Can You Feel It? Evaluation of Affective Expression in Music Generated by MetaCompose

Marco Scirea[1], Peter Eklund[1], Julian Togelius[2], and Sebastian Risi[1]
[1]IT University of Copenhagen, Denmark
[2]Department of Computer Science and Engineering, New York University, NY, USA
msci@itu.dk,petw@itu.dk,julian@togelius.com,sebr@itu.dk

ABSTRACT

This paper describes an evaluation conducted on the MetaCompose music generator, which is based on evolutionary computation and uses a hybrid evolutionary technique that combines FI-2POP and multi-objective optimization. The main objective of MetaCompose is to create music in real-time that can express different mood-states. The experiment presented here aims to evaluate: (i) if the perceived mood experienced by the participants of a music score matches intended mood the system is trying to express and (ii) if participants can identify transitions in the mood expression that occur mid-piece. Music clips including transitions and with static affective states were produced by MetaCompose and a quantitative user study was performed. Participants were tasked with annotating the perceived mood and moreover were asked to annotate in real-time changes in valence. The data collected confirms the hypothesis that people can recognize changes in music mood and that MetaCompose can express perceptibly different levels of arousal. In regards to valence we observe that, while it is mainly perceived as expected, changes in arousal seems to also influence perceived valence, suggesting that one or more of the music features MetaCompose associates with arousal has some effect on valence as well.

CCS CONCEPTS

•**Applied computing → Sound and music computing;**

KEYWORDS

Music generation, Affective Computing, Quantitative evaluation

ACM Reference format:
Marco Scirea[1], Peter Eklund[1], Julian Togelius[2], and Sebastian Risi[1]. 2017. Can You Feel It? Evaluation of Affective Expression in Music Generated by MetaCompose. In *Proceedings of the Genetic and Evolutionary Computation Conference 2017, Berlin, Germany, July 15–19, 2017 (GECCO '17),* 8 pages.
DOI: 10.475/123_4

1 INTRODUCTION

Computer music generation is an active research field encompassing a wide range of approaches [39]. The motivations for building a computer system that can competently generate music are manifold. Most importantly music has the power to evoke moods and emotions – even music generated algorithmically [23]. In some cases, the main purpose of a music generation algorithm is to evoke a particular mood. This is particularly true for music generators that form part of highly interactive systems, such as those supporting computer games. A common goal of such systems is create music that elicits a particular mood, which suits the dynamic state of the game play. Music generated for computer games can be understood as experience-driven procedural content generation (*EDPCG*) [57], in which music generation adapts to the game, with particular moods or affects expressed in response to player actions.

MetaCompose [50] is a music generator designed to create background music for games in real-time that can express different mood-states. While Scirea *et al.* describe an evaluation of the music-generation technique [50], they do not provide proof of the claimed affective expression, which is one of the main points of interest of MetaCompose.

This paper addresses this by providing a quantitative study based on human annotation of clips of music produced with the generator. Previous evaluations of the same mood expression theory used by MetaCompose seem to suggest that listeners can reliably recognize perceived levels of arousal, but in some cases valence seems to be more ambiguous [48, 49]. The previous version of MetaCompose was only able to play its music in real-time, we expanded the system to make it create pieces (and transitions within the pieces) in real-time, as this is a step forward needed to apply this generator to the intended media of video-games. To better scrutinize the perceived valence we have introduced a real-time annotation task, where the participants report changes in valence in real-time while listening to the piece of music. In is important to underline that there is a difference between perceived and evoked emotion [17], this study focuses on how people perceive the emotional expression of the music produced by MetaCompose, and not if and what kind of emotional response it can arouse in them.

2 BACKGROUND

2.1 Music Generation and Games

Procedural generation of music is a field that has received much attention in the last decade [36].

Wooller *et al.* [55] identifies two categories of procedural music generation, namely *transformational* and *generative* algorithms. MetaCompose [50], falls in the latter category. *Transformational* algorithms act upon an already prepared structure (audio clips, MIDI files, etc.), for example by having music recorded in layers that can be added or removed at a specific time to change the feel of the music. Note that this is only one example and there are a great number of transformational approaches [1, 5], but a complete study of these these is beyond the scope of this paper. *Generative*

DOI: 10.475/123_4

algorithms instead create the musical structure themselves, which leads to a higher degree of complexity in keeping the produced music of consistent quality and coherence, especially when wanting to connect the music to game events. Such an approach requires more computing power, as the musical content has to be created dynamically and on the fly. An example of this approach can be found in the game *Spore*: the music generators were created by Brian Eno with the *Pure Data* programming language [41], in the form of many small samples that assemble to create the soundtrack in real-time.

MetaCompose adopts the latter approach, in particular focusing on generative procedural music generation in games for emotional expression. While the topics of affect [6], semiotics [16] and mood-tagging [31] are also interesting and significant, the focus of this system is *real-time generation of background music able to express moods during game play.*

Many projects focus on expressing one (or more) affective states; an example is described by Robertson [43], where a music generator is developed to express fear. There are parallels between Robertson's work and MetaCompose, for example musical data is represented via an abstraction (in Robertson's case via the CHARM representation [51, 54]), yet Scirea *et al.* [50] claim their system has a higher affective expressiveness since it is designed to express multiple moods in music. A more extensive example of a generative music system targeted at expressing particular emotions is described by Monteith *et al.* [38] using Markov models, *n*-grams and statistical distributions from a training corpus of music. Chan and Ventura's work [10] focuses on expressing moods; yet their approach relies on changing the harmonization of a predefined melody, while MetaCompose generates the complete musical piece.

There are many examples of evolutionary algorithmic approaches to generating music, two notable examples are the methods to evolve piano pieces by Loughran *et al.* [32] and Dahlstedt [12], although many more can be found in the *Evolutionary Computer Music* book [37]. Other examples of real-time music generation can be found in patents. Two examples are a system that allows the user to play a solo over some generative music [42], and another that creates complete concerts in real-time [34]. An interesting parallel between the second system [34] and MetaCompose [50] is the incorporation of a measure of "distance" between music clips in order to reduce repetition. Still, neither of the patented systems present explicit affective expression techniques.

As the final objective, MetaCompose [50] is designed to be employed to create computer game music. It is therefore important to mention the work by Livingstone [31], which defines a dynamic music environment in which music tracks adjust in real-time to the emotions of the game character (or game state). While this work is interesting, it is limited by the use of predefined music tracks for affective expression. Finally, another notable project in affective expressive music in games is *Mezzo* [8], a system that composes neo-Romantic game soundtracks in real-time and creates music that adapts to emotional states of the character, mainly through the manipulation of *leitmotifs*.

2.2 Emotions and moods

Emotions have been extensively studied within psychology, although their nature (and what constitutes the basic set of emotions) varies widely. Numerous models of emotion have been developed since the seminal studies of the early 20th Century [25, 45], arguably one of the most influential is the theory of basic or discrete emotions devised by Ekman [15].The theory of basic emotions hypothesizes that all affective experiences derive from a core set of basic emotions that are distinct and independent.

An alternate approach to the study of emotions has been the development of dimensional models of affect, which assert that all emotions derive from the combination of two or more underlying psychological "dimensions" [40, 46].Lazarus argues that "emotion is often associated and considered reciprocally influential with mood, temperament, personality, disposition, and motivation" [27]. Therefore, the approach presented in MetaCompose [50] aims to produce scores with an identifiable mood, and in so doing, induce an emotional response from the listener.

Affect is generally considered to be the experience of feeling or emotion. Brewin states that affect is post-cognitive [7]; namely emotion arises only after an amount of cognitive processing has been accomplished. With this assumption in mind, every affective reaction (e.g., pleasure, displeasure, liking, disliking) results from "a prior cognitive process that makes a variety of content discriminations and identifies features, examines them to find value, and weighs them according to their contributions" [7]. Another view is that affect can be both pre- and post-cognitive, notably [28]; here responses are created by an initial emotional response that then leads to an induced affect.

Mood is an affective state. However, while an emotion generally has a specific object of focus, mood tends to be more unfocused and diffuse [33]. Batson writes that mood "involves tone and intensity and a structured set of beliefs about general expectations of a future experience of pleasure or pain, or of positive or negative affect in the future"[3]. Another important difference between emotions and moods noted by Beeddie *et al.* [4] is that moods, being diffuse and unfocused, often persist longer than emotions.

2.3 A taxonomy of moods in music

The set of adjectives that describe music mood and its emotional response is immense and there is no accepted standard vocabulary as such. For example, in the work of Katayose [21], the emotional adjectives include *Gloomy, Serious, Pathetic* and *Urbane*.

Russell [44] proposed a model of affect based on two bipolar dimensions: *pleasant-unpleasant* and *arousal-sleepy*, theorizing that each affect adjective can be mapped into a bi-dimensional space (Figure 1). Thayer [52] applied Russell's model to music using the dimensions of *valence* and *stress*; although the names of the dimensions are different from Russell's, their meaning is identical. Also, we find different terms among different authors [46, 56] for the same moods. Scirea *et al.* [50] use the terms *valence* and *arousal*, most commonly used in affective computing research. This way, affect in music can be divided into quadrants based on the dimensions of valence and arousal: *Anxious/Frantic* (Low Valence, High Arousal), *Depression* (Low Valence, Low Arousal), *Contentment* (High Valence, Low Arousal) and *Exuberance* (High Valence, High Arousal).

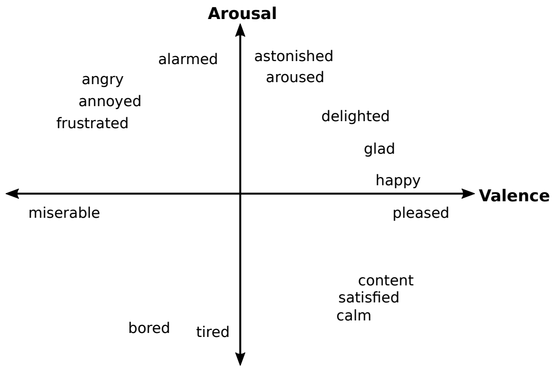

Figure 1: The Valence-Arousal space, labeled by Russell's [44] direct circular projection of affect-adjectives.

These quadrants have the advantage of being explicit and discriminate; they are also the basic music-induced emotions described in [24, 29].

In their previous work, Scirea *et al.* [50] designed their system MetaCompose on these theories to evaluate affective expression in music through a crowd-sourced quantitative experiment: participants were asked to evaluate the affective expression perceived in the music proposed through free-form answers [49]. Subsequently the words were stemmed (to group all the variations of similar words) and positioned in the bi-dimensional affective space through a best-localized criteria: the closer the words describing a part of the space are clustered, the more descriptive they are considered to be of that space.

3 MOOD EXPRESSION THEORY

Scirea *et al.* previously described their model for mood expression [47, 49]. It's important to note that this is a tentative theory used as a starting point, and this study aims at finding out how effective it is. In this section we present a summary of this theory for the purpose of better understanding how the MetaCompose composer works.

Four features that influence perceived mood in music are presented: *volume*, *timbre*, *rhythm*, and *dissonances*. Scirea *et al.* state how these are mainly inspired by Liu *et al.*'s work [30]. While Liu *et al.*'s research focused on mood classification via machine learning, so their approach is applied and expanded to generate music instead. *Volume* is defined by how strong the volume of the music is. It is an arousal-dependent feature: high arousal corresponds to high volume; low arousal to low volume. Intuitively, high volume music results in increased stress. In a similar way, lower volume music, being less intense, is less arousing.

Timbre is defined as the combination of qualities of a sound that distinguish it from other sounds of the same pitch and volume. For example, timbre is what makes the C4 chord sound different when played on a piano compared to a guitar. It is often associated with *"how pleasing a sound is to its listeners"* [2]. One of timbre's most recognizable feature is what is called "brightness", that is, how much of the audio signal is composed of bass frequencies.

In previous literature, MFCC (Mel-Frequency Cepstral Coefficients [26]) and spectral shape features [18] (among other audio features) have been used to classify music on the basis of its timbral feature. Timbre is often associated with valence: the more positive the valence, the brighter the timbre.

Rhythm is divided into three features: *strength*, *regularity* and *tempo* [30]. *Rhythm strength* is defined as how prominent is the rhythmic section is (drums and bass). This feature influences arousal and MetaCompose acts by regulating the volumes of the instrument currently considered the "bass" to be proportionally higher or lower in the general mix. *Regularity* is defined as how regular the rhythm is. This feature influences valence. *Tempo* is defined how fast the rhythm is. This feature influences arousal and is expressed as the beats-per-minute (BPM) that the instruments follow.

As an example, in a high valence/high arousal piece of music, we observe that the rhythm is strong and steady. On the other hand, in a low valence/low arousal piece, the tempo is slow and the rhythm not as easily recognized.

Dissonance is the juxtaposition of two pitches where the frequency ratio between two tones is not close to a simple harmonic ratio. This appears in notes that are very close to each other (but can appear between further apart notes), for example C and C♯. The distance between these two notes is only a semitone, which gives the listener a generally unpleasant sensation. But a dissonant interval does not always have to sound bad. In fact most music contains dissonances, they can be used as cues expressing something amiss. The listener's ear can also be trained to accept dissonances through repetition, which explains why some musical genres rely on dissonant intervals that are otherwise avoided in others.

Meyer [35] observes that the affect-arousing role of dissonances is evident in the practice of composers as well as in the writings of theorists and critics, remarking how the affective response is not only dependent on the presence of dissonances *per se*, but also upon conventional association. This means that depending on the conventions of the musical style, dissonances might be more or less acceptable to the listener, and so can arouse different affective reactions in the listener.

A study of listening preferences of infants, conducted by Trainor and Heinmiller [53], shows that even these young listeners, with no knowledge of musical scale, have an affective preference for consonance. This feature is connected to valence, hypothesizing that introducing more and more dissonances creates a more negative affect expression.

4 METACOMPOSE

Scirea *et al.*'s MetaCompose [50] consists of three main components: (i) *composition generator*, (ii) *real-time affective music composer*. This section presents a summary of the music generation method employed by MetaCompose, a more complete description can be found in [50].

The *composition generator* (i) creates the basic abstraction of a score that will be used by the *real-time affective music composer* in order to (ii) create the final score according to a specific mood or affective state. In other words, the *composition generator* (i) serves as a composer that only writes the basic outline of a piece, while the *real-time affective music composer* (ii) acts as an ensemble, free to interpret the piece in different ways. The system also has an *archive* which maintains a database of all the previous compositions

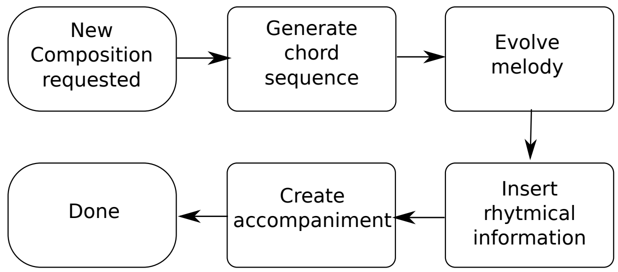

Figure 2: Steps for generating a *composition*.

connected to the respective levels/scenes of the game-state while also allowing a rank to be computed that measures the novelty of future compositions compared to those previously generated. MetaCompose is designed to react to game events depending on the effect desired. Examples of responses to such events include: a simple change in the affective state, a variation of the current composition, or an entirely new composition.

Composition in the context of MetaCompose refers to an abstraction of a music piece composed by a *chord sequence*, a *melody* and an *accompaniment*. It is worth noting that the term *accompaniment* denotes another abstraction (a simple rhythm and an *arpeggio*), not the complete score of a possible accompaniment. The main reason for the deconstruction of compositions is to produce a general structure (an abstraction) that we believe makes music recognizable and provides identity. Generating abstractions, which themselves lack some information that one would include in a classically composed piece of music (e.g. tempo, dynamics, etc) allows MetaCompose to modify the music played in real-time depending on the affective state the interactive media wishes to convey through the mood expression theory. The generation of compositions is a process with multiple steps: (i) creating a chord sequence, (ii) evolving a melody fitting this chord sequence, and (iii) producing an accompaniment for the melody/chord sequence combination (see Figure 2).

Scirea *et al.* [47, 49] define a number of features to include (objectives) and to avoid (constraints) in melodies, these are based on classical music composition guidelines and musical practice. The constraints define that a melody should: i) *not have leaps between notes bigger than a fifth*, ii) *contain at least a minimum amount of leaps of a second* (50% in the current implementation) and iii) *each note pitch should be different than the preceding one.* Three objectives are used to compose the fitness functions: a melody should i) *approach and follow big leaps (larger than a second) in a counter stepwise motion (explained below)*, ii) where the melody presents big leaps the *leap notes should belong to the underlying chord* and finally iii) *the first note played on a chord should be part of the underlying chord.*

When dealing with constrained optimization problems, the approach is usually to introduce penalty functions to act as constraints. Such an approach strongly favors feasible solutions over the infeasible ones, potentially removing infeasible individuals that might lead to an better solutions. There have been many examples of constrained multi-objective optimization algorithms [9, 14, 19, 20]. MetaCompose's approach to melody generation uses a combination of the Feasible/Infeasible two-population method (FI-2POP [22]) and NSGA-II [13] dubbed Non-dominated Sorting Feasible-Infeasible 2 Populations (NSFI-2POP [50]). This approach combines the benefits of maintaining an infeasible population, which is free to explore the solution space without being dominated by the objective fitness function(s), and finding the Pareto optimal solution in the presence of multiple objectives. The algorithm takes the structure of FI-2POP, but the objective function of the feasible function is substituted with the NSGA-II algorithm.

5 EXPERIMENT DESIGN

The main objective of this study is the evaluation of the affective expression in the music produced by MetaCompose. A secondary objective is evaluating in real-time changes in valence in order to better understand what music characteristics influence the listener's perception.

An experiment was designed where participants, while listening to a piece of generated music, would annotate changes in valence via manipulating an annotation wheel. By "annotation wheel" we mean a physical knob that the participants could turn clockwise to indicate an increase in valence and counterclockwise for a decrease. The annotation was conducted using software written by Phil Lopez[1] (and inspired by the work of Clerico *et al.* in annotating fun [11]) with the use of a Griffin Technology PowerMate programmable controller. Afterward participants were tasked with annotating the mood perceived at the start and end of the piece and provide an overall assessment of the music quality.

The questions asked were all in the form of 5-point Likert scales:

- How would you rate the quality of the music you just listened to? *Very low/Somewhat low/Moderate/Somewhat high/Very high*
- How positive/negative was the music at the beginning of the piece? *Very negative/Somewhat negative/Neither negative nor positive/Somewhat positive/Very positive*
- How tense/calm was the music at the beginning of the piece? *Very calm/Somewhat calm/Neither calm nor tense/ Somewhat tense/Very tense*

The last two questions are duplicated for the end of the piece.

A survey was developed with HTML and PHP, using a MySQL database to hold the data collected. The real-time annotation tool is a *C#* program which uses VideoLan's VLC to play the musical clips. The PHP code invokes the annotation tool through the *exec()* function, which effectively stops the execution of the PHP until the annotation terminates.

The experiment was designed to present the participants with 10 randomly chosen music clips (5 static and 5 with a transition, repetitions of the same piece were not allowed). As each clip has length of one minute the experiment was designed to last between 15 and 20 minutes for each participant.

5.1 Music clip generation

For the purpose of this experiment 19 music clips were generated using MetaCompose: 10 that exhibited a transition in affective expression, and 9 that did not. Of the 10 pieces with transitions:

[1] https://github.com/WorshipCookies/RealTimeAnnotation

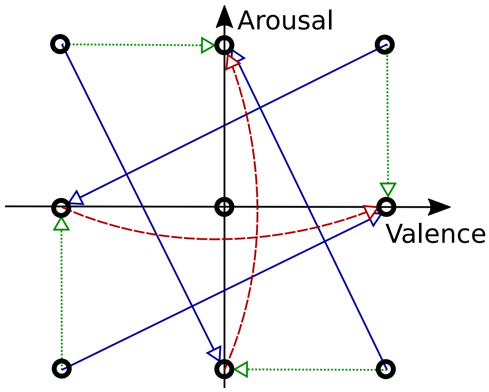

Figure 3: Visual representation of the mood expression of the generated clips: in red/dashed, the 2 large mono-dimensional transitions; in green/dotted, the 4 small mono-dimensional transitions; in blue/solid, the bi-dimensional transitions. Vertices represent the affective expression of the static clips. A list of which clips correspond to each transition can be accessed at http://msci.itu.dk/gecco/clip_list.txt**.**

2 present large changes in only one dimension, 4 present smaller changes in only one dimension, and the remaining 4 present a combination of changes in both valence and arousal (see Figure 3). The music clips are one minute long and the transitions occur half-way through the clip. The pieces themselves are synthesized using Java's MIDI synthesis, the current default method for MetaCompose.

5.2 Experiment setup

Two computers were used for the experiment, with identical setup of software (HTML+PHP survey running locally on an Apache webserver) and hardware (Sony headphones and Griffin Technology PowerMate controllers). The volume of the computer audio was adjusted beforehand to the same level on each PC. All tests were conducted in the meeting rooms at *[University name redacted for blind review]*, which present comparable levels of light and room layout.

6 RESULTS AND ANALYSIS

The data collected corresponds to 200 answers and real-time annotations, from 20 participants. Recall that each participant was presented with a randomized selection of 5 music clips (from a possible 10) containing a transition in expressed mood state and 5 clips with static mood expression (from a possible 9). The clips were also presented in random order.

6.1 Survey analysis

6.1.1 Transition perception. Table 1 shows the differences in the annotations the participants provided for the start and end of the clips. The clips that presented a static mood-state (clips 10-18) present little variation in annotation. In the *transition* group, two clips have been labeled as having almost no perceivable change in expression (clips 2 and 9). Both these clips have no change in arousal (this seems to align with the results to be discussed in Section 6.2). Furthermore, the average variation in valence in these two cases is higher than any of the variations observed in the *static* group, leading us to hypothesize that listeners can indeed perceive variations in affective expression.

Table 1: Variations in arousal and valence from survey. Given the categorical nature of the data we include variation in average and mode. The possible answers are on a 5-point Likert scale (range 0-4), transition data is the difference in how participants annotated the affective expression at the start and end of clips. Clips 0-9 present a transition in affective expression, clips 10-18 are static.

Clip No.	Valence average variation	Arousal average variation	Valence mode variation	Arousal mode variation
0	-0.444	-0.556	-1	0
1	2	-2.833	3	-4
2	0.538	-0.154	0	0
3	0.25	1	2	2
4	1.615	-0.154	2	1
5	0.625	2.125	0	3
6	-1.125	-1.875	-2	-1
7	-1	-1.417	-2	-2
8	1.25	1.75	1	3
9	-0.545	-0.182	0	0
10	0.111	0.333	0	0
11	0	-0.333	0	0
12	0	0	0	0
13	0.231	-0.077	0	0
14	0	0.071	0	0
15	0.111	-0.111	0	-1
16	0.111	-0.222	0	1
17	0	0.214	-1	0
18	0.25	-0.125	0	0

It is important to notice however, that while most perceived transitions reflect what would be expected based on the generator parameters, there are three notable exceptions in annotating valence. In clip 3, a transition to a more positive mood has been annotated, while the clip would have been expected to maintain the same valence; in this case it is noteworthy that while the variation in mode makes it seem like a very strong misclassification (+2), the variation in average scores present a much better score (+0.25). Clip 7 shows a decrease in valence where there would be expected to be none, and clip 8 shows an increase in valence where there would rather be expected to be a small decrease. All of these cases connect to, and find a possible explanation, in the results and discussion that follow in Section 6.2.

6.1.2 Valence analysis. The raw answers given by the participants can be represented in categorical values from 0 to 4 (answers on a Likert scale). Observing the contingency Table 2, it can be observed that there is only a small variation in how clips, that should express neutral and positive valence, are categorized by the participants. Performing a χ^2 test of independence on this data returns a p-value of $2.822e^{-10}$ ($\chi^2 = 61.11, \nu = 8$), so the null hypothesis that the annotations are independent from the expressed valence can be rejected. A series of tests has been conducted on each coupled

Table 2: Valence, raw answers contingency table. Shows how many times an answer was chosen in respect of the intended valence expression.

Intended/Chosen	0	1	2	3	4
Negative	15	61	34	27	2
Neutral	5	22	27	50	16
Positive	5	21	38	58	19

Table 3: Valence contingency table, shows how many times an answer was chosen with what was intended. In this case the answers identifying a negative/positive valence are grouped, no matter the perceived intensity, creating three possible answers: positive, negative and neutral.

Intended\Chosen	Negative	Neutral	Positive
Negative	76	34	29
Neutral	27	27	66
Positive	26	38	77

valence-expressions of this experiment to test the independence of the answers' distributions.

- **Negative vs Neutral** Fisher's exact test: $p = 1.188e^{-08}$. Chi-squared $p = 3.082e^{-08}$ ($\chi^2 = 40.713, \nu = 4$)
- **Neutral vs Positive** Fisher's exact test: $p = 0.9039$. Chi-squared $p = 0.9019$ ($\chi^2 = 1.0517, \nu = 4$)
- **Negative vs Positive** Fisher's exact test: $p = 8.69e^{-11}$. Chi-squared $p = 3.994e^{-10}$ ($\chi^2 = 49.79, \nu = 4$)

Because very small numbers appear in Table 2 χ^2 might not be producing precise estimates of the p-value. To check the correctness of the results a categorization of {Positive, Neutral, Negative} is achieved (Table 3) by grouping the "somewhat positive/negative" and "very positive/negative" answers. Although this removes some of the answers' granularity. Repeating the same tests as before, chi-squared test of independence on this data returns a p-value of $6.419e^{-12}$ ($\chi^2 = 58.358, \nu = 4$). Performing the tests on the coupled data we obtain:

- **Negative vs Neutral** Fisher's exact test: $p = 5.264e^{-09}$. Chi-squared $p = 7.826e^{-09}$ ($\chi^2 = 37.332, \nu = 2$)
- **Neutral vs Positive** Fisher's exact test: $p = 0.5992$. Chi-squared $p = 0.5934$ ($\chi^2 = 1.0437, \nu = 2$)
- **Negative vs Positive** Fisher's exact test: $p = 3.79e^{-11}$. Chi-squared $p = 8.17e^{-11}$ ($\chi^2 = 46.456, \nu = 2$)

While we have a very strong statistical significance between Negative valence and the other two levels, the Neutral and Positive levels appear too similar to consistently distinguished between them.

6.1.3 Arousal analysis. As with valence, a contingency table can be created showing how the participants rated the arousal present in the pieces (Table 4). This time a clear difference between the distributions emerges. Applying the chi-squared test a p-value of $2.2e^{-16}$ ($\chi^2 = 152.11, \nu = 8$) can be calculated, which sustains the hypothesis that the answers are not independent of the expressed arousal. Performing the tests on the coupled arousal-expressions we obtain:

Table 4: Arousal raw answers contingency table. Shows how many times an answer was chosen in respect of the intended arousal expression.

Intended\Chosen	0	1	2	3	4
Low	78	41	11	12	5
Neutral	23	41	54	10	1
High	11	23	31	54	5

Table 5: Arousal contingency table showing how many times an answer was chosen with respect to what we intended. In this case, answers that identify a calm/tense arousal are grouped no matter the perceived intensity, creating three possible answers: high, low and neutral.

Intended\Chosen	Low	Neutral	High
Low	119	11	17
Neutral	64	54	11
High	34	31	59

- **Low vs Neutral** Fisher's exact test: $p = 1.506e^{-13}$. Chi-squared $p = 2.475e^{-12}$ ($\chi^2 = 60.328, \nu = 4$)
- **Neutral vs High** Fisher's exact test: $p = 1.149e^{-10}$ Chi-squared $p = 7.947e^{-10}$ ($\chi^2 = 48.358, \nu = 4$)
- **Low vs High** [2] Chi-squared $p = 2.2e^{-16}$ ($\chi^2 = 90.451, \nu = 4$)

Again, small numbers can be found in Table 4, so the "slightly tense/calm" and "very tense/calm" are combined to obtain Table 5. With Chi-squared a p-value smaller than $2.2e^{-16}$ ($\chi^2 = 125.61, \nu = 4$), consistent with the previous result. Performing the same tests on the coupled data we obtain:

- **Low vs Neutral** Fisher's exact test: $p = 3.386e^{-11}$. Chi-squared $p = 1.47e^{-10}$ ($\chi^2 = 45.281, \nu = 2$)
- **Neutral vs High** Fisher's exact test: $p = 7.894^{-12}$. Chi-squared $p = 3.346^{-11}$ ($\chi^2 = 48.242, \nu = 2$)
- **Low vs High** Fisher's exact test: $p < 2.2e^{-16}$. Chi-squared $p < 2.2e^{-16}$ ($\chi^2 = 78.57, \nu = 2$)

A statistically significant difference of the participants' answer given the three arousal levels can be shown for each of the groups, moreover by looking at the answers distributions we can confirm that the arousal levels are perceived as expected. Still we notice that there seems to be a bias towards low arousal.

6.2 Real-time annotation

The data recorded with the real-time annotation tool consists of a score representing how much higher/lower people are rating the valence of the clip from the origin (the valence at the start of the clip)[3]. As there is no limit to how high/low people could score changes, each raw log is pre-processed with min-max normalization. This way each of the measurements will range between 0-1 and the new data will account for personal perception of changes (e.g. one participant might annotate each change with a double-value scale

[2] Fisher's exact test couldn't be calculated because of a lack of memory

[3] The raw data can be accessed at http://msci.itu.dk/gecco/alllogs.zip.

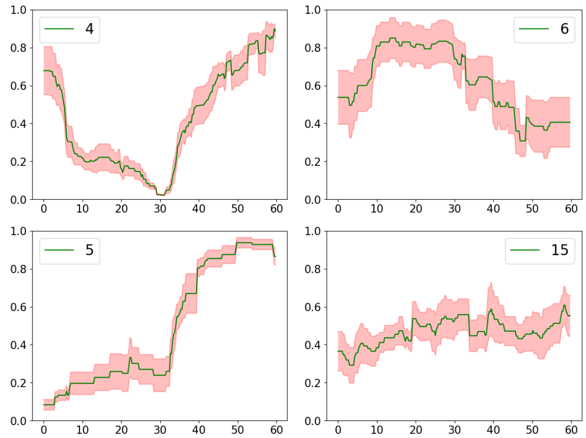

Figure 4: Examples of averaged real-time annotation for valence, the standard deviation is displayed as the red zone, the complete set can be accessed at http://msci.itu.dk/gecco/graphs.zip**. These showcase the main types of annotation that can be observed from the data: e.g., clip 4 presents a correctly annotated clip, an increase in valence halfway through the clip, clip 6 presents a correctly annotated decrease in valence, clip 5 presents an incorrectly annotated increase in valence, and clip 15 correctly shows no transition. The x-axis represents seconds after the start of the clip.**

compared to another participant). Finally, for each clip the average and standard deviation has been calculated to obtain the graphs that can be seen in Figure 4[4].

The first thing that can be noticed is that the clips that present transitions in mood expression present a change in affect trend halfway through the clip (where the change in affective expression happens). Yet in some cases (clips 3, 5 and 8) we observe an increase in valence which should not be there. This increase in participant-observed valence is accompanied by an increase in expressed arousal, which might suggest that one or more of the features that we associate with arousal has an effect on valence as well. Interestingly, clip 7, which should not present any change in arousal, shows a very small negative transition which might correspond with the decrease in expressed arousal. Yet the high standard deviation in observed data present throughout the piece, is not as indicative of misclassification of arousal as the previously mentioned clips.

6.3 Demographics

From the 20 users that participated in the experiment, 14 are males, 5 females, and 1 participant did not express gender. The participants' age has an average of 27.2 years ($stdev \approx 6.3$). In regards to the other demographic answers, expressed in 5-point Likert scale (0–4), most people self-reported very little experience with playing an instrument ($avg = 1.2$, $stdev \approx 1.2$, $mode = 0$), very little knowledge of music theory ($avg = 1.1$, $stdev \approx 1.1$, $mode = 0$), and a considerable experience with video-games ($avg = 2.5$, $stdev \approx 1.27$, $mode = 3$). No matter how we divide the population the results are not significantly different, possibly because of the limited number of participants.

[4]the complete set can be accessed at http://msci.itu.dk/gecco/graphs.zip

7 CONCLUSIONS

This paper describes a study to evaluate the affective expression of the music generated by MetaCompose, based on the human annotation of the clips' produced by MetaCompose.

The main question of the paper is: *can MetaCompose reliably express mood states?* In response, the paper describes an experimental evaluation in which we create music clips from MetaCompose (either containing a transition in affective state or not), and asked participants to annotate the pieces, both with in real-time and after a complete first listening.

Analysis of the data supports the hypothesis that transitions in affective expression intended in the compositions produced by MetaCompose can be recognized by the listeners, and moreover that the sampled levels of arousal are correctly detected with a strong statistical significance. Valence expression seems less well-defined: (i) from the survey answers we see no strong difference between the annotations provided for *Neutral* and *Positive* pieces, (ii) from the analysis of transition perception we observe some incorrect annotations, and (iii) in the real-time annotation some incorrectly perceived changes can be noticed in affect static clips.

To explain point (i) we hypothesize that the fault lies in the introduction of dissonances: MetaCompose seems to only start to include dissonances when expressing negative valence. This means that dissonance-wise there is no difference between *Positive* and *Neutral* valence levels. Points (ii) and (iii) however seem to uncover a more systematic flaw in the expression theory used by Scirea *et al.*: it seems that one (or more) of the features that they associate with arousal have also an effect on valence, as we can observe perceived increases/decreases in valence in response to relative changes in expressed arousal. We need to acknowledge that our sample size is not very large, yet considering the very strong statistical significance of the results we obtained on arousal it seems likely that MetaCompose does indeed present some deficits in valence expression. A more systematic analysis of each music feature would be recommended to amend the mood expression theory to reliably express valence.

In summary, we show how MetaCompose expresses, in a reliably and perceivable way, affect arousal in the music clips it generates. However, there are emergent issues in affect valence expressions, very likely due to some interplay between the musical features associated with arousal and the ones associated with valence.

8 ACKNOWLEDGEMENTS

We would like to offer our special thanks to Professor Georgios Yannakakis and Phil Lopez for the discussions that led to the design of this experiment, and for putting to our disposal the real time annotation tool used in this study.

REFERENCES

[1] Steven Abrams, Daniel V Oppenheim, Don Pazel, James Wright, and others. 1999. Higher-level composition control in music sketcher: Modifiers and smart harmony. In *Proceedings of the ICMC*. Citeseer.

[2] J-J Aucouturier, François Pachet, and Mark Sandler. 2005. " The way it Sounds": timbre models for analysis and retrieval of music signals. *Multimedia, IEEE Transactions on* 7, 6 (2005), 1028–1035.
[3] C Daniel Batson, Laura L Shaw, and Kathryn C Oleson. 1992. Differentiating affect, mood, and emotion: Toward functionally based conceptual distinctions. (1992).
[4] Christopher Beedie, Peter Terry, and Andrew Lane. 2005. Distinctions between emotion and mood. *Cognition & Emotion* 19, 6 (2005), 847–878.
[5] John Biles. 1994. GenJam: A genetic algorithm for generating jazz solos. In *Proceedings of the International Computer Music Conference*. INTERNATIONAL COMPUTER MUSIC ACCOCIATION, 131–131.
[6] David Birchfield. 2003. Generative model for the creation of musical emotion, meaning, and form. In *Proceedings of the 2003 ACM SIGMM Workshop on Experiential Telepresence*. 99–104.
[7] Chris R Brewin. 1989. Cognitive change processes in psychotherapy. *Psychological review* 96, 3 (1989), 379.
[8] Daniel Brown. 2012. Mezzo: An Adaptive, Real-Time Composition Program for Game Soundtracks. In *Proceedings of the AIIDE 2012 Workshop on Musical Metacreation*. 68–72.
[9] Deepti Chafekar, Jiang Xuan, and Khaled Rasheed. 2003. Constrained multi-objective optimization using steady state genetic algorithms. In *Genetic and Evolutionary ComputationfiGECCO 2003*. Springer, 813–824.
[10] Heather Chan and Dan A Ventura. 2008. Automatic composition of themed mood pieces. (2008).
[11] Andrea Clerico, Cindy Chamberland, Mark Parent, Pierre-Emmanuel Michon, Sebastien Tremblay, Tiago Falk, Jean-Christophe Gagnon, and Philip Jackson. 2016. Biometrics and classifier fusion to predict the fun-factor in video gaming. In *IEEE Conference on Computational Intelligence and Games*. IEEE, 233–240.
[12] Palle Dahlstedt. 2007. Autonomous evolution of complete piano pieces and performances. In *Proceedings of Music AL Workshop*. Citeseer.
[13] Kalyanmoy Deb. 2001. *Multi-objective optimization using evolutionary algorithms*. Vol. 16. John Wiley & Sons.
[14] Kalyanmoy Deb, Amrit Pratap, and T Meyarivan. 2001. Constrained test problems for multi-objective evolutionary optimization. In *Evolutionary Multi-Criterion Optimization*. Springer, 284–298.
[15] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.
[16] Mirjam Eladhari, Rik Nieuwdorp, and Mikael Fridenfalk. 2006. The soundtrack of your mind: mind music-adaptive audio for game characters. In *Proceedings of Advances in Computer Entertainment Technology*.
[17] Alf Gabrielsson. 2001. Emotion perceived and emotion felt: Same or different? *Musicae Scientiae* 5, 1_suppl (2001), 123–147.
[18] John M Grey and John W Gordon. 1978. Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America* 63, 5 (1978), 1493–1500.
[19] Amitay Isaacs, Tapabrata Ray, and Warren Smith. 2008. Blessings of maintaining infeasible solutions for constrained multi-objective optimization problems. In *IEEE Congress on Evolutionary Computation*. IEEE, 2780–2787.
[20] Fernando Jimenez, Antonio F Gómez-Skarmeta, Gracia Sánchez, and Kalyanmoy Deb. 2002. An evolutionary algorithm for constrained multi-objective optimization. In *Proceedings of the Congress on Evolutionary Computation*. IEEE, 1133–1138.
[21] H Katayose, M Imai, and S Inokuchi. 1988. Sentiment extraction in music. In *Proceedings of the 9th International Conference on Pattern Recognition*. 1083–1087.
[22] Steven Orla Kimbrough, Gary J Koehler, Ming Lu, and David Harlan Wood. 2008. On a Feasible–Infeasible Two-Population (FI-2Pop) genetic algorithm for constrained optimization: Distance tracing and no free lunch. *Eur. J. Operational Research* 190, 2 (2008), 310–327.
[23] Vladimir J Konečni. 2008. Does music induce emotion? A theoretical and methodological analysis. *Psychology of Aesthetics, Creativity, and the Arts* 2, 2 (2008), 115.
[24] Gunter Kreutz, Ulrich Ott, Daniel Teichmann, Patrick Osawa, and Dieter Vaitl. 2008. Using music to induce emotions: Influences of musical preference and absorption. *Psychology of Music* 36, 1 (2008), 101–126.
[25] Carl Georg Lange and William James. 1922. *The emotions*. Vol. 1. Williams & Wilkins.
[26] Thibault Langlois and Gonçalo Marques. 2009. A Music Classification Method based on Timbral Features.. In *ISMIR*. 81–86.
[27] Richard S Lazarus. 1991. *Emotion and Adaptation*. Oxford University Press.
[28] Jennifer S Lerner and Dacher Keltner. 2000. Beyond valence: Toward a model of emotion-specific influences on judgement and choice. *Cognition & Emotion* 14, 4 (2000), 473–493.
[29] Erik Lindström, Patrik N Juslin, Roberto Bresin, and Aaron Williamon. 2003. "Expressivity comes from within your soul": A questionnaire study of music students' perspectives on expressivity. *Research Studies in Music Education* 20, 1 (2003), 23–47.
[30] Dan Liu, Lie Lu, and Hong-Jiang Zhang. 2003. Automatic mood detection from acoustic music data. In *Proceedings of the International Symposium on Music Information Retrieval*. 81–7.
[31] Steven R Livingstone and Andrew R Brown. 2005. Dynamic response: Real-time adaptation for music emotion. In *Proceedings of the 2nd Australasian Conference on Interactive Entertainment*. 105–111.
[32] Roisin Loughran, James McDermott, and Michael O'Neill. 2015. Tonality driven piano compositions with grammatical evolution. In *IEEE Congress on Evolutionary Computation (CEC)*. IEEE, 2168–2175.
[33] Brett AS Martin. 2003. The influence of gender on mood effects in advertising. *Psychology & Marketing* 20, 3 (2003), 249–273.
[34] Sidney K Meier and Jeffrey L Briggs. 1996. System for real-time music composition and synthesis. (March 5 1996). US Patent 5,496,962.
[35] Leonard B Meyer. 2008. *Emotion and meaning in music*. University of Chicago Press.
[36] Eduardo Reck Miranda. 2013. *Readings in music and artificial intelligence*. Routledge.
[37] Eduardo Reck Miranda and Al Biles. 2007. *Evolutionary computer music*. Springer.
[38] Kristine Monteith, Tony Martinez, and Dan Ventura. 2010. Automatic generation of music for inducing emotive response. In *Proceedings of the International Conference on Computational Creativity*. 140–149.
[39] George Papadopoulos and Geraint Wiggins. 1999. AI methods for algorithmic composition: A survey, a critical view and future prospects. In *AISB Symposium on Musical Creativity*. Edinburgh, UK, 110–117.
[40] Jonathan Posner, James A Russell, and Bradley S Peterson. 2005. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology* 17, 03 (2005), 715–734.
[41] Miller Puckette and others. 1996. Pure Data: another integrated computer music environment. *Proceedings of the Second Intercollege Computer Music Concerts* (1996), 37–41.
[42] Alexander P Rigopulos and Eran B Egozy. 1997. Real-time music creation system. (May 6 1997). US Patent 5,627,335.
[43] Judy Robertson, Andrew de Quincey, Tom Stapleford, and Geraint Wiggins. 1998. Real-time music generation for a virtual environment. In *Proceedings of ECAI-98 Workshop on AI/Alife and Entertainment*. Citeseer.
[44] James A Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178.
[45] Klaus R Scherer, Angela Schorr, and Tom Johnstone. 2001. *Appraisal processes in emotion: Theory, methods, research*. Oxford University Press.
[46] Harold Schlosberg. 1954. Three dimensions of emotion. *Psychological review* 61, 2 (1954), 81.
[47] Marco Scirea. 2013. Mood Dependent Music Generator. In *Proceedings of Advances in Computer Entertainment*. 626–629.
[48] Marco Scirea, Yun-Gyung Cheong, Byung Chull Bae, and Mark Nelson. 2014. Evaluating musical foreshadowing of videogame narrative experiences. In *Proceedings of Audio Mostly 2014*.
[49] Marco Scirea, Mark J Nelson, and Julian Togelius. 2015. Moody Music Generator: Characterising Control Parameters Using Crowdsourcing. In *Evolutionary and Biologically Inspired Music, Sound, Art and Design*. Springer, 200–211.
[50] Marco Scirea, Julian Togelius, Peter Eklund, and Sebastian Risi. 2016. Meta-Compose: A Compositional Evolutionary Music Composer. In *International Conference on Evolutionary and Biologically Inspired Music and Art*. Springer, 202–217.
[51] Alan Smaill, Geraint Wiggins, and Mitch Harris. 1993. Hierarchical music representation for composition and analysis. *Computers and the Humanities* 27, 1 (1993), 7–17.
[52] Robert E Thayer. 1989. *The Biopsychology of Mood and Arousal*. Oxford University Press.
[53] Laurel J Trainor and Becky M Heinmiller. 1998. The development of evaluative responses to music:: Infants prefer to listen to consonance over dissonance. *Infant Behavior and Development* 21, 1 (1998), 77–88.
[54] Geraint Wiggins, Mitch Harris, and Alan Smaill. 1990. *Representing music for analysis and composition*. University of Edinburgh, Department of Artificial Intelligence.
[55] Rene Wooller, Andrew R Brown, Eduardo Miranda, Joachim Diederich, and Rodney Berry. 2005. A framework for comparison of process in algorithmic music systems. In *Generative Arts Practice 2005 — A Creativity & Cognition Symposium*.
[56] Wilhelm Wundt. 1980. *Outlines of psychology*. Springer.
[57] Georgios N. Yannakakis and Julian Togelius. 2011. Experience-driven procedural content generation. *IEEE Transactions on Affective Computing* 2, 3 (2011), 147–161.