Ambient Music Co-player: Generating Affective Video in Response to Impromptu Music Performance

Issei Fujishiro (member)†, Anri Kobayashi †

Abstract  When a musical instrument player performs music, the accompanying visual information can have a significant effect on the performance. In this paper, we present an ambient music co-player (AMP) as a system that generates background videos in response to the impromptu performance of a single musical instrument player. The AMP system evaluates the performance to interpret the real player’s emotional impression and generates an influential video based on the results of the evaluation. The player tends to change their performance while being inspired by the generated video, further triggering the system to modify the video. The AMP system aims to establish an affective loop where the system continues applying stimuli to the performance of the real player. The final goal of this study is to make the system act as a “co-player” of the player and to amplify the quality of the player’s performing experience entirely through interactions between the two. By conducting a user evaluation, it was proven that the AMP system was able to inspire an amateur guitarist as the subject through affective video generation and to make his performance better than when playing alone.

Key words: Affective computing, music visualization, video generation, musical performance, classifier, cellular automata.

1. Introduction

Many known approaches adopted by prior systems cooperate with musical instrument players and make their performance better. Showing the right pitch or timing and judging the quality of the player's performance in terms of harmony and rhythm are typical approaches. In recent years, however, music genres and styles have diversified; hence, the question of what is a good performance is difficult to answer. However, all musical players have hoped for unexperienced expressions, which could possibly be realized by co-acting with other players and by being inspired by them. In our study, we attempt to construct a virtual “co-player” who provides inspiration to the real player in the performance.

When we listen to a musical performance, we may call up an image that is associated with its auditory information. This sometimes allows us to make the musical experience more deeply emotional. Such visual information is imperative not only for listeners of music, but also for instrument players themselves. For example, several players in a jam session may change their style of playing immediately by closely examining the co-players’ expressions and behaviors and predicting their emotions and intentions on the fly. In this paper, we propose a system called ambient music co-player (AMP for short hereafter) that provides an influential background video for a musical instrument player and attempts to inspire the player to make changes to their style of playing.

Figure 1 shows the interaction loop between the AMP system and a player. The player plays music in front of a display monitor and their performance is inputted into the system as audio signals via an audio interface. The system evaluates these signals in an affective way and generates a corresponding video based on the results of the evaluation. Then, the player tends to change their performance while being inspired by the resultant video, further triggering the system to modify the video. The final goal of this study is to establish an affective loop where the system continues amplifying the performance of the real player.

The remainder of this paper is organized as follows. The next section gives an overview of the related work. Section 3 describes the requirements of inputting into the AMP system. The processes of evaluating musical performances and generating videos are presented in Section 4 and Section 5, respectively. Empirical evaluations of the system are discussed in Section 6, and Section 7 concludes this paper with a few remarks on future work.
What differentiates the current version from our prior one can be summarized in the following five points:

1. The number of emotional categories the system deals with was increased from four (2 × 2) to nine (3 × 3), though Russell’s 2D arousal–valence model is still respected (Section 4).

2. The musical feature vector construction of the new version was sophisticated in the temporal direction through introducing the concept of phrase (Section 4.1) and in the spatial direction through augmentation with 12 elements in total (Section 4.2).

3. Accordingly, the SVM (support vector machine) network employed by the current version was enriched with twelve classifiers rather than two classifiers in the previous one (Section 4.3).

4. To make the rules of cell state transition and coloring work more autonomously, the active mode was newly introduced to the current version, instead of the async mode of the previous version (Section 5.4).

5. A comparative user evaluation was conducted to prove empirically the enhanced quality of the AMP system as “co-player” (Section 6.3).

2. Related Work

2.1 Cooperation with Musical Performance

Several systems attempt to improve players’ performances through intervening actively in the performance, including Reidsma et al.2) and Brown3), and through other attempts to assist players in mastering musical instruments, including Rogers et al.4). Our work can be categorized into the former, as its novel point lies in its style of intervention, which is characterized by affective interaction during an impromptu performance.

2.2 Music Analysis

Some researchers have attempted to elucidate relationships between music and the emotions experienced by listeners. For example, Berg et al.5) studied how the feature quantity extracted from audio signals of a music piece relates to the change in listeners’ emotions. Meanwhile, Chiang et al.6) found a way to classify music pieces based on their emotional impressions using machine learning. In our study, we employed Russell’s two-dimensional model of emotion7) to deal with various kinds of emotions and designed a classification algorithm for various music performances based on the emotions that listeners experience, with reference to Chiang et al.5), Lu et al.6), and the like. For the classification purpose, we adopted SVM, as it has been reported that the scheme shows the best performance in classifying music pieces9).

2.3 Video Analysis and Generation

Similar to those between music and emotions, there are many known studies on the relationships between videos and emotions experienced by viewers10)–13). Xu et al.13) developed a system that predicts what kind of emotional impressions a certain film gives an audience by extracting the visual feature quantity from fragments of said film. In contrast, some researchers attempted to affect emotions through video expression. Omata et al.14) developed a system that actually influences viewers’ emotions by showing videos containing such factors related to viewers’ changes in emotions as colors, shapes, and animations. Our work refers to these and attempts to build an algorithm that actually influences players’ emotions through video generation. Moreover, there exist some attempts to design interactive systems that synchronize music and video15)–16). The best-known system of this kind is the visualizer of Windows Media Player17). These systems generate a video in response to the input sound. The AMP system generates a video based on a dedicated cellular automaton model. The reason we adopted the model is that it can be considered the best means to inspire a player through emotional video expressions. Several video generation algorithms of existing interactive systems15)–17) could generate impressive videos, whereas their ways of reacting to sound remain fixed; thus, it is difficult to make a fresh and surprising impression such that the players want to continue playing. On the other hand, the cellular automaton model has a notable ability to generate various spatial patterns, even with simpler rules.
3. Requirements of Input to AMP

By restricting the format of the musical performance to a solo improvisation with an electric guitar, we intended to allow our video generation algorithm to inspire the player more effectively.

3.1 Electric Guitar

In musical performances using an electric guitar, a device called an effector is sometimes used for processing the original sound to distort and to add a reverb sound, for instance. This allows guitarists to have a wide range of expressions with the guitar. By comparing changes in music features extracted from audio signals of such a guitar performance, the AMP system can analyze changes in a performance and listeners' emotional impressions with relative ease. Indeed, Juslin\textsuperscript{18}) attempted to examine the relationship between musical performances with the electric guitar and the emotions experienced by listeners.

3.2 Solo

We simply assume one-to-one interactions between the AMP system and a solo player. It is expected that the AMP system would be more sophisticated to enable the establishment of an affective loop with several players if we could also consider interactions among players and the global atmosphere forged during a jam session.

3.3 Improvisation

We do not deal with the case in which a player is playing existing music pieces. This is because the player must dutifully obey the instructions indicated by the composer; thus, the flexibility of expression is restricted more than that in impromptu performances. The AMP system allows a player to express the inspiration that they feel under the auspices of a generated video.

4. Musical Performance Evaluation

To evaluate musical performances based on their emotional impressions, we respect Russell’s two-dimensional model of emotion\textsuperscript{7}) as our evaluation criterion. This model evaluates emotions using a pair of valuation bases: “arousal,” indicating the emotional intensity, and “valence,” indicating whether an emotion is positive or negative. To enhance the discriminative effect compared with the model in our previous paper\textsuperscript{1)}, we define the nine categories of emotion that correspond exactly to the nine subspaces of the two-dimensional space by dividing each valuation base into three classes: high, medium, and low, as shown in Fig.2. As mentioned earlier, referring to Lu et al.\textsuperscript{8}) and others, we designed a procedure of the performance evaluation for the AMP system. First, the system preprocesses the input signals to make “performance data” (Section 4.1). Then, it extracts music features from each frame in the performance data, arranges them into features that represent the characteristics of the performance data, and makes a feature vector consisting of them (Section 4.2). Finally, the system utilizes SVM classifiers to classify the feature vector into one of the nine categories of emotion in Fig. 2 (Section 4.3). We expanded each of the processes from our previous work\textsuperscript{1)} and will describe them in greater detail in the next subsections, respectively.

4.1 Performance Data

Figure 3 shows the conceptual organization of the performance data. The AMP system evaluates the player’s performance based on audio signals. The system takes as input only audio signals with a sampling rate of 44.1 kHz and bitrate of 16 bit. The system first starts to store signal samples into a buffer that has a limited capacity, and then if it becomes full, the system exports the current content of the buffer into the main memory as a “frame.” When a considerable number of frames is stored in the memory, the system sets these frames as “performance data.” The system handles performance data as the base unit of the performance that can represent a change in its style. On a trial basis, the buffer or frame size was fixed to 1,024 samples.

In the previous version of the AMP system\textsuperscript{1)}, the number of frames in the performance data varied inorganically as the system updated its response to a change in the tempo of the performance. In the new version, to evaluate the performance data from a more musical perspective, it is adjusted dynamically to the length of each “phrase” in the performance, which is a musical term meaning an array of several notes that has a musical function. In a musical performance, various phrases may be performed repeatedly in the same or in a different style. As such, the new version introduced
phrase as a minimum unit that represents changes in emotional impression in the performance. The system executes onset detection of the audio signals to find timings when notes are played. Thereafter, the system extracts music features as shown in Section 4.2 (except for onset and autocorrelation features) and detects change points in their time sequence using a cumulative sum control chart\(^\text{19}\). Finally, the system defines a phrase as a section between two change points.

### 4.2 Music Feature Extraction

Referring to Chiang et al.\(^\text{16}\), Juslin\(^\text{18}\), and others, the AMP system extracts music features related to listeners’ emotional impressions from the performance data. In general, musical features referring to intensity (sound level) are related to the degree of arousal, and those referring to timbre are related to the degree of valence. The system processes performance data, extracts musical features as shown below, and then constructs the feature vector. The feature vector contains extracted musical features as its elements, which are divided into two subsets: arousal and valence subvectors, as shown in Fig. 4. Note that from the previous version of the system\(^\text{1}\), we added several features related to rhythm strength and rhythm regularity to improve the accuracy of the evaluation. The first is related to the degree of arousal while the second is related to the degree of valence.

#### (1) Arousal Features

**Intensity:** The AMP system calculates the intensity of the signals in each frame, and then represents the average and variance of these as features of the performance. Simultaneously, the system calculates the intensity of each of the octave-scale bands (subbands), which are defined as

\[ [0, \frac{\omega_0}{2^n}], \left(\frac{\omega_0}{2^n}, \frac{\omega_0}{2^{n-1}}\right), \ldots, \left(\frac{\omega_0}{2^n}, \frac{\omega_0}{2^1}\right), \text{ (1)} \]

where \( \omega_0 \) denotes the sampling rate and \( n \) the number of subband filters, the latter of which was set to 6 in the system on an empirical basis.

**Onset:** The onset features indicate the onset of notes in the musical performance and are known as the features that represent the strength of the rhythm. To detect onsets, we utilized spectral flux, which is the L2-norm distance of the frame-to-frame power spectrum difference. By finding peaks in the time sequence of the spectral flux values calculated from each of the frames, we can obtain the onset frames in the performance data\(^\text{20}\). To find the peaks, the system applies the weighted median filter to the time sequence of spectral flux values and seeks their local maximum points. The AMP system finds the intensity of the signals in frames at the moment of onset and at intervals between consecutive onset frames (equivalent to the length of the note). Then, the system calculates the average of the intensity values and interval lengths.

#### (2) Valence Features

**Spectral Contrast:** The spectral contrast is often used as a feature that represents the outline of the power spectrum for music analysis. This is considered superior to Mel-Frequency Cepstrum Coefficients, as there is no loss of the relative spectral information in each subband. The system calculates this for each frame and averages the value for each subband feature. To extract spectral contrasts, we define \( x_{k,i} \) as the value of a frequency-bin on the \( k \)-th subband. We herein put indexes \( i \) in descending order, as follows

\[ x_{k,1} > \ldots > x_{k,i} > \ldots > x_{k,N_k}, \quad \text{(2)} \]

where \( N_k \) denotes the number of frequency-bins in each subband. Thus, the strengths of the spectral peaks and the spectral valleys are estimated as

\[
\text{Peak}_k = \log \left\{ \frac{1}{\alpha N_k} \sum_{i=1}^{\alpha N_k} x_{k,i} \right\}, \quad \text{(3)}
\]

\[
\text{Valley}_k = \log \left\{ \frac{1}{\alpha N_k} \sum_{i=1}^{\alpha N_k-1} x_{k,N_k-i+1} \right\}, \quad \text{(4)}
\]

where \( \alpha \) denotes a neighborhood factor. Then, the spectral contrast \( SC_k \) is defined as:
SC_k = Peak_k − Valley_k. \hspace{1cm} (5)

Note that \( \alpha \) was set to 0.4 in the system on an empirical basis.

**Centroid and Roll-Off Frequency:** The AMP system calculates the value of the centroid frequency and 95th percentile frequency, called the roll-off frequency, of the power spectrum in each frame. These values correlate to the degree of deviation of the spectral distribution to high frequency bands, and they represent the brightness of timbre. The system calculates their averages.

**Spectral Flux:** The spectral flux correlates not only to the onsets, but also to the spectrum variations between adjacent frames, and it represents the articulation of a musical instrument sound. The AMP system calculates the spectral flux values for each of the frames and averages them.

**Autocorrelation:** The AMP system uses an autocorrelation of the time sequence of the spectral flux values to examine the rhythm regularity of a phrase. The system extracts autocorrelation values at points of each onset as autocorrelation peaks and minimum autocorrelation values between each pair of adjacent autocorrelation peaks as autocorrelation valleys, and it calculates the averages of the peaks and valleys and a ratio of the two averages.

### 4.3 Performance Data Classification

The system evaluates the performance data using a dedicated classification algorithm that uses SVM to sort the obtained feature vectors. In the previous version of the AMP system\(^1\), the algorithm used two classifiers: the arousal classifier and valence classifier, corresponding to each respective subvector of the feature vector. For the new version, we increased the number of classifiers to 20 to introduce ECOC (error correcting output coding)\(^2\) to the algorithm because it allows SVM-based classification algorithms to deal with multi-class classification and to improve its robustness.

Figure 5 gives an overview of the new performance data classification. Six of the classifiers correspond to the arousal subvector while the others correspond to the valence subvector. Each of the six classifiers outputs a binary value in response to the expected class of the subset, as shown in Table 1. By inputting the feature vector into all of the classifiers simultaneously and arranging their result values with a fixed order, the AMP system generates a binary code that represents a summarized result. Then, the system compares the system-generated code with the predefined codes corresponding to that class value shown in Table 2 and finds the one with the minimum humming distance. Finally, the system determines the category of emotion to which the data belongs by referring to the arousal/valence class corresponding to the predefined code. If the obtained code is unclassifiable, the system cannot determine the category and it outputs an exception that will be ignored by the video generation process.

The performance data used here is generated by using the electric guitar with an effector that distorts input sounds. The effector is often used by a player who wants a powerful and aggressive sound. In general, distorted guitar sounds contain high frequency components largely, and the average intensity is higher than that of clean guitar sounds. This characteristic of the performance data is represented as values for the extracted music features of its feature vector, and then the result binary code “110100 (represents High class)” is obtained from the arousal classifiers and the code “110000 (represents High class)” is obtained from the valence ones. Finally, the system combines the obtained codes and classifies the performance data into Passion, the category indicating high arousal and high valence.

In our study, we used hundreds of audio datasets of the electric guitar performance included in Logic Pro, a digital audio workstation software developed by Apple Inc., as training datasets for the classifiers. These have already been tagged with meta data, such as Cheerful-Dark and Intense-Relaxed. When a cheerful tag is attached, the data has a positive (high) valence value, while a dark tag has a negative (low) valence value, intense a high arousal value, and relaxed a low arousal value. If the data has no tag corresponding to an arousal and/or valence value, the system interprets that it has a medium value for arousal and/or valence.

![Fig. 5](image-url) Overview of performance data classification.
5. Video Generation

By referencing the results of the musical performance evaluation described in the previous section, the AMP system generates an influential background video based on a dedicated cellular automaton model.

5.1 Cell Map

The system embodies the internal states of the video generation with a two-dimensional array of cells, called a “cell map.” Each cell on the cell map decides its next self-state by referring to its neighbors’ states/parameters at every time step. The rule for associating them and determining the cell’s next self-state is referred to as “the rule of cell state transitions.” As all the cells on the cell map transit from one state to another simultaneously, the motions on the cell map appear as a whole. The system uses “the rule of coloring” to associate a cell state with the color of the corresponding pixel, and it generates a snapshot image. In this way, the system generates an animation in which the cell colors change at every moment. If the rule of cell state transitions itself changes, the motion on the cell map.

5.2 Reaction-Diffusion System

To arouse the player’s inspiration, we introduced a reaction-diffusion system to the video generation process. This system is a mathematical model that formulates the following situation: when one or more chemical substances initiate a local chemical reaction or diffusion, it causes the substances to spread into the entire space. Turing showed that some reaction-diffusion systems fulfilling certain conditions can generate a wide variety of spatial patterns called Turing patterns. By applying this system to the rule of cell state transitions, we can produce videos with various styles. To implement this, we adopted the Gray-Scott model, which is simple enough to generate patterns by setting only four constant parameters in the model, but can generate numerous style patterns. The AMP system updates the self-state of cells following the Gray-Scott model. Responding to performance changes, the system also updates the constants in the model to control the spatial patterns appearing on the cell map. Thus, the system generates a wide variety of videos that may exceed the player’s expectations. The system of equations of the Gray-Scott model is formulated as:

\[
\frac{\partial u_{i,j}}{\partial t} = D_u \Delta u_{i,j} - u_{i,j} v_{i,j}^2 + F(1 - u_{i,j}), \tag{6}
\]

\[
\frac{\partial v_{i,j}}{\partial t} = D_v \Delta v_{i,j} - v_{i,j} u_{i,j}^2 + F(1 + k)v_{i,j}, \tag{7}
\]

where \(u_{i,j}\) and \(v_{i,j}\) denote the density of two chemical substances at a certain point of the 2D rectangular space. In the video generation process, they are represented as internal parameters of the cell at \((i, j)\) on the cell map. \(D_u \Delta u_{i,j}\) and \(D_v \Delta v_{i,j}\) are called diffusion terms, to represent influences of the parameters \(u\) and \(v\) of four neighbors of the cell. \(D_u\) and \(D_v\) are diffusion coefficients for \(u\) or \(v\), while \(\Delta u_{i,j}\) and \(\Delta v_{i,j}\) represent second derivatives of the parameters \(u_{i,j}\) and \(v_{i,j}\) and are given respectively as:

\[
\Delta u_{i,j} = \frac{u_{i+1,j} - 2u_{i,j} + u_{i-1,j} + u_{i,j+1} - 2u_{i,j} + u_{i,j-1}}{\Delta x^2}. \tag{8}
\]

\[
\Delta v_{i,j} = \frac{v_{i+1,j} - 2v_{i,j} + v_{i-1,j} + v_{i,j+1} - 2v_{i,j} + v_{i,j-1}}{\Delta y^2}. \tag{9}
\]

where \(-u_{i,j}v_{i,j}^2 + F(1 - u_{i,j})\) and \(u_{i,j}v_{i,j}^2 + F(1 + k)v_{i,j}\) are called reaction terms with \(F\) and \(k\) as control parameters. By setting these parameter values suitably, we can generate a continuous time evolution on the entire cell map.

The AMP system uses the constant parameters \(D_u\), \(D_v\), \(F\), and \(k\) for the rule of cell state transitions. By changing these parameter values, the system controls the spatial patterns appearing on the cell map. The system determines the color of a pixel referring to the \(v\) value of the cell. The system uses this colorization as the rule of coloring. Then, in the same way as the rule of cell state transitions, the system changes the rule of

Table 1 Relationship between expected class and classifier output values.

| No. | Abbreviation | High | Mid | Low |
|-----|-------------|------|-----|-----|
| 1   | H           | 1    | 0   | 0   |
| 2   | HvM         | 1    | 0   | 0 or 1 |
| 3   | M           | 0    | 1   | 0   |
| 4   | MyL         | 0 or 1 | 1 | 0   |
| 5   | L           | 0    | 0   | 1   |
| 6   | LvH         | 0 or 1 | 1 | 1   |

Table 2 Predefined codes corresponding to class values.

| Class | H | HvM | M | MyL | L | LvH |
|-------|---|-----|---|-----|---|-----|
| High  | 1 | 1   | 0 | 0   | 0 | 0   |
| Medium| 0 | 0   | 1 | 1   | 0 | 0   |
| Low   | 0 | 0   | 0 | 1   | 1 | 1   |
| Unclassifiable | 0 | 1 | 1 | 1 | 0 | 1 |

\[
\frac{\partial u_{i,j}}{\partial t} = D_u \Delta u_{i,j} - u_{i,j} v_{i,j}^2 + F(1 - u_{i,j}), \tag{6}
\]

\[
\frac{\partial v_{i,j}}{\partial t} = D_v \Delta v_{i,j} - v_{i,j} u_{i,j}^2 + F(1 + k)v_{i,j}, \tag{7}
\]
5.3 Rules of Cell State Transitions and Coloring

Referring to Xu et al.\cite{13}, Omata et al.\cite{14}, and others, we designed the video generation process by considering several visual features related to the correlation between a video and the emotions experienced by viewers. Similar to the musical features, the visual features can be sorted into two types: arousal and valence features. Through reproducing the features by controlling the rules of cell state transitions/coloring, the system can generate videos that express a particular emotion. In practice, by setting the parameters suitably, we defined dozens of rules of cell state transitions/coloring that span the nine categories of emotion in Fig.2.

(1) Arousal Features

**Motion Intensity**: The motion intensity is the amount of change between consecutive snapshot images. If the motion intensity is higher, the video reflects higher arousal. In the video generation process, if we set the parameters $F$ and $k$ to lower values, the spatial patterns in the cell map change quickly. Thus, we can make a video reflect higher arousal.

**Shot Duration**: In the field of film editing, a “shot” means a fragment of video filmed without any break. In the system, we define a shot as a fragment of video that is generated from the same rules of cell state transitions/coloring until the system switches them to another. If the shot is shorter, the video reflects higher arousal. In other words, by increasing a frequency of rule switching, we can generate a video that reflects higher arousal.

(2) Valence Features

**Brightness and Saturation**: If the brightness and saturation of colors in a snapshot image are higher, the video reflects higher valence. Using the rule of coloring that uses many colors with high brightness and saturation, we can generate a video that reflects higher valence.

5.4 Passive and Active Modes

The video generation process of the previous version\cite{15} has two modes of expression: sync and async. Instead, the new version of the AMP system has two modes of expression: passive and active. Through the implementation of these two modes, we attempt to allow the system to take initiative in the interaction with the player and to inspire the player more actively than the previous one.

In the case of the passive mode, the system decides the emotion to be expressed through video generation in accordance with the resultant emotion of the musical performance evaluation, which is the same as the sync mode in the previous version. In this mode, the system always selects the rules of cell state transitions/coloring that allow the system to express the corresponding emotion with the result of the musical performance evaluation. For example, when the performance is classified into Passion, the system expresses Passion using the rule that makes the motions of the cell map aggressive and that makes the colors in the video brighter and more vivid.

In active mode, the new AMP system generates videos through which the system takes a lead role in playing the performance with an emotional impression corresponding to a certain category. In practice, the system sets the emotion using the following procedure:

(1) From the nine categories of emotion in Fig.2, the system selects the “destination category” to which it leads the player.

(2) The system selects the “leading category,” which is the nearest category from the destination category and a neighbor of the category to which the performance has been classified.

(3) If the category of the performance changes and coincides with the destination category, the system restarts the procedure over again from Step 1. If the category of the performance coincides with the leading category, the system goes back to Step 2.

(4) If it does not coincide with the leading category within a fixed period, the system selects another category from the neighbors of the category of the performance as the leading category. At this time, the leading category is chosen from the neighbors in short-distance order. Then, the system returns to Step 3.

Note that in active mode, the system always selects the rules of cell state transitions/coloring related to the emotion of the leading category. Accordingly, the system generates videos having emotions that are always different from those related to the category of the current performance in an attempt to inspire the player.

The system selects the passive mode basically. However, when the category of the player does not change in a fixed period, the system autonomously selects the active mode. The “fixed period” is determined in proportion to the average length of onset intervals (notes).
in the performance, which has been calculated in the musical performance evaluation.

By selecting either the passive or active mode depending on the situation of the performance, the system either expresses cooperativeness with the player at one stage or behaves differently in an attempt to change the player’s performance style at another stage.

6. Experiments

In this section, we will evaluate the current AMP system in three ways: measuring the accuracy of the performance data classification algorithm, testing with an actual player, and conducting a user study. We implemented the system with openFrameworks, an open source C++ toolkit for creative coding. Our experiment environment consists of a MacBook Pro (Retina, 13-inch, Late 2013) computer, a Zoom TAC-2R audio interface, and a Fender Stratocaster electric guitar.

6.1 Evaluation of Performance Data Classification

(1) Overview

We conducted a test to measure the accuracy of the performance data classification algorithm. We applied a five-fold cross-validation to the arousal/valence classifiers in the algorithm, and the entire algorithm. We used audio data samples of an electric guitar performance included in Logic Pro, a digital audio workstation software developed by Apple Inc.

(2) Results and Discussion

Table 3 shows the numbers of training samples for each classifier and compares all the classifiers in terms of accuracy. The accuracy of arousal (and valence) subset classification was estimated by comparing the proper arousal (or valence) class of the given data and the class predicted by the arousal (or valence) ECOC algorithm. Similarly, the accuracy of the entire algorithm was estimated by comparing the proper category and the category predicted by the entire algorithm. The performance data classification algorithm seems to classify the samples in the datasets fairly well. Note here that we cannot simply compare the result with the results of existing algorithms because there are differences in the kind of musical pieces and aims of the proposed algorithm and existing ones.

Then, the appearance of the video changed, as can be seen in Fig.6(e). Figure 6(f) shows a snapshot when the player played harder after Fig.6(e) was presented. At this moment, the system revised its evaluation of the performance to Low-valence, so the leading category was changed to Anxiety. Through the session, we can observe that changes in the system-generated video reflect changes in the player’s performance. See the accompanying video for more details.
6.3 User Evaluation

To prove the effectiveness of the AMP system empirically, we conducted a user evaluation after having an amateur guitarist (male, Japanese, age: 25) use the system and asked him to complete a questionnaire on whether the system achieved the intended goals. The subject is different from the one in the previous user evaluation (male, Japanese, age: 24), but was carefully chosen to ensure almost the same level of skill.

(1) Overview
First, we had the subject sit down in front of the laptop computer running the system and hold the electric guitar connected with the audio interface to input audio signals of the performance to the system. Then, we had the subject perform impromptu with the guitar for six minutes freely with the instruction of “looking at the screen of the display monitor.” Figure 7 is a snapshot of the subject performing for the evaluation. After the performance was finished, we asked the subject to answer 14 questions to evaluate whether the system achieved the goals, as shown in Table 4, on a five-point evaluation scale: “Strongly agree,” “Partially agree,” “Not sure,” “Partially disagree,” and “Strongly disagree.”

(2) Results and Discussion
As shown in Table 4, Questions 7, 8, and 10 were answered “Partially agree.” Therefore, a loop of expressions was formed between the system and the subject during the session. Questions 12 and 14 were also answered “Partially agree.” From these answers, we found that the interactions with the system influenced the subject favorably. On the other hand, the subject answered “Not sure” to Questions 2 and 9. Moreover, referring to the result of the previous user evaluation, the answer to Questions 3 was changed from “Partially agree” to “Partially disagree.” Based on these answers, we can assume that the system was unable to aid the
subject perfectly in recognizing that the responses of the system are based on affective concepts or humans’ perception of music. Compared with the previous user evaluation\(^1\), we found an improvement in the answer to Questions 5 from “Partially disagree” to “Partially agree.” Thus, we can argue that it was a success to introduce the new active mode because it allowed the system to take initiative in the interaction with the player. The current answers to Questions 1, 3, and 4, however, failed in comparison to the previous ones. We guessed one of the reasons for this result is that the system in the active mode may make the impression that it did not react as the subject intended.

Besides, the answer to Question 6 was also improved from “Not sure” to “Partially agree.” From this, we can suggest that the modifications to the processes in the musical performance evaluation and the video generation allowed the current version of the system to express more emotionally than the previous one.

To summarize the results, we suggest that the AMP system achieved the primary goal of inspiring a player to a certain degree. Meanwhile, it is still unclear whether this was achieved due to the “affective” functions of the system. To clarify this point, we will have to conduct an experiment to compare usability between the AMP system and a system that removes the affective functions from the AMP system. In addition to the qualitative user evaluations, we should conduct quantitative evaluation experiments. For example, by studying the transitions in the subjects’ emotions by measuring and recording physiological observations, such as changes in skin conductivity and the pulsations of subjects, we can validate whether the system is able to establish an affective loop with the player. In addition, we can consider whether the implementation of the passive/active mode might enable the new version of the AMP system to take initiative in the session with the player; however, there is a problem that the system in the active mode can make the player doubt whether it is reacting to the player’s performance. To solve this, we need to adjust the behaviors of the system in the passive/active mode and the way to switch them.

7. Conclusion

In this paper, we presented an improved system that generates videos in response to an impromptu performance of a single musical instrument player. The system respects the architecture adopted in the previous version\(^1\) to interpret the musical expressions of the player but its organization is more sophisticated in terms of emotional categorization, feature vector construction, and the complexity of the SVM network. The introduction of active mode synergetically might work with the sophisticated organization; hence, the partially enhanced discrimination capability was proven empirically through a comparative user evaluation. We need to enhance the method of musical performance evaluation and video generation further by reconsidering which music features can be extracted from performance data to improve the accuracy of the musical performance evaluation and by adopting other models of a reaction-diffusion system to improve the expressiveness of the video generation.

The current system still has a limited mechanism to coact with a player. At present, it has a function that switches its mode between passive or active and decides only whether it should synchronize emotions with the results of the musical performance evaluation. However, concerning real jam sessions, there are many ways for players to coact with one another, and these are more complicated than our approach. We need to devise a means to inspire a player better by considering the practical cases of jam sessions.

We also plan to address the problem with the restricted format of the musical performance. If the system was able to evaluate performances of musical instruments other than the electric guitar, including vocals, the electric bass, and the drums, suitably and to analyze several players’ performances at the same time, then the system could be expected to join a real session of a band and to live up to the expectations of the co-players. By inputting into the system not only the sound of musical instruments but also the cheers of an audience listening to the performance in a concert, the generated video would be entirely synchronized with the players and the audience in the full atmosphere of the live environment. This is indeed the real implication of the word “ambient” in the system name AMP.

Acknowledgments

This work has been supported in part by MEXT KAKENHI under the Grant-in-Aid for Scientific Research on Innovative Areas No.25120014 and JSPS KAKENHI under the Grant-in-Aid for Challenging Research (Pioneering) No. 20K20481.
Table 4 The questions in the user evaluation and a comparison of subjects’ answers to experiments with current/previous versions of the AMP system.

| Questions                                                                 | Current                                      | Previous                                      |
|---------------------------------------------------------------------------|----------------------------------------------|-----------------------------------------------|
| 1: Did you think the system reacted with your performance?                | Partially agree                               | Strongly agree                                 |
| 2: Did you think the system interpreted the emotional parts of your performance? | Not sure                                     | Not sure                                      |
| 3: Did you think the system interpreted your performance in the same way a human does? | Partially disagree                            | Partially agree                                |
| 4: Did you think the system-generated videos correspond to your performance? | Partially agree                               | Strongly agree                                 |
| 5: Did you think the system took the lead in interactions with you through generating the video? | Partially agree                               | Partially disagree                            |
| 6: Did you think the system expressed emotions by generating the video?   | Partially agree                               | Not sure                                      |
| 7: Were you inspired by the system-generated videos?                      | Partially agree                               | Partially agree                                |
| 8: Did you have any actual feelings that you communicated with the system through expressions? | Not sure                                     | Partially disagree                            |
| 9: Did you have any actual feelings that you communicated with the system emotionally? | Not sure                                     | Partially disagree                            |
| 10: Did your interactions with the system affect your performance?        | Partially agree                               | Partially agree                                |
| 11: Did you regard the system as a “co-player” while you were playing?    | Not sure                                     | Not sure                                      |
| 12: Did you enjoy your interactions with the system?                      | Partially agree                               | Partially agree                                |
| 13: Did your interactions with the system make your performance better than when playing alone? | Partially agree                               | Partially agree                                |
| 14: Did you have a better performing experience than when playing alone?  | Partially agree                               | Partially agree                                |

References

1) A. Kobayashi and I. Fujishiro: “An affective video generation system supporting impromptu musical performance,” Proc. 2016 International Conference on Cyberworlds, pp. 17-24, Sept. 2016.

2) D. Reidsma, A. Nijohlt, and P. Bos: “Temporal interaction between an artificial orchestra conductor and human musicians,” Computers in Entertainment—Special Issue: Media Arts (Part II), Vol. 6, No. 4, Article No. 53, Dec. 2008.

3) A. R. Brown: “Visualizing digital media interactions: Providing feedback on jam2jam AV performances,” Proc. 22nd Conference of the Computer-Human Interaction Special Interest Group of Australia on Computer-Human Interaction, pp. 196-199, Nov. 2010.

4) K. Rogers, A. Röhlig, M. Weing, J. Gugenheimer, B. Könings, M. Klepsch, P. Schaub, E. Rukatö, T. Seuffert, and M. Weber: “P.I.A.N.O.: Faster piano learning with interactive projection,” Proc. Ninth ACM International Conference on Interactive Tabletops and Surfaces, pp. 149-158, Nov. 2014.

5) J. Berg and J. Wingstedt: “Relations between selected musical parameters and expressed emotions, extending the potential of computer entertainment,” Proc. 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology, pp. 164-171, June 2005.

6) W.-C. Chiang, J.-S. Wang, and Y.-L. Hsu: “A music emotion recognition algorithm with hierarchical SVM based classifiers,” Proc. 2014 International Symposium on Computer, Consumer and Control, pp. 1249-1252, June 2014.

7) J. A. Russell: “A circumplex model of affect,” Journal of Experimental Psychology: Human Perception and Performance, Vol. 39, No. 6, pp. 1161-1178, Dec. 1980.

8) L. Lu, D. Liu, and H.-J. Zhang: “Automatic mood detection and tracking of music audio signals,” IEEE Trans. on Audio, Speech, and Language Processing, Vol. 14, No. 1, pp. 5-18, Jan. 2006.

9) Y. Yang and H. H. Chen: “Machine recognition of music emotion: A review,” ACM Trans. on Intelligent Systems and Technology, Vol. 3, No. 3, Article No. 40, May 2012.

10) P. Mulhem, M. S. Kankanhalli, J. Yi, and H. Hassan: “Pivot vector space approach for audio-video mixing,” IEEE MultiMedia, Vol. 10, No. 2, pp. 28-40, Mar. 2003.

11) X. Xiang and M. S. Kankanhalli: “Affect-based adaptive presentation of home videos,” Proc. 19th ACM International Conference on Multimedia, pp. 553-562, Nov. 2011.

12) Y.-Y. Xiang and M. S. Kankanhalli: “Video retargeting for aesthetic enhancement,” Proc. 18th ACM International Conference on Multimedia, pp. 919-922, Oct. 2010.

13) M. Xu, J. S. Jin, S. Luo, and L. Duan: “Hierarchical movie affective content analysis based on arousal and valence features,” Proc. 16th ACM International Conference on Multimedia, pp. 677-680, Oct. 2008.

14) M. Omata, K. Moriwaki, X. Mao, and D. Kamaka: “Affective rendering: Visual effect animations for affecting user arousal,” Proc. 2012 International Conference on Multimedia Computing and Systems, pp. 737-742, May 2012.

15) SJQ and kezzardrix: “[SJQ + [SJQ × kezzardrix]] at UrBAN-GUILD (Kyoto), 2012,” June 2012. Available at https://www.youtube.com/watch?v=ZIA3ajh5RN8

16) d.v.d: “d.v.d - CET - Japan - 20071202,” December 2007. Available at https://youtu.be/YMr3fTLZtw

17) Microsoft Inc.: “Windows Media Player.” Available at https://support.microsoft.com/en-us/help/18612/windows-media-player

18) P. N. Juslin: “Cue utilization in communication of emotion in music performance: relating performance to perception,” Journal of Experimental Psychology: Human Perception and Performance, Vol. 26, No. 6, pp. 1797-1812, Dec. 2000.

19) W. A. Taylor: “A Pattern Test for Distinguishing Between Atoregressive and Mean-Shift Data,” Sept. 2016. Available at http://www.variation.com/cpa/tech/pattern.html

20) S. Dixon: “Onset detection revisited,” Proc. 9th International Conference on Digital Audio Effects, pp. 133-137, Sept. 2006.

21) T. G. Dietrich and G. Bakiri: “Solving multiclass learning problems via error-correcting output codes,” Journal of Artificial Intelligence Research, Vol. 2, pp. 263-286, Jan. 1995.

22) A. M. Turing: “The chemical basis of morphogenesis,” Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences, Vol. 237, No. 641, pp. 37-72, Aug. 1952.

23) P. Gray and S. K. Scott: “Autocatalytic reactions in the isothermal, continuous stirred tank reactor: Isolas and other forms of multistability,” Chemical Engineering Science, Vol. 38, No. 1, pp. 29-43, Dec. 1983.

24) P. Gray and S. K. Scott: “Autocatalytic reactions in the isothermal, continuous stirred tank reactor: oscillations and instabilities in the system A + 2B → 3B; B → C,” Chemical Engineering Science, Vol. 39, No. 6, pp. 1087-1097, Dec. 1984.

Issei Fujishiro is a professor at the Department of Information and Computer Science, Faculty of Science and Technology, Keio University. He is also an adjunct professor at the School of Computer Science and Technology, Hangzhou Dianzi University. He received his Doctor of Science in information science from the University of Tokyo in 1988. His research interests include smart ambient media with multi-modal displays. Fujishiro is a member of the Science Council of Japan, a fellow of the IFES, an honorary member of the IEEEI, and a senior member of the IEEE, ACM, and IPSJ.

Anri Kobayashi received his B. E. and M. E. in information and computer science in 2015 and 2017, respectively, both from Keio University. His research interests in school included interactive systems using music and video representations. He is currently working for DMM.com Ltd.