An Emotion-aware Interactive Concert System: A Case Study in Realtime Physiological Sensor Analysis

Adinda van 't Klooster and Nick Collins

Abstract

An artistic concert system called *BioCombat* devised by the co-authors puts two performers' physiological states head to head, making their attempt to feel particular emotions the object of a live competition. Each performer provides sonic materials for each of eight emotional states; generative visuals are projected in concert, reactive to audio and physiological features, alongside an emotional scoring system also visible to the audience. This article describes the physiological sensing and emotion classification machinery required to support such a challenging real-time task, with both quantitative and qualitative evaluation results. The system draws upon recent research in the field. The designers offer solutions to the problems, and discuss the potential, of such current generation technology for artists.

Keywords: emotion classification, physiological sensing, interactive music system, live concerts, audio-visuals

1. Introduction

This article describes a live interactive performance system that uses physiological sensing technology and machine learning to attempt realtime emotion classification; it is a digital art projects that productively engages with the cutting edge of scientific research. The live concert setting in which this system was deployed supplied a strong challenge in building a real-world application outside the laboratory. Artistic practice here provided an alternative way of evaluating technology's current abilities, parallel to more typical engineering led studies. We reveal practical implementation details we hope of great benefit to future projects from artists and researchers building related systems. A primary artistic goal was to interrogate the inner world of human emotion, whilst acknowledging the limitations of current scientific knowledge and of the technologies available for detecting emotion, especially in live contexts. The cross-disciplinary placement of this research is at the nexus of new musical interfaces informed by physiological sensing and emotion recognition, and fine arts practice. In some places, artistic and pragmatic requirements won out over traditional laboratory-based scientific evaluation, hopefully to the good of the artistic experience, though effort was put into contributing back to the field of emotion research as well.

The heritage of artistic and musical applications of interactive biosensing systems is well covered (van 't Klooster, 2011; Miranda & Wanderley, 2006; Wilson, 2002), and a roll call of artists would stretch from Alvin Lucier (with his famous *Music for Solo Performer* (1965) where amplified neurosignals perturb a network of percussion) to such notaries as David Rosenboom, Atau Tanaka, Tina Gonsalves, Brigitta Zics and George Khut. A tendency in recent systems has been to engage with developments in machine learning to cope with interpretation of the complex data from biosensors (Thorogood & Pasquier, 2013; Vermeulen, 2014), a trend we follow here.

BioCombat is a work by the co-authors for two performers, each of whose emotional state is tracked via physiological sensing, and who compete to better feel certain emotions on demand and thus take command of electroacoustic output. A game setting for the emotion detection machinery provides a play on the functioning and evaluation of physiologically aware systems.

The research backdrop for the emotion recognition is that of affective computing (Picard, 1997), especially emotion modelling and recognition within music information retrieval and music psychology (Kim et al., 2010; Eerola, 2012), and realtime interactive music systems (Collins, 2007; Rowe,

2001). Research on reading emotional state from physiological signals has been maturing; some authors report high percentage success rates on classification tasks distinguishing four, or even eight separate emotional states (van den Broek, Lisy, Westerink, Schut, & Tuinenbreijer, 2009; Kim & André, 2008; Picard, Vyzas, & Healey, 2001). Nonetheless, these references utilise 1-min windows of feature data per calculation in feature extraction and subsequent classification; the concert task reported here requires much more frequent decisions, on the order of once every second based on the previous 5 s. Janssen, van den Broek, and Westerink (2012) introduced a personalised music player driven by physiological state, though they use slowly varying skin temperature to detect longer-term mood rather than more short-term emotion. Affective sensing technology has been brought to concert systems with limited success both in terms of classifier performance and aesthetic experience (Thorogood & Pasquier, 2013; Vermeulen, 2014); live emotion classification for a music system tests the cutting edge of computational emotion recognition. Beyond session data for training, parameter tuning and testing of machine learning classification algorithms within a conventional evaluation framework, generalisation performance is also qualified by a potential for performance nerves to impact on physiological state. Concert conditions here lead to a novel evaluation setting, away from the domain of conventional laboratory study.

We proceed by describing *BioCombat* in more detail as an artwork, before examining the underlying engineering challenge of emotion recognition from physiological signals. We finish the article with discussion of the system as delivered in concert, and some advice for future artists who can gain from the work herein.

2. *BioCombat* as live interactive system

BioCombat is a live interactive system which juxtaposes biofeedback art with performative video gaming. A continuous two-dimensional model of emotion (Russell, 1980) is used in combination with a discrete model, by placing eight distinct emotions in Russell's arousal-valence space. Affective physiological sensing and machine learning are combined to create a 'judging agent' for a contest between two performers. Live scoring is seen on the left projection in Fig. 1; the visuals on the right projection are driven by live physiological input and electro acoustic sound via animations for each of eight emotions: happiness, sadness, anger, calmness, excitement, annoyance, tenderness and fear. The eight emotions were selected to provide a broad cross section of emotional life, across all quadrants of the arousal-valence plane. All eight were seen as achievable states within human experience for the attempt to reproduce feeling them in concert. Two performers take part, and their heart rate, galvanic skin response (GSR) and (single channel) EEG data are tracked live.

As Fig. 1 shows, the screen on the left provides the game status, with instructions and scores visible to the audience. Every minute, the computer demands a new emotion to be felt by the performers, and the music changes. Within the available 60 s, after a 20 s climatisation period for the performers, scoring begins based on who is best at feeling the requested emotion, as indicated by trained classifier models for emotional state; live updates on the score continue every second. Points are gained when the correct arousal (high or low) and valence (positive or negative) are measured, according to quadrants in the standard circumplex two-dimensional model of emotion (Juslin & Sloboda, 2001; Russell, 1980). The status is continually updated until the next emotion is requested, with points accumulating. Each participant provides electroacoustic sound material for each of the eight emotion classes; the current winner is rewarded by their own compositional contribution playing louder than that of their opponents'.

The right screen in Fig. 1 is the visual output of the system, with abstract animations designed for each target emotion; 'scared' is depicted here. The visuals can provide a stimulus to help a performer reach a requested emotional state, though are probably more useful for the audience, who face the screen rather than the performers (who often have their eyes closed to maximise being able to feel the emotions, and minimise muscle side effects in reading EEG data). There are two animated shapes within the right projection: the left one is mapped to the left performer's biosignals, and the right one to the right performer's data. Live features in the sound, such as loudness, rhythm, dissonance and frequency distribution in the energy spectrum, also influence the animations. The graphics are based on drawings by the artist Adinda van 't Klooster. She evaluated the emotional expressiveness of the graphics through an online survey (https://www.affectformations.net/research/visualaffects) and in the concert system used predominantly those graphics that people could most unambiguously map to the target emotions. For this survey, she created four abstract graphics for each of the eight emotions: fifty online participants rated each graphic in terms of its emotional expression using a slider interface and the eight emotion labels. Happy, calm and sad images were most accurately rated in line with their intended expression, whilst tender, annoyed and scary images received much less agreement (van 't Klooster, 2016). The animations generated from these images used motion paths, speed, rotation and size changes to add further dynamism to the graphics. Input from the biosignals also influenced the animations; for example, the tender animation used heart rate and size to let the graphics pulse in time with the participants' heartbeats.

Each target emotion in concert has its own associated pair of soundscapes, one provided by each performer. Adinda van 't Klooster created electronic sound compositions between 1 and 3 mins duration, for each emotion. The tracks were created from recorded sound and manipulated to varying degrees. The main rule used in the composition process was that the emotions had to be expressed via modulations in timbre rather than pitch or rhythm although pitch

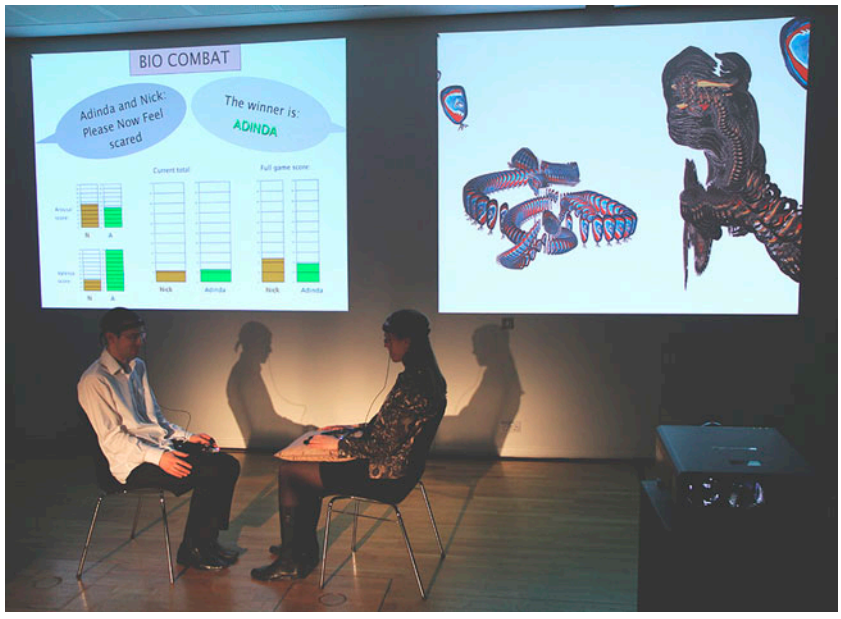

Fig. 1. *BioCombat*: an audiovisual physiological sensor battle to feel emotions competitively; the co-authors compete, image © Adinda van 't Klooster and Nick Collins, 2015, photograph by Simone Tarsitani.

and rhythm still played a subsidiary part in terms of the source material. The individual files can be listened to via the Affect Formations website: https://www.affectformations.net/projects/sound-affects. Some sound files use associative meaning (such as the happy track that combined a contented babbling baby with a street organ) and others, such as for example the sadness track, are very abstract.

Nick Collins created generative SuperCollider patches for each emotion, where frequently used sound synthesis elements include an auditory hair cell model as filter, and feedback loops (the auditory system modelling of these UGens was a coincidence and not meant to reflect the physiological sensing theme, except for a general biological connection). As an example of the generative work, the following is an 'annoyed' synthesis patch. Two auditory hair cell models (HairCell and Meddis) are used alternately and serially as filters, and the patch includes various other distorting components including feedback, to make a very rich and time varying sound:

```
SynthDef(\BioCombatannoyed,{|out = 0 amp = 0.0|
   var a, strength, sound;
   a =
Saw.ar(LFNoise0.kr(LFNoise0.kr(2.0).exprange(0.2,12.7)).exprange(1.0,3.0).round(0.125))*0.2 + (LFNoise0.kr(LFNoise0.kr
   ([0.1,0.15]).exprange(0.7,10.0)).range(0,0.5)*0.99*LocalIn.ar(2));
//rich modulated sound source including feedback (LocalIn)
//serial alternation of hair cell models as compressor/filters
   6.do{|i|
      if(i%2==0) {
         a = HairCell.ar(a,0,5000,5000,LFNoise0.kr([0.07,0.04]).range(0.7,0.9));
      } {
         a = Meddis.ar(a)*3.0;
      };
   };
   //distortion
   a = tanh(LeakDC.ar(LPF.ar(a,LFNoise0.kr([4,7]).exprange(100,10000).lag(0.1))));
   //send feedback
   LocalOut.ar(a);
   //reverberation and further slight distortion
   sound = FreeVerb.ar(a.distort,0.4,0.9,0.1);
   strength = amp.lag(0.05);
   //amplitude controlled low pass filter cutoff
   Out.ar(out,LPF.ar(sound,100 + (20000*strength.squared))*strength)
}).add;
```

Fig. 2. Physiological sensing input, analysis and output.

The relative amplitude of sounds associated with each emotional class depends upon how much better their creator is at feeling a specific target emotion than their game rival. No performers' sounds can disappear entirely, though they can drop by up to 8 dB.

Fig. 2 provides an overview of the technical stages of the *BioCombat* system. The physiological sensors were I-CubeX (http://infusionsystems.com) BioVolt (single channel electroencephalogram (EEG) and electrocardiogram (ECG)) and BioEmo (GSR) sensors plugged into a Wi-microSystem, sampled at 500 Hz via Max/MSP. Subsequent feature extraction and machine learning utilised the SCMIR library in SuperCollider (Collins, 2011), and visuals were created with Processing, with data sent internally between the multiple applications via Open Sound Control messaging (Wright, 2005). Using SCMIR (http://composerprogrammer.com/code.html) had the benefit that it works well for preparing machine listening and learning tasks like signal classification, and once trained up the classifiers are easily deployed in concert live (SuperCollider is an inherently realtime performance-oriented system). The live setting demands calculation upon features aggregated within short-term windows of 5 s, and is substantively different to many previous research investigations utilising minute long windows; fast reaction time is sought here rather than a highly delayed if more stable decision.

The use of multiple programmes and the delicate sensors themselves makes running the piece non-trivial, but not so demanding that it couldn't be organised within the running of a longer concert with other pieces. The performers donned the sensors during a 5-min gap allowed by a different piece from another performer, and the onstage set-up of the programmes took around a minute. Execution order of the various processes was critical, and could possibly be automated via a timed applescript, but was safest run one by one on the day to confirm loading.

3. Live emotion classification

At the core of *BioCombat* are emotion classifiers driven by physiological signals, one classifier trained for each performer, recognising the disparity between human participants of their physiological data and the need for personalisation. In trialling systems for the concert, we investigated classification based on eight emotions ('calm', 'sad', 'annoyed', 'scared', 'angry', 'excited', 'happy' and 'tender'), on the four quadrants within arousal-valence space (thus grouping sad, anger-annoyed-scared, excited-happy, tender-calm), and binary classification for arousal (high versus low, as for example excited against calm) and valence (positive versus negative, e.g. happy as opposed to sad). The specific positions taken for each emotion class in the valence-arousal plane are plotted in Fig. 3.

The system's 'emotional intelligence' was developed through machine learning, with algorithms trained from example sessions with the participants. Machine learning classification attempted to identify a performer's emotional state, first explicitly to one of the eight emotion labels, but with more success less specifically to one of the four quadrants of the arousal-valence space. In creating the training data, initially the performers listened to specific music for each of the eight emotions, that they judged made them feel the particular emotion, but later (and for the evaluation scores reported here) musical cues were taken as the exact music created for particular emotional states for the concert by each of the two performers. In addition the performers could attempt to feel specific emotions by accessing relevant personal memories.

Fig. 3. Mapping of the emotions used in *BioCombat* in the arousal-valence space.

EEG, GSR and heart rate were recorded during listening. Five times, at different points throughout various days, two-minute recordings were created for each emotion. This data, or rather, features derived from the raw physiological signals, was then used to train classifiers for each of the performers. We compared the following machine learning algorithms: neural net, naive Bayes and nearest neighbour. The machine learning inputs could use up to fifteen derived features from the three channels of physiological data. Features were extracted at 43 Hz, max–min normalised with respect to extrema within the whole data-set, then mean aggregated within five-second windows with a step size of a tenth of a second. The full list of 15 features is in Table 1.

Bottom-up feature selection (Guyon & Elisseeff, 2003) was used to find through exhaustive search the best performing subset of features for a given algorithm and participant; one classifier was trained for each candidate feature subset. The feature data was split into training, parameter tuning and final test set for these purposes in the proportions 50/25/25. Classes had an equal number of representatives for the eight emotions in each split, and unequal for the quadrant learning across quadrants due to the unequal distribution of emotional labels to quadrants (though consistently distributed within each split). The data consisted of 5490 feature windows (training examples) for each of the eight emotion classes.

Table 2 compares the performance for the two participants across different classification tasks. For the eight-class problem of predicting the emotional state, results are provided including and excluding the GSR feature, since GSR was so critical to effective performance. The disparate ability of the two participants to evoke target emotions (or in other words, to reproduce physiological signals consistently) is apparent. Algorithms perform well above chance, but always with some misclassifications; participant 1 is harder to model, and likely finds the act of physiological reproduction underlying this task more challenging. Generalisation performance, as measured on the unseen test set, is encouraging, and often comparable with the training and tuning sets. Though different algorithms were investigated, in almost all cases, neural nets achieved much better scores. All neural nets were trained over 1000 epochs. Outputs were encoded with one output per class; class activation was a 1 in the respective output (the n^{th} output for the n^{th} class), and zeroes in other outputs. The near equivalent performance for participant 2 on the four-class versus eight-class problems is notable, and may be due to the unequal aggregation of examples between quadrants. The performance percentages, whilst hardly perfect, hold up well to existing studies in the literature, particularly considering the short-time window decisions. Previous research suggests that arousal is more straightforward to discriminate than valence, though participant 2 performed relatively equivalently between the two dimensions.

To provide a comparison, Table 3 shows performance of a naive Bayes classifier across the two- and four-class tasks; the greater challenge of modelling participant 1 remains clear. Naive Bayes was normally worse than a neural net; the latter was the predominant best performing algorithm as shown in Table 2. The nearest neighbour algorithm typically achieved around chance performance, and its results are not reported further here.

To refine the sense of accuracy per emotion class, Table 4 breaks down success and failure of prediction within an overall confusion matrix for participant 1's valence, and Table 5 for participant 2; Table 6 shows a per emotion count of accuracy for the same scenario, comparing both participant 1 and participant 2. For participant 1, negative valence emotions are harder to predict, and scared is the most confused (operating around chance) whilst tender is most accurately discerned. For participant 1, there was a qualitative sense that fear was the hardest emotion to feel on demand; conversely, the evaluation here also suggests, and experience confirmed, that they could feel tenderness on demand with relative ease. Participant 2 is

Table 1. Features extracted from the three physiological signals.

Feature number	Source signal	Feature
0	GSR	Root mean square amplitude
1	GSR	Running sum of sample by sample absolute amplitude difference
2	GSR	Running sum of 1/23 s absolute amplitude difference
3–5	ECG	Onset detection statistics (heartbeat onset analysis); heartbeats per second, mean and standard deviation of inter-heartbeat interval
6–9	ECG	Beat detection statistics. Measures on the metrical beat histogram: entropy, ratio of the largest to the second largest entries, diversity (Simpson's D measure), metricity (consistency of high energy histogram entries to integer multiples or divisors of strongest entry)
10	EEG	Spectral centroid
11	EEG	Spectral entropy
12–14	EEG	Band-wise energy around centre frequencies 5, 20 and 40 Hz, with half octave bandwidth

Table 2. Training/parameter tuning/test scores across different participants, conditions and machine learning algorithms, with the best scores discovered and associated selected feature subsets.

Participant	Task	Algorithm	Scores (over training/tuning/test sets)	Features selected
1	Arousal left/right quadrant correct (2 classes)	Neural Net (15/15/2 input/ hidden/output units)	70.1/69.8/70.2	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]
1	Valence top/bottom quadrant correct (2)	Neural Net (11/11/2)	69.4/69.5/68.7	[0, 1, 2, 3, 5, 6, 7, 8, 9, 12, 14]
1	Quadrant correct (4)	Neural Net (13/13/4)	53.9/53.9/54.1	[0, 1, 2, 3, 4, 5, 6, 7, 9, 10, 12, 13, 14]
1	Emotion class correct (8), GSR excluded	Naive Bayes	33.9/33.5/34.0	[3, 4, 5, 9, 12, 13, 14]
1	Emotion class correct (8)	Neural Net (12/12/8)	50.1/43.9/46.8	[0, 2, 3, 4, 5, 6, 9, 10, 11, 12, 13, 14];
2	Arousal left/right quadrant correct (2 classes)	Neural Net (13/13/2)	83.0/82.9/83.1	[0, 1, 2, 3, 5, 6, 7, 9, 10, 11, 12, 13, 14]
2	Valence top/bottom quadrant correct (2)	Neural Net (11/11/2)	85.8/85.0/85.9	[0, 2, 3, 4, 5, 7, 8, 11, 12, 13, 14]
2	Quadrant correct (4)	Neural Net (14/14/2)	74.6/73.9/74.2	[0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]
2	Emotion class correct (8), GSR excluded	Neural Net (11/11/8)	61.3/62.2/61.2	[3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14]
2	Emotion class correct (8)	Neural Net (15/15/8)	75.0/75.2/75.2	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]

Table 3. Training/parameter tuning/test scores across different participants and tasks, with feature subsets selected; comparative performance of Naïve Bayes algorithm on arousal, valence and quadrant classification tasks.

Participant	Task	Algorithm	Scores (training/tuning/test)	Features selected
1	Arousal left/right quadrant correct (2 classes)	Naive Bayes	62.7/62.9/63.3	[0, 4, 7, 9]
1	Valence top/bottom quadrant correct (2)	Naive Bayes	55.5/55.4/56.0	[0, 3, 5, 9, 10]
1	Quadrant correct (4)	Naive Bayes	32.1/32.6/32.1	[6,7,9,10, 12]
2	Arousal left/right quadrant correct (2 classes)	Naive Bayes	72.2/72.6/72.6	[2, 3, 4, 5, 6, 12]
2	Valence top/bottom quadrant correct (2)	Naive Bayes	70.4/71.0/71.5	[0, 1, 2, 6, 13]
2	Quadrant correct (4)	Naive Bayes	53.5/54.3/53.9	[0, 1, 5, 6, 9, 12, 13]

Table 4. Confusion matrix for modelling participant 1: positive/negative valence discrimination, broken down in terms of overall success.

Valence	Predicted negative	Predicted positive
True label: negative valence	3684	2044
True label: positive valence	1542	4186

generally much superior, including substantially so for scared, though marginally less for tender, and finds happiness and annoyance toughest to achieve. Across both participants the overall spread is by no means aberrant, and the system makes a generally good attempt across emotions. Similar profiles were observed for other conditions.

The most critical feature, consistently appearing in feature selection experiments, was feature 0, the GSR amplitude. If feature 0 was excluded from feature selection for participant 2's classifier, the best performing subset achieved 61% accuracy on the final test set for eight emotion classes, rather than 75% gained including feature 0; for participant 1 this was 34% versus 47%, respectively. If data was aggregated between participants to make a single classifier, the different baseline GSR levels of the two performers derailed effective classifier performance to around chance: this was a situation where individualised models were absolutely essential to success.

Table 5. Confusion matrix for modelling participant 2: positive/negative valence discrimination, broken down in terms of overall success.

Valence	Predicted negative	Predicted positive
True label: negative valence	5026	718
True label: positive valence	903	4841

Table 6. Correct prediction and confusion matrix information for modelling participant 1 and 2 on positive/negative valence discrimination, broken down per emotion.

Emotion	Negative or positive valence?	Participant 1 predicted negative/positive	Participant 1 percentage Correct	Participant 2 predicted negative/positive	Participant 2 percentage correct
Calm	Positive	419/1013	70.7	72/1364	95
Sad	Negative	1063/369	74.2	1358/78	94.6
Annoyed	Negative	1001/431	69.9	1042/394	72.6
Scared	Negative	703/729	49.1	1267/169	88.2
Angry	Negative	917/515	64.0	1359/77	94.6
Excited	Positive	479/953	66.6	153/1283	89.3
Happy	Positive	422/1010	70.5	411/1025	71.3
Tender	Positive	222/1210	84.5	267/1169	81.4

Training was initially carried out with eight pieces of music chosen by each participant, which they felt would help them evoke well the particular emotional states. However, despite strong training and generalisation test scores, the qualitative performance in practice runs of *BioCombat* left much to be desired; a performer would be convinced they were doing well at feeling a particular emotion, only to discover no points were being scored! After extensive checking of the code, no problems were found in the signal analysis; the only conclusion was that the task of feeling emotional states on cue in the concert situation did not sufficiently match the preparatory circumstances, with the concert audio itself a confound.

The solution was to train the networks with the emotion-specific audio compositions used in the *BioCombat* performance itself, created by each artist/composer to be played for each of the emotions if they were winning. By recording training data against these sounds, which would be present in the concert conditions (albeit to varying degrees), a more quantitatively and qualitatively robust system was established. One might argue that rather than distinguishing between feeling eight different emotions the system is actually distinguishing the patterns in the biosignals triggered by listening to the chosen eight soundscapes. This is unlikely to be the case, however. The difference between listening to the soundscapes versus listening to the soundscape *and* inducing the emotions through remembering life events was tested; it was concluded that accessing personal memories was essential in gaining usable datasets. Potential for muscle memory and EEG to skew the data was minimised; eyes were kept closed most of the time, and this condition led to the most successful system, where performers felt the system judged them fairly in terms of their ability to feel each particular emotion on demand.

In trial rehearsal sessions, the most successful machine learning algorithms for the live situation were selected. It was found after various testing sessions that the discrimination amongst eight emotion classes felt problematic, despite attempts to stabilize results. For the concert performance, the task of distinguishing the four quadrants of arousal-valence space was more robust, and separate algorithms for arousal and for valence worked best. The eight target emotions were placed in Russell's arousal-valence circle (Russell, 1980), as depicted in Fig. 3, and when both arousal and valence were in the right quadrant, the performer gained a point. If only arousal or valence was in the right quadrant, half a point could be scored.

The ultimate system deployed in concert was robust enough to feel 'honest' to the performers, and as far as could be introspected, performance nerves did not skew the final competition situation (no formal evaluation of nerves was carried out). It is interesting to note how different the qualitative performance of the classifier was from the predicted (generalised) success rates suggested by the quantitative machine learning analysis. The situation necessarily demanded personalisation to the two participants, running their own individualised classifiers in concert.

4. *BioCombat* in concert

BioCombat has been performed twice so far: it was trialled at Durham University in February 2015 and thereafter presented in final shape at the Sage Gateshead in March 2015. It was made as part of a larger body of work produced

during a residency of Adinda van 't Klooster at the Durham University Music Department. She worked with various members of staff on a variety of projects that can all be found on the project website <https://www.affectformations.net>. A video of the *BioCombat* performance can be found here <https://www.youtube.com/watch?v=qZmdpDISdUQ&feature=youtu.be> at 34 mins and 18 s in.

Both performers reported finding it hard but not impossible to feel the emotions on demand in concert set-up (perhaps some theatre actors would find this a more familiar task, though the situation is rather different to a conventional theatre production). The time span of one minute per emotion was about right but the first 20 s were not counted in scoring to allow the performers some time to get into the requested emotion. The graphics and audio were only sporadically used by the performers to help feel the emotions, as the use of personal memories was a more powerful way to access emotional states, especially in the concert set-up where the audio was out of a performer's control and depended on the score total. The competitive nature of the performance made it naturally harder to feel the emotions, as keeping an eye on the score would often be counterproductive to feeling the emotions. Preparing the data for the classifiers, especially feeling the negative emotions repetitively, was fairly taxing.

There were some interesting comments from the audience: one person reported wanting to perform the piece herself and another reported finding the performance 'intriguing', and 'making impressive use of the technology' whilst a third person said: 'For *BioCombat*, I felt that when the performers were told what to feel I also started to experience those feelings by wondering how they were forcing themselves to'. Naturally, the only people who can assess the effectiveness of the system are the two performers and they can only assess their own model; they have no access to the true feelings of the other performer. With the success rates of detection of the performers being unequal, there always remains a possibility of the system being unfair, in the sense of biased to score more for one competitor than the other. However, when the system was less robust this level of unfairness was clearly perceived by both performers. The more limited classification task of detecting the four quadrants rather than the eight emotions was required to make it fit for performance and the perceived fairness of the system was thus much increased as well as the joy of performing with it.

We finish this section with a list of advice for future artists exploring this field who might wish to build related interactive systems.

(1) The training examples should reflect as much as possible the final concert situation (time was lost in this project pursuing an earlier alternative physiological data collection set-up based on 'music that the participants thought cued different emotional states' rather than the final concert audio itself).
(2) It is possible to create short-term time window-based sensing with relatively good quantitative performance. Nonetheless, the final arbiter is qualitative, and robustness of an algorithm to concert nerves may override evaluation metrics prepared outside of concert conditions.
(3) The GSR sensors used in this system were found to be unreliable, with both of them breaking and needing to be re-soldered just before the concert. The authors recommend the use of a more robust GSR sensor, though the EEG sensor and heart rate sensor were found to be reasonably stable.
(4) There is no easy technological solution at present, and the coupling of multiple applications may be required. Extremely careful programming and awkward multi-stage concert set-up is necessary.
(5) There is a real excitement, however, for artists to work with such technologies, essential as they are to gain better access to the interior world of human emotional state.

5. Conclusions

Artistic projects provide strong opportunities to test research work in physiological emotion recognition, in an ecologically valid setting. Artistic endeavour away from lab-based study remains exacting, with artists demanding users of such technology. Although medical grade sensors are not necessarily required, the physiological sensing work at the core of the artwork must be shown to work honestly and accurately for artistic integrity. Concert conditions establish an immensely challenging scenario for realtime sensor data acquisition and processing, far beyond offline situations in a requirement for fast reaction to user state.

This article has described a realtime concert system, *BioCombat*, innately requiring participant-specific machine learning following causal short-term window signal processing (such a set-up is different to the luxury of one minute data windows deployed in much previous research). The final system was more trustworthy when set up with a more limited classification task, distinguishing the four quadrants of arousal-valence space, rather than eight distinct emotional states. We described tensions between qualitative and quantitative success less often reported in engineering literature; the gamification of the emotion classification task highlighted the qualitative performance of algorithms as much as their quantitative accuracy over training/tuning/test materials. It is likely that performance nerves and excitement played a powerful part in taking the system outside of its trained circumstances. Future work might prepare a test database with concert data recordings, though the premiere of *BioCombat* could not utilise such data; instead, future systems might boot strap from personalised data collected in concert as the ultimate ecologically valid stimulus (much validation work would remain to be done in order to achieve this convincingly).

Funding

This work was supported by the Leverhulme Trust Artist in Residence grant allowing Adinda van 't Klooster to be hosted at the Durham University Music Department, and by a small grant for the arts from the Arts Council England.

References

Collins, N. (2007). Musical robots and listening machines. In N. Collins & J. d'Escriván (Eds.), *The Cambridge companion to electronic music* (pp. 171–184). Cambridge: Cambridge University Press.

Collins, N. (2011). *SCMIR: A supercollider music information retrieval library*. Proceedings of ICMC2011, International Computer Music Conference, Huddersfield.

Eerola, T. (2012). Modeling listeners' emotional response to music. *Topics in Cognitive Science, 4*, 607–624.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research, 3*, 1157–1182.

Janssen, J. H., van den Broek, E. L., & Westerink, J. H. D. M. (2012). Tune in to your emotions: A robust personalized affective music player. *User Modeling and User-Adapted Interaction, 22*, 255–279.

Juslin, P. N., & Sloboda, J. A. (2001). *Music and emotion: Theory and research*. Oxford: Oxford University Press.

Kim, J., & André, E. (2008). Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*, 2067–2083.

Kim, Y. E., Schmidt, E. M., Migneco, R., Morton, B. G., Richardson, P., Scott, J., … Turnbull, D. (2010). *Music emotion recognition: A state of the art review*. Proceedings of ISMIR, Utrecht.

Miranda, E. R., & Wanderley, M. M. (2006). *New digital musical instruments: Control and interaction beyond the keyboard*. Middleton, WI: A-R Editions.

Picard, R. W. (1997). *Affective computing*. Cambridge, MA: MIT Press.

Picard, R. W., Vyzas, E., & Healey, J. (2001). Toward machine emotional intelligence: Analysis of affective physiological state. *Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23*, 1175–1191.

Rowe, R. (2001). *Machine musicianship*. Cambridge MA: MIT Press.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161–1178.

Thorogood, M., & Pasquier, P. (2013). *Impress: A machine learning approach to soundscape affect classification for a music performance environment*. Proceedings of NIME, Daejeon, South Korea.

van den Broek, E. L., Lisy, V., Westerink, J. H., Schut, M. H., & Tuinenbreijer, K. (2009). *Biosignals as an advanced man-machine interface*. (Research report IS15-IS24).

van 't Klooster, A. R. (2011). *Balancing art and technology: The aesthetic experience in body and technology mediated artworks* (PhD thesis). Sunderland University.

van 't Klooster, A. (2016). *Creating emotion-sensitive interactive artworks: Three case studies*. Leonardo, 7–8. doi:10.1162/LEON_a_01344

Vermeulen, V. (2014). Affective computing, biofeedback and psychophysiology as new ways for interactive music composition and performance. *eContact!, 16*. Retrieved from http://cec.sonus.ca/econtact/16_3/vermeulen_affectivecomputing.html

Wilson, S. (2002). *Information arts: Intersections of science, art and technology*. Cambridge, MA: MIT Press.

Wright, M. (2005). Open sound control: An enabling technology for musical networking. *Organised Sound, 10*, 193–200.