretically) all imaginable emotions can be placed within the two dimensional plot. This may be why in the past ten years the circumplex model of affect has overtaken the model of discrete emotions in popularity in the area of music and emotion research [7].

One of the exciting developments in music and emotion research is its interdisciplinary. Advances in psychology, computing and audio processing and technologies such as biosensors have all contributed to the field. In comparison, in the field of the visual arts, over the past few decades the debate on emotion and art has remained largely philosophical and has focussed more on defining what the aesthetic experience is than on how this might be measured. An important exception to this statement would be Berlyne, who proposed a 'new experimental aesthetics' that focussed on arousal as a measurable emotional response to art. Berlyne suggested that four elements were of particular importance in achieving optimum arousal: complexity, novelty, ambiguity and 'puzzlingness'. [3]. Over the years Berlyne changed his mind about what the ultimate level of arousal is [2,3] but his arousal theory is weakened by its emphasis on arousal as the main indicator of an individual's preference [26,27].

In current emotion psychology the theory of appraisal has surpassed the above outlined arousal theory. The appraisal theory proposes that not objects or events themselves but our appraisal of them are the cause of our particular emotional experience [23]. This is diagonally opposite to what upholders of a formalist view of art, like Roger Fry, believe, namely that the formal components within the artwork generate the aesthetic experience [6]. Whilst a thorough discussion of the nature of aesthetic emotion is outside the remit of this paper, it should be clear that as of yet there are no clear answers. Interestingly though, neuroscience has taken an interest in the subject matter and recent research suggests that the individuals' taste in art is linked with their sense of identity [31].

2. ARTWORKS THAT TRACK EMOTION

The technologies used to track emotion are diverse and complex. As input, a combination of physiological signals can be used, but facial recognition and voice analysis are also popular candidates and gesture recognition is another technique to obtain emotional content [17].

For a review of artworks that use physiological sensing to obtain more information on the viewers' internal state, see [30, chapter 3]. A selection of artworks that track emotion via

other means are listed below; these works show a prevalence for the model of discrete emotions. The non-live work *Cheese* by Christian Moeller picks up on the authenticity of a smile. The work consists of six flatbed monitors each showing a video of a different actress who has been instructed to keep smiling for an hour and a half. The smile is analysed by bespoke software and judged on its sincerity. When a smile does not pass the 'sincerity threshold', an alarm goes off and alerts the actress to be more authentic [20].

Tina Gonsalves made a body of work based on emotion. One work, *Chameleon*, is a collaboration with neuroscientists Chris Frith and Hugo Critchley and uses facial recognition software to assess the emotional state of the gallery visitor [13]. A visitor's emotion is mirrored back to them with the distortion that the emotions of the people standing in front of the two other screens are averaged with theirs. Gonsalves based her system on the discrete model of the six basic emotions by Paul Ekman, with the difference that she replaced fear with a neutral state. This may be because of Ekman's observation that the difference between fear and surprise was not easily distinguishable across different cultures [9].

Naoko Tosa, Joy Nicholson and Ryohei Nakatsu explored the area of using emotional content in speech to control the behaviour of virtual characters. They created an emotional recognition algorithm by collecting a large speech database and training a neural network using this database. They obtained 50% accuracy for speaker and content independent emotion recognition. They worked with eight emotional states: anger, sadness, happiness, fear, surprise, disgust, playfulness and neutrality. They claimed they could develop realistic computer characters with an ability for spontaneous interaction [21], however, the level of accuracy left something to be desired.

Naoko Tosa applied this research to the 'Network Neuro-Baby with robotics hand', which she classes not as an artwork, but as an 'automatic facial expression synthesizer that responds to expressions of feelings in the human voice and handshake.' This performance tool generates a baby face with facial expressions that respond to the human voice. If the speaker's tone is gentle and soothing, the baby smiles and laughs. If the speaker's voice is low or threatening, the baby responds with a sad or angry expression and voice. Loud, sudden or disproving sounds will make the baby cry. The baby is programmed to keep eye contact through the use of an Active Eye Sensing System. A hand-shaking device gathers further information on the emotional state of the interactant [29].

3. LIVE EMOTION SYSTEMS FOR MUSIC

Live performance systems in music based on emotion have mostly been based on physiological sensing [16,22], attempting to access inner emotional state. The emosynth project of Valery Vermeulen [32], for instance, investigated audiovisual material control from biosignals, with a training phase measuring a user's response to provided material, and a synthesis stage for new content.

There is a role too, however, for models of perceived emotion. Here, the natural detection modality is to work on an audio signal itself. Detection of emotion via audio content has been investigated within the music information retrieval community [15,19]. The majority of work has concentrated on the offline case, and a commercial app such as MoodAgent relies on previously established mood profiles on a company server deriving from hand-annotation (http://www.moodagent.com/faq; the app in question here is somewhat strange in listing tempo as an emotion!).

Thorogood and Pasquier [28], presented a soundscape affect classifier system that uses two axes, one for arousal and one for valence, with multiple linear regression; the evaluation presently seems to have been conducted offline based on data from one expert user. Tuomas Eerola has previously created a realtime emotion visualizer using a cartoon-style 2D emotion mapping from 5 acoustic features [8]. The aim of such work is to create an 'emotional agent' [1,4] that can detect the emotion expressed in music. Such a system can be trained from annotated audio files and then applied to live audio in a concert situation to give decisions on observed emotional content.

4. A PROTOTYPE DISCRETE EMOTION DETECTION SYSTEM

As a step towards our emotion-aware system, we first investigated the detection of perceived discrete emotions from audio alone. Four acoustic piano improvisations of five and a quarter minutes each for the emotional states of happiness, anger, sadness and tenderness were recorded with two different microphones (capacitor and dynamic) simultaneously. The piano improvisations were created (improvised) by the second author following the guide of Eerola [8] for the 'typical' characteristics of these states combined with the pianist's personal opinion of what these emotions should sound like. Eerola [8] suggests that 'Anger…is characterized by a high sound level, fast tempo, high pitch variability, high-frequency energy content, and fast tone attacks' [8, p.611]) and provides similar data for the other three emotions. Training of a machine learning algorithm took place offline on half the material, with the other half used for testing of generalisation performance. Implementation used the SCMIR Music Information Retrieval Library for SuperCollider [5], which allows straightforward translation from offline training to a live system.

The ten audio features were perceptual loudness, spectral centroid, tempo, musical key clarity (the confidence in one key over others in a key detector), musical key detected (especially important for the major/minor distinction), attack slope (of detected onsets, as an articulation cue), and four energy bands split equally within 7 octaves (as a measure of registral area on the piano). These features were selected on the basis of Eerola's study [8] concerning effective audio features used in an offline study of affect in soundtracks, that linked themselves to the four emotional states of happy, sad, angry, and tender. Features were extracted based on a frame rate of approximately 43 per second, with an overlap factor of 2 for windows of 2048 samples at 44100Hz sampling rate. Features were normalized with respect to global minimum and maximum values extracted from the corpus (the feature extraction stage including normalization factors carries forwards to the live system).

After extraction, feature vectors were aggregated by means within one-second windows, calculated with a hop size of 0.1 seconds. This hop size ensures that a larger number of training and testing examples are created, and avoids biasing the system to too gross a grid (robust generalisation was found to depend on this step). The mean vectors were then aggregated further into four one-second windows (spaced by one-second gaps), covering the last four seconds with a bag of frames combination of 40 features. A NeuralNet with 40 inputs, 40 hidden units and 4 outputs (one for each emotion class) was then training on roughly half the data (12724 instances), and tested on the other half (12728); presentation order of instances was randomised to avoid class bias in training, and training was over 1000 epochs (other parameters were explored empirically without any increase in performance; the implementation was the SuperCollider NeuralNet class).

The generalisation performance recorded was 64.9% accuracy over the test set (8260 instances correctly classified out of 12728), competitive with results for offline systems reported in [15]. Training set accuracy was 100%, showing that the neural net had perfectly attuned itself to the training examples provided, but it is generalisation ability on the separate test set that determines robustness for live

performance. The confusion matrix appears in Table 1. We see that the classifier spots the main categories, but that the classes of tender and sad engender more confusion, probably due to the signal characteristics of lower volume and more muted playing in both cases (though minor/major should have been a distinction for sad/tender, key is not a perfect discriminator).

Table 1 Confusion Matrix for Neural Net over test set

		Predicted Class			
		Angry	Happy	Sad	Tender
True Class	Angry	2473	575	98	26
	Happy	333	2294	174	335
	Sad	386	340	1366	1124
	Tender	7	1	1069	2127

A live classifier was built simply by translating the offline SCMIRAudioFile feature extractor to an SCMIRLive instance with some appropriate feature aggregation code to match the procedures detailed above; the NeuralNet had already been saved and could be brought in ready to use in native SuperCollider code. Once built, qualitative analysis of the discrete emotional state based classifier indicated that it could recognise emotional states, though with some momentary errors in classification, and a reduced effectiveness in the condition of novel input. We must be careful to admit that the machine is most likely picking up on different playing styles associated with certain emotional tags, via the audio feature data, and not itself understanding emotion at a high level.

Hochenbaum et al. [14] suggest in multimodal musician recognition that larger windows, in their case 15 seconds, assist recognition (although such large windows may not be plausible in human-like performance given the potential quick reactions of human beings to musical stimuli within at least short-term memory). We further tested using eight-second windows, achieving a top accuracy over the test set of 71% (and a similar confusion matrix), and reduced performance for shorter windows. Nonetheless, the four second reacting system above is the one used in a live prototype so far, providing a compromise of performance and latency.

5. *IN A STATE* PERFORMANCE INTERFACE

As this performance projects aims to address some of the ambiguities in emotion research we wanted to use multiple models of emotion. We expanded the system to include an arousal-valence detection system from piano input and are further developing capacity for the computer to analyse its own electroacoustic output via discrete and continuous models. For the piano-led arousal-valence detection system, a GUI was built for the collection of arousal-valence data (see Figure 1). This was used by the second author whilst listening to old piano improvisations by himself (15 minutes worth), obtaining continuous arousal and valence readings. Similar machine learning to section 4 was applied. Training and test instances were created in a 50/50 split, based on a window size of four seconds for averaging features and arousal-valence, with a hop size of 0.1 seconds. In order to measure accuracy, we used a basic tolerance for the correct region of detection; a predicted arousal-valence output from the neural net was marked as a correct prediction if it was within a Euclidean distance under 0.05 of the real value. Current performance is around 88% on training data, but 25% on test. For comparison, on test data random choice of a point in arousal-valence space performs at 0.02%, and always choosing the centre (0,0) at 4%. Accuracy could be improved in future by working with a larger corpus.

For the final concert system the two emotion detection systems (discrete and continuous), are used alongside each other as inputs to generate animated graphics created in Processing. For this artistic part of the interface, the arousal-valence data is mapped to background colour, using the computers' color wheel (see Figure 2).

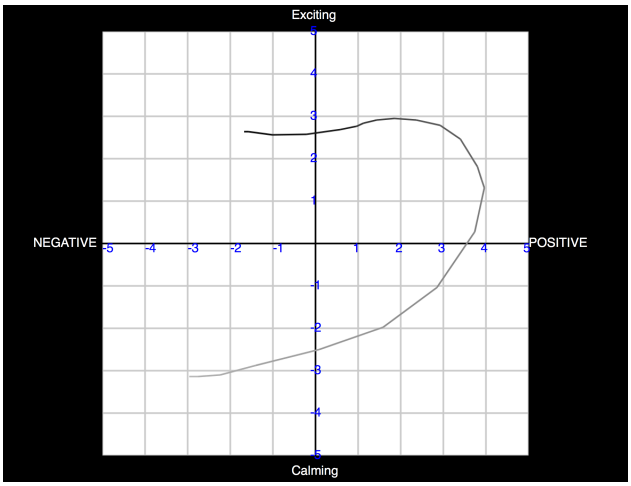

Figure 1 Arousal-valence collection GUI

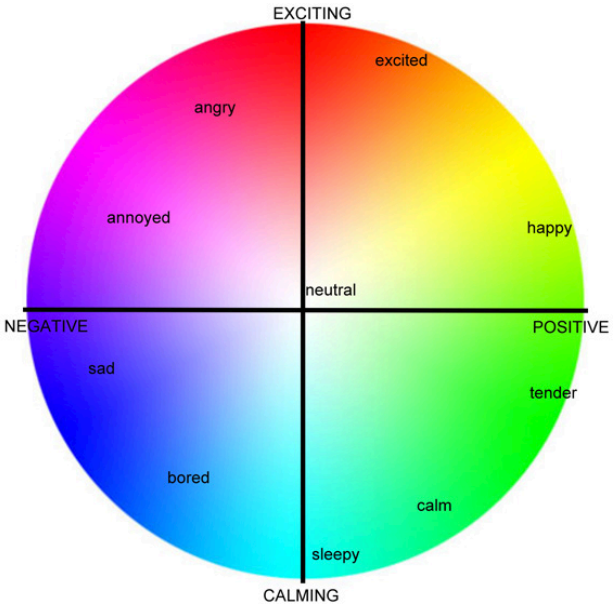

Figure 2 Arousal-valence space by Russell (1980) with added colour and emotion mapping

To give an idea of where different emotions fit roughly within Russell's [24] arousal-valence space, nine emotions and a theoretical 'neutral' state are layered on top. When the arousal-valence data is used as main input, coming within a certain range of these positions of the arousal-valence space can trigger different abstract animations to visualise the nine emotions: excited, happy, tender, calm, sleepy, bored, sad, annoyed or angry. The abstract animations are based on digitised ink drawings by the first author and take up the majority of the screen size. There are also smaller number boxes to display live accuracy readings for the four previously mentioned discrete emotional states and a small graphic display to show the detected position within the arousal valence space. At certain parts these models may compete for deciding the output.

There are two projection screens planned for the premiere performance at NIME 2014. One visualises the perceived emotion in the piano music and the other visualises the perceived emotion in the electronic response from the computer agent. The computer's

sounding response is based on a set of scenes, each one a generative system. Responses include live feature-led audio processing of audio, synthesis of new gestures, and manipulation of sounds evocative of particular emotional states. The scenes are themselves motivated by musical emotion literature, and use available parameters of the emotion models (e.g. the current discrete emotion and arousal-valence value). The pianist can use the visualisation of the computer agent's music to determine approaches to improvisation. A feedback loop is established via multiple modalities; aural feedback of musician to computer to musician and round again, with visual feedback depicting tracked emotional state as a further channel. The whole system becomes a complex audiovisual system mediated by models of emotion.

The idea of using such an interface in the context of live performance is to visualize emotion in multiple ways and to highlight the problematic nature of emotion research with its many ambiguities. As accuracies reported above suggest, the computer will not always get it right when estimating the emotions expressed in the music. This must not be viewed as a shortcoming of the interface but rather as representative of how far the field of emotion research has come and will still have to travel. Artistic licence ensures the visual experience is rewarding in its own right and goes beyond a bare scientific representation.

6. REFERENCES

[1] J. Bates. The Role of Emotion in Believable Agents. *Communications of the ACM* 37, July 1994, 122-125.

[2] D. E. Berlyne (Ed.). *Studies in a new experimental aesthetics: steps toward an objective psychology of aesthetic appreciation*. Hemisphere, Washington DC, 1974.

[3] D. E. Berlyne. *Aesthetics and psychobiology*. Appleton-Century-Crofts, New York, 1971.

[4] A. Camurri and A. Coglio. An Architecture for Emotional Agents. *IEEE Multimedia*, October-December 1998, 24-33

[5] N. Collins. SCMIR: A SuperCollider Music Information Retrieval Library. *Proceedings of the International Computer Music Conference*, Huddersfield, 2011.

[6] J. Dewey. *Art as Experience*. Capricorn Books, New York, 1934.

[7] T. Eerola and J. K. Vuoskoski. A Review of Music and Emotion Studies: Approaches, Emotion Models, and Stimuli. *Music Perception* 30, 2013, 307-340.

[8] T. Eerola. Modeling Listeners' Emotional Response to Music. *Topics in Cognitive Science* 4, April 2012, 607-624

[9] P. Ekman. *Emotions Revealed*. Times Books, New York, 2003.

[10] P. Ekman. An argument for basic emotions. *Cognition and Emotion* 6, 1992, 169-200.

[11] R. Fry. *Vision and Design*, Chatto & Windus, London, 1920.

[12] E. H. Gombrich and R. Saw. Symposium: Art and the Language of the Emotions. *Proceedings of the Aristotelian Society* 36 (1962), 215-246

[13] T. Gonsalves (2009) *The Chameleon Project*, [online], available at: <http://www.tinagonsalves.com/chamselectframe02.htm>, [Last accessed 9- 12-2010]

[14] J. Hochenbaum, A. Kapur, and M. Wright. Multimodal Musician Recognition. *Proceedings of New Interfaces for Musical Expression*, Sydney. 2010.

[15] Y. E. Kim, E. M. Schmidt, et al.. Music emotion recognition: A state of the art review. In *Proceedings of ISMIR*, Utrecht. 2010.

[16] B. R. Knapp and P. R. Cook. The integral music controller: introducing a direct emotional interface to gestural control of sound synthesis. *Proceedings of the International Computer Music Conference*, Barcelona, 2005.

[17] S.G. Koolagudi and K.S Rao. Emotion recognition from speech: a review. *International Journal of Speech Technology* 15, 2 (2012) 99-117.

[18] C. L. Krumhansl. An exploratory study of musical emotions and psychophysiology. *Canadian journal of experimental psychology* 51, 1997, 336-353.

[19] L. Lu, D. Liu, and H-J. Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing* 14, January 2006, 5-18.

[20] C. Moeller. *A Time and a Place: Cheese*, [online], 2003, available at: <http://www.christian-moeller.com/display.php?project_id=36>

[21] R. Nakatsu, J. Nicholson, and N. Tosa. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. *Knowledge based systems* 9, London, Springer-Verlag, 2000, 290-296.

[22] G. Ouzounian. The Biomuse Trio in Conversation: AN INTERVIEW WITH R. BENJAMIN KNAPP AND ERIC LYON, *econtact* 14.2, 2012, [online], available at: <http://cec.sonus.ca/econtact/14_2/ouzounian_biomuse.html>

[23] I. J. Roseman and C. A. Smith. Appraisal theory; Overview, assumptions, varieties, controversies. In Scherer, K. R., Schor, A. & Johnstone, T. (Eds.), *Appraisal process in emotion: Theory, methods, research*. Oxford University Press, New York, 2001, 3-19.

[24] J. A. Russell. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 1980, 1161-1178.

[25] K.R. Scherer and M. Zentner. Music-evoked emotions are different – more often aesthetic than utilitarian. *Behavioral and Brain Sciences* 31, 2008, 595-596.

[26] P.L. Silvia. *Exploring the psychology of interest*. Oxford University Press, New York, NY, 2006.

[27] P.J. Silvia. Emotional responses to art: From collation and arousal to cognition and emotion. *Review of General Psychology* 9, 2005, 342.

[28] M. Thorogood and P. Pasquier. Impress: A Machine Learning Approach to Soundscape Affect Classification for a Music Performance Environment. *Proceedings of NIME*, Daejeon, South Korea, 2013.

[29] N. Tosa, H. Hashimoto, et al. Network-Based Neuro-Baby with Robotic Hand. *Proc. of IJCAI'95 Workshop on Entertainment and AI/Alife*, Montreal, Canada, 1995.

[30] A. van 't Klooster, *Balancing Art and Technology*. PhD Thesis, University of Sunderland, 2011.

[31] E.A.Vessel, G. B. Starr and N. Rubin. Art reaches within: aesthetic experience, the self and the default mode network. *Frontiers in Neuroscience*, 30 December 2013

[32] V. Vermeulen. The EMO-Synth, an emotion-driven music generator. *eContact* 14, February 2012, [online], available at:<http://cec.sonus.ca/econtact/14_2/vermeulen_emosynth.html>

[33] R. M. Winters, I. Hattwick, M. M. Wanderley. Emotional Data In Music Performance: Two Audio Environments for the Emotional Imaging Composer. *Proceedings of the 3rd International Conference on Music & Emotion*, Jyväskylä, Finland, 2013.

[34] M. Zentner, D. Grandjean and K. R. Scherer. Emotions Evoked by the Sound of Music: Characterization, Classification, and Measurement. *Emotion* 8, 2008, 494-52