Multimedia Tools and Applications (2021) 80:1793–1812
https://doi.org/10.1007/s11042-020-09636-5

Scene2Wav: a deep convolutional sequence-to-conditional SampleRNN for emotional scene musicalization

Gwenaelle Cunha Sergio[1] · Minho Lee[1]

Received: 30 September 2019 / Revised: 4 July 2020 / Accepted: 13 August 2020 /
Published online: 10 September 2020

Abstract
This paper presents Scene2Wav, a novel deep convolutional model proposed to handle the task of music generation from emotionally annotated video. This is important because when paired with the appropriate audio, the resulting music video is able to enhance the emotional effect it has on viewers. The challenge lies in transforming the video to audio domain and generating music. Our proposed encoder Scene2Wav uses a convolutional sequence encoder to embed dynamic emotional visual features from low-level features in the colour space, namely Hue, Saturation and Value. The decoder Scene2Wav is a proposed conditional SampleRNN which uses that emotional visual feature embedding as condition to generate novel emotional music. The entire model is fine-tuned in an end-to-end training fashion to generate a music signal evoking the intended emotional response from the listener. By taking into consideration the emotional and generative aspect of it, this work is a significant contribution to the field of Human-Computer Interaction. It is also a stepping stone towards the creation of an AI movie and/or drama director, which is able to automatically generate appropriate music for trailers and movies. Based on experimental results, this model can effectively generate music that is preferred to the user when compared to the baseline model and able to evoke correct emotions.

Keywords Sequence-to-conditional SampleRNN · Convolutional neural network · Deep recurrent neural network · Domain transformation · Emotional music generation

1 Introduction

Art may very well be one of the defining characteristics of the human species, and are key to effective human interactions. Its many forms are practiced by almost all human cultures

✉ Minho Lee
mholee@gmail.com

Gwenaelle Cunha Sergio
gwena.cs@gmail.com

[1] School of Electronics Engineering, Kyungpook National University, 80 Daehakro,
Bukgu, Daegu 41566, South Korea

and in all modern societies, visual arts and music are intimately intertwined [17]. A deeper investigation into the relationship between those two modalities of arts [36] opens up a whole new range of possibilities. For instance, it gives an opportunity for visually and/or hearing impaired people to appreciate the field of arts they are unable to perceive. A second application is automatic movie and/or drama directors which can generate appropriate music given muted videos. Both applications are able to evoke stronger emotion from users with the addition of another modality.

Understanding the appeal of emotionally aware systems, researchers in the Affective Computing community have put together efforts in trying to estimate emotion induced by watching videos for various applications [7, 26, 29]. Visual and audio stimuli play important roles in affecting the user's state of emotion [20]. Researchers in visual stimuli, have built a machine with emotional characteristics by considering image features, EEG signals and interaction with subjects [41, 42]. Others have used the IAPS dataset to train a Support Vector Classifier to classify 1D emotion in paintings, successfully demonstrating the potential of machines deriving emotion from images [40], or used it to collect "descriptive emotional category data" with the aim of identifying images that evoke one emotion more than others [16]. Researchers in audio stimuli, such as Shan et al. [27], have proposed a novel emotion-based music recommendation system that ranks film music emotion and features against the available music database. Researchers in [34] have also done research on a variety of hand-crafted music features, and Jaimovich et al. [10] have exposed subjects to a variety of different musical excerpts, both with the goal of understanding how those stimuli affect the user's emotional state.

It is clear now that there is a relationship among visual stimuli, audio stimuli and emotion, and the fact that together they're able to elicit even stronger emotions from the user [1]. Given that, an important direction for this area of research is to find a method that is able to establish a relation between art of a single modality, scene, into art of a different modality, audio. This allows the two modalities to complement one another in the absence of the other.

Thus, the goal of this research, is to develop a conditional model that considers the important aspect of emotion in dynamic scenes, or videos, and music alike to complement one in the absence of the other. In order to achieve that goal, the network must incorporate the properties to model sequences well and to correctly transfer emotion from video to music. We propose *Scene2Wav*, a deep conditional neural network that understands the emotion aspect of a visual scene and, from that, is able to compose music which can elicit similar emotion. An unconditional model is not able to generate music corresponding to a desired video, instead simply generating random music. So it's imperative that our model be conditional so as to generate appropriate music given a specific video. By conditioning the model on emotional visual features, it is able to correctly transfer emotion from one domain to the other. Our proposed model consists of three modules. First, a Convolutional Neural Network (CNN) is tasked with the emotional visual feature extraction stage. This is followed by an Encoder Deep Recurrent Neural Network with Gated Recurrent Units (GRUs) responsible for sequentially encoding the extracted emotional visual features, transforming an emotional scene into a more abstract embedding. Lastly, a conditional Scene2Wav decoder, composed of a proposed conditional SampleRNN, takes the emotional visual vector obtained from the encoder as a condition to generate novel emotional music signals.

A few notes about our work. Researchers and developers can have access to our code,[1] including data pre-processing and proposed model, and replicate it with their own dataset.

[1] Code available at https://github.com/gcunhase/Scene2Wav

Another aspect worth mentioning is that our work manipulates raw audio data. It is motivated by three main reasons: information quality, scalability, and contribution to the raw audio computing community. Most work in this area uses representations such as MIDI or ABC notation because of its lower dimensionality when compared to raw audio. However, this also results in loss of information, which can be heard through less consistent playback quality [22]. We also want our model to be scalable, in future works, for natural sounds and not just piano, which is not possible with lower dimensionality data such as MIDI [22]. Lastly, we believe that our research is crucial to further develop works involving raw audio manipulation, which are still at the early stages given the increased complexity in the task. The gap between raw audio and notation data research has been arguably getting smaller due to the democratization of GPUs through cloud technology, but it's still very expensive to run a model there for extended periods of time. To the best of our knowledge this is the first work to involve deep learning and novel music generation based on emotional scene features.

In summary, our contribution is two-fold:

- *Scene2Wav*: End-to-end deep neural network model able to generate long, rich, raw music signals from encoded emotional visual features.
- *Conditional Music Generator*: Scene2Wav decoder using a proposed conditional SampleRNN that considers emotional visual embedding obtained from a convolutional recurrent neural network encoder.

The remainder of this paper is organized as follows. Section 2 starts by further explaining the related works. Section 3 describes the dataset and pre-preocessing steps. Section 4 covers the proposed method, including visual feature extraction, visual feature encoding, and music generation. Section 5 follows with experimental results and discussion, which includes an extensive performance analysis. Finally, we conclude this paper in Section 6 on conclusion and future works.

2 Related works

Convolutional Neural Network (CNN) [13] and its variations are the most widely used deep learning models in tasks such as feature extraction [3, 39], image classification [12, 18], and object recognition [44]. This model mimics the organization of the human visual cortex system by having a structure of stacked convolutions and non-linear functions, thus being able to effectively extract feature from raw images. On top of excelling in the mentioned tasks, another advantage of using this network is the fact that it takes as input minimally processed images, abolishing the need of hand-crafted features and making it general for a plethora of datasets.

Sequence-to-Sequence (Seq2Seq) is a general term for models that map one sequence to another, with it's main application being in Neural Machine Translation [19, 25]. A typical Seq2Seq model has a encoder-decoder structure, each with a Recurrent Neural Network (RNN) responsible for dynamically modeling a sequence of samples. Considering that our video data involves audio, it's of utmost importance that we consider its dynamic time properties with RNN, which considers information in previous time steps, unlike feedforward networks.

A few applications able to take an image and generate music are currently available: Photosounder [21], Paint2Sound [28], and SonicPhoto [38]. However, a common limitation of all mentioned available applications is that they fail to consider emotion when generating music and the visual features are not extracted with deep learning techniques, but with statistical hand-crafted methods. A closer attempt to an emotion based image musicalization is made by researchers in [43], however they do not consider dynamic information in time since they are musicalizing image and not video. Additionally, they do not delve into machine learning techniques for their model, instead simply using a comparison and matching algorithm and focusing on feature extraction. More recently, we attempt to generate emotional audio from input scene, but that work has the limitation of not generating novel music, the focus of this current work, instead following a comparison and concatenating procedure [6, 24].

Convolutional Sequence-to-Sequence (ConvSeq2Seq) models have been widely used with improved performance in video captioning tasks [35] and domain transformation tasks such as music transcription [32, 37]. SampleRNN [15] is the state-of-the-art, alongside WaveNet [33], for music generation according to the authors human evaluation performance. This model hierarchically combines sample-level modules as multilayer perceptrons and frame-level modules as RNNs in order to capture long-term dependencies in the temporal sequences, and it does so on three different datasets.

The original SampleRNN model has the limitation of being unconditional, meaning that it produces random music when instructed to generate audio, but researchers have recently extended it for applications such as voice conversion and text-to-speech [30, 45]. To alleviate this limitation, we propose a conditional SampleRNN that decodes emotional visual information into audio eliciting the same emotion from users. The conditional SampleRNN is trained on sequentially encoded visual features as initial hidden state and target audio, generating emotionally charged music. The conditional aspect of the model allows for music to be generated according to a given video, characteristic the the unconditional version is unable to. For the encoder part of our model we propose to use a Convolutitonal Sequence Encoder due to its proven effectiveness and performance in multimodal tasks. The previously mentioned ConvSeq2Seq model is then chosen as our baseline model to evaluate the effect our proposed conditional decoder SampleRNN has over a regular decoder RNN.

3 Dataset and pre-processing

One of the challenges in this work is finding an appropriate dataset with all the required information needed for our task. The COGNIMUSE database [46] is a multimodal video database annotated with emotion amongst others. In other words, it is a collection of movie extracts containing synchronized scene, music and emotion scores, making it ideal for our case study. There are 7 fully annotated videos in the COGNIMUSE database, shown in Table 1, where each video is an approximate 30 minute snippet of a full movie. Each video is originally annotated with valence values ranging from -1 to 1 in regular time intervals. In our work, we consider valence scores below 0.00 to be negative (0) and above that threshold to be a positive emotion (1). The dataset is divided in such a way as to guarantee disjoint train and test datasets, presented in the top and bottom half of the table respectively, so as to prevent overfitting of the model.

The pre-processing of the dataset[2] consists of splicing the data (scene, audio, and emotion scores) into chunks of 3 seconds, duration that was kept short due to the high-dimensional characteristic of audio signals. By splicing the data in such manner, every 3

[2] https://github.com/gcunhase/AnnotatedMV-PreProcessing

Table 1 COGNIMUSE annotated videos, divided into train (top half) and test (bottom half)

Video	Movie Name	Duration(min)	Splices
BMI	A Beautiful Mind	31.3	625
CHI	Chicago	30.15	602
FNE	Finding Nemo	30.3	605
GLA	Gladiator	30.05	600
LOR	Lord of the Rings	37.57	750
CRA	Crash	26.63	532
DEP	The Departed	30.48	609
Total	-	215.70	4,323

seconds of data have a visual scene and an audio signal of similar duration and a mutually corresponding emotion score. It's important to mention that splices are shuffled so as to make the model unbiased from the previous sample seen during training.

The original videos are saved at 30 frames per seconds (fps) with size of (640×480), however, due to memory limitations and to reduce complexity, we downsample them to 10 fps and (100×100), meaning that every chunk of 3 seconds of video had 30 frames. Lastly, illustrated in Fig. 1, we convert the frames from RGB to HSV color scheme. The reason for that is that RGB is a device-oriented color space, meant to be used in a digital setting, whereas the HSV color space is a user-oriented non-linear transform of RGB, able to better represent perceptual color relationship to humans [4]. The RGB to HSV is explained mathematically in (1) to (7). First, consider (1) to (3):

$$M = \max(R, G, B) \tag{1}$$

$$m = \min(R, G, B) \tag{2}$$

$$\delta = \text{range}(R, G, B) = M - m \tag{3}$$

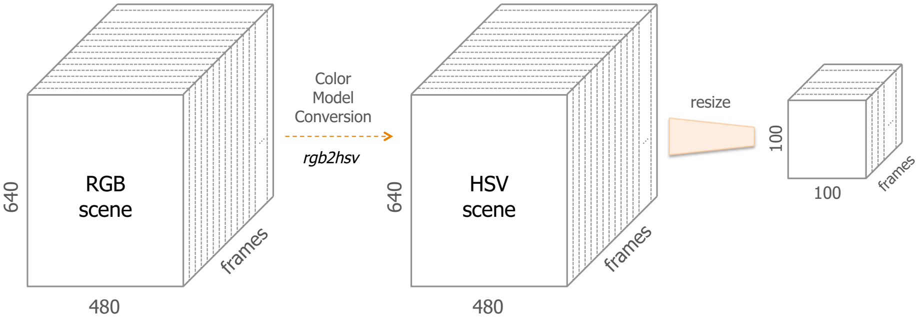

Fig. 1 Scene data pre-processing

M, m, and δ can now be used to obtain the HSV values of each RGB frame, consider (4) to 7:

$$H' = \begin{cases} \frac{G-B}{\delta}, & \text{if M=R} \\ 2. + \frac{B-R}{\delta}, & \text{if M=G} \\ 4. + \frac{R-G}{\delta}, & \text{if M=B} \end{cases} \tag{4}$$

$$H = \mathrm{mod}(H'/6., 1.) \tag{5}$$

$$S = \delta/V \tag{6}$$

$$V = M \tag{7}$$

In an attempt to further reduce complexity, all audios are being used at 16 kHz sample rate, mono channel, and 16 bit depth. Additionally, all the audio samples have the same size. After pre-processing the data we have a total of 4,323 video splices, with 2,147 (49.7%) positively annotated and 2,176 (50.3%) negatively so. When dividing it into train and test, the total is 3,182 (1,778 positive and 1,404 negative) and 1,141 splices (369 positive and 772 negative) respectively, shown in Table 2.

4 Proposed model

4.1 Scene2Wav

Our proposed model, Scene2Wav, is an end-to-end deep neural network composed of a Convolutional Neural Network (CNN), an Encoder Deep Recurrent Neural Network (RNN) and a proposed conditional Decoder SampleRNN [15] with Gated Recurrent Units (GRUs).

The proposed architecture, shown in Fig. 2, consists of three modules. The first is the emotional visual feature extraction with CNN. This is followed by further sequence encoding with an Encoder Deep RNN framework. The last module consists of music generation through a decoding process with our proposed Scene2Wav decoder. This decoder is our proposed conditional SampleRNN that conditionally considers emotional aspects in music. In other words, the proposed decoder is conditioned on the encoded emotional visual sequence features vector **v**. The original unconditional SampleRNN model is not able to generate music corresponding to a desired video, instead simply generating random music. Because of that limitation, the conditional aspect of our model is essential to allow for emotion consideration and effective influence of the encoded information into our desired output.

The scene input is given to the CNN module, responsible for the frame by frame feature extraction, and these features are given to the Encoder RNN, responsible for encoding the visual features as a sequence. These encoded features are then used as input to the Scene2Wav decoder in order to generate audio samples of the same duration as the scene

Table 2 Dataset distribution: train and test, positive and negative emotion scores

	Positive	Negative	Total
Train	1,778	1,404	3,182
Test	369	772	1,141
Total	2,147	2,176	4,323

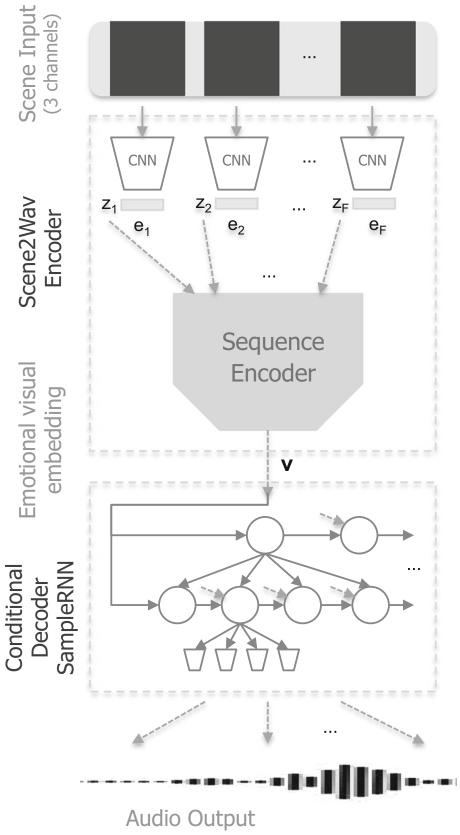

Fig. 2 Proposed model architecture for Scene2Wav. It consists of three modules: emotional visual feature extraction with CNN, sequence encoding with an Encoder Deep RNN framework, and music generation with our proposed conditional SampleRNN, conditioned on the encoded emotional visual sequence features vector **v**

and that evoke similar emotions to the original visual input. The following subsections delve into the details of each module.

4.2 Emotional visual feature extraction with CNN

The Convolutional Neural Network (CNN) module is a deep neural network formed by sequences of Convolutional and Max Pooling layers stacked one after the other, with architecture illustrated in Fig. 3. This module is trained with the emotion score as target (see Table 2) and scene frames information in a supervised learning manner, thus being responsible for the scene classification according to its emotion score and the feature extraction by utilizing the information obtained before the fully connected layer.

The scene input has shape $(f, 3, 100, 100)$, where f is the number of frames in a scene. Since the scenes we're considering have duration of 3 seconds at 10 frames per second, $f = 30$, and since we're considering frames in the HSV colour model, each frame has 3 channels of size 100×100. This scene information is given to a model with 3 sets of Convolution, Batch Normalization, ReLU and Max Pooling layers in such a manner that there are $(f, 108)$ features by the end of it. This is followed by a Fully Connected layer and a Softmax activation function, with the output being the binary representation of the emotion classes.

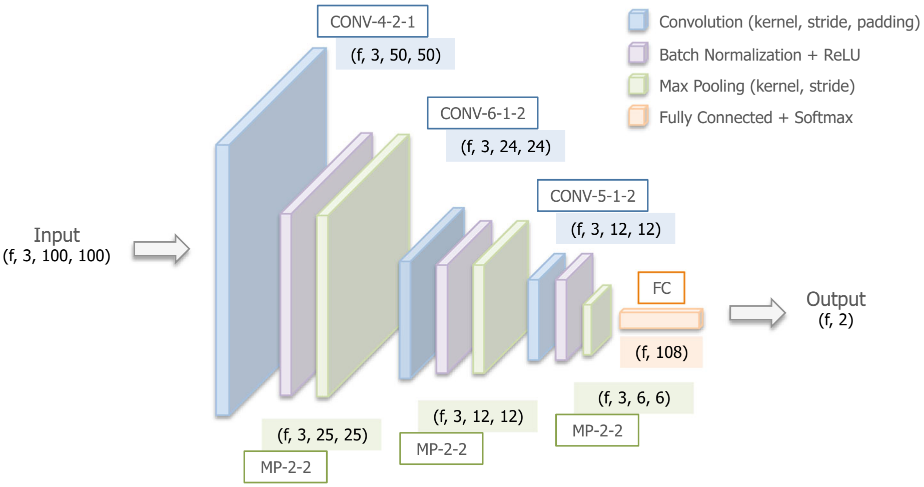

Fig. 3 Convolutional Neural Network (CNN) for visual emotion feature extraction

4.3 Domain transformation with deep encoder RNNs

After obtaining each frames features, we need to consider our data's dynamic time properties. The model chosen for this task is a Recurrent Neural Network (RNN), since it considers dynamic properties of data in previous timesteps unlike feedforward networks. Since our aim is to generate music from scene, two different domains, we need a framework that allows that. In order to perform the mentioned domain transformation, we train the sequence-to-sequence (seq2seq) model shown in Fig. 4. Note that the illustrated seq2seq is used as is in the baseline model. Whereas in our proposed model, we substitute the decoder part for a conditional SampleRNN (see Section 4.4) while still keeping the pre-trained encoder.

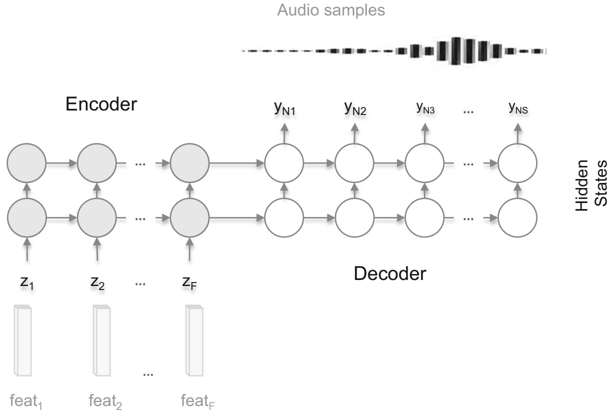

Fig. 4 Encoder-Decoder Recurrent Neural Network (Seq2Seq). This framework is used as is to perform domain transformation in the baseline model. Whereas in the proposed model, the decoder part is substituted for a conditional SampleRNN (see more in Section 4.4)

Vanilla RNNs, however, have limitations when dealing with data with long term dependencies, such as audio. To remedy that, our model uses Gated Recurrent Unit (GRU) cells [5], mathematically described in (8):

$$\begin{aligned} r_t &= \sigma(W_{er}e_t + W_{hr}h_{t-1}) \\ z_t &= \sigma(W_{ez}e_t + W_{hz}h_{t-1}) \\ u_t &= tanh(W_{eu}e_t + W_{hu}(r_t \odot h_{t-1})) \\ h_t &= (1 - z_t)h_{t-1} + z_t u_t \end{aligned} \tag{8}$$

where r_t and z_t are reset and update gates respectively, e is the input, u_t is the candidate activation after considering what should be reset and h_t is the final activation or hidden state of the cell. For simplicity, this description can be generalized as (9):

$$h_t = \mathcal{H}(h_{t-1}, e_t) \tag{9}$$

During the training of the baseline model, this module receives the features extracted from CNN as input, encodes it with a 2-layer GRU encoder, and decodes it with another 2-layer GRU decoder into the target audio. In our proposed model, the 2-layer GRU decoder is substituted for a conditional SampleRNN, which is explained in more detail in the following section. Our final Scene2Wav encoder is thus composed of a CNN for emotional visual feature extraction and a 2-layer GRU encoder for sequential encoding of these features.

4.4 Conditional SampleRNN for music generation

The original SampleRNN [15] is originally proposed to generate random audio samples by modeling the probability of a sequence of audio samples $X = x_1, x_2, ..., x_T$ as the product of the probabilities of each sample conditioned on all previous samples, as shown in (10):

$$p(X) = \prod_{i=0}^{T-1} p(x_{i+1}|x_1, ..., x_i) \tag{10}$$

However, this model has the limitation of not being conditioned on any feature set. In our case, this means that the model is not able to generate corresponding music given a video, instead generating random audio signals. In this section, we propose a conditional SampleRNN to be used as our Scene2Wav decoder for music generation with the scene visual features. It is critical to consider the emotional visual vector in the SampleRNN to generate music that carries the same emotional aspect as the visual input. So, needless to say, this module is essential for the completion of the task suggested in this work, and it does so by additionally modeling the probability of X conditioned on the extracted visual features $\mathbf{v}$, as indicated in (11):

$$p(X|\mathbf{v}) = \prod_{i=0}^{T-1} p(x_{i+1}|x_1, ..., x_i, \mathbf{v}) \tag{11}$$

The proposed model, shown in Fig. 5, is divided into three tiers, or layers, with Tiers 3 and 2 each being a frame-level module and Tier 1 being a sample-level module. The frame-level modules are deep RNNs, and they are responsible for modeling chunks of audio conditioned on the sequentially encoded emotional visual features $\mathbf{v}$. As can be seen in Fig. 5, Tiers 3 and 2 are both frame-level modules, but each operates at different temporal resolutions: Tier 3 processes longer windows of audio, allowing for extraction of global

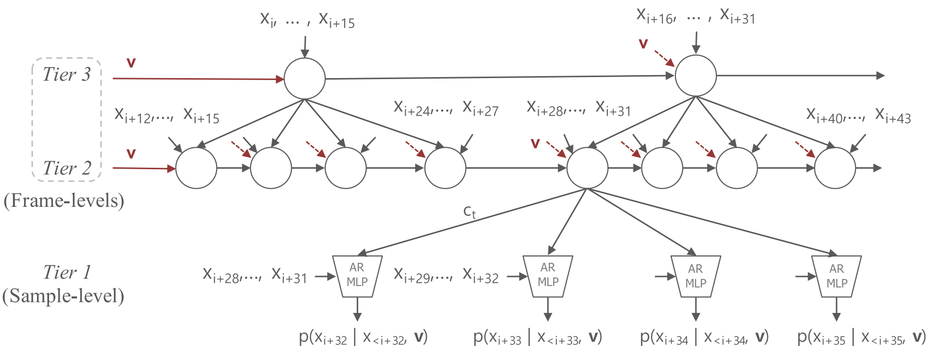
Tier 3
Tier 2
(Frame-levels)
Tier 1
(Sample-level)
v
v
$x_i, \dots, x_{i+15}$
$x_{i+16}, \dots, x_{i+31}$
$x_{i+12}, \dots, x_{i+15}$
$x_{i+24}, \dots, x_{i+27}$
$x_{i+28}, \dots, x_{i+31}$
$x_{i+40}, \dots, x_{i+43}$
c_t
$x_{i+28}, \dots, x_{i+31}$
$x_{i+29}, \dots, x_{i+32}$
AR MLP
$p(x_{i+32} \mid x_{<i+32}, \mathbf{v})$
$p(x_{i+33} \mid x_{<i+33}, \mathbf{v})$
$p(x_{i+34} \mid x_{<i+34}, \mathbf{v})$
$p(x_{i+35} \mid x_{<i+35}, \mathbf{v})$

Fig. 5 The proposed conditional SampleRNN consists of two modules: frame-level and sample-level. The frame-level is composed of two deep RNNs and is responsible for the audio modeling conditioned on the sequentially encoded emotional visual features **v**. The sample-level is composed of an MLP and it is responsible for the next sample prediction given the previous samples and conditioned features

characteristics of the audio, and Tier 2 processes smaller windows, allowing for modeling of more refined audio characteristics. Finally, Tier 1 is a sample-level module responsible for the next sample prediction, with a Multilayer Perceptron (MLP), given the previous samples and conditioned features. Note that the conventional model is structured similarly except that frame-level modules are not conditioned.

The model receives an audio sequence as input, which is divided into chunks of non-overlapping frames of size $FS^{(k)}$, where k is the tier number. The first level (Tier 3) receives $FS^{(3)} = 16$ audio samples and the second and third levels each receive $FS^{(2)} = FS^{(1)} = 4$. This means that the higher level has a larger receptive field and is able to model longer dependencies, while the lower levels are responsible for modeling samples that are closer together in time. In our proposed Scene2Wav decoder, unlike in the conventional model, Tiers 2 and 3's cells also take into consideration the emotional visual embedding **v** obtained from the CNN and RNN encoder. This is mathematically formalized in (12):

$$\begin{aligned} r_t' &= \sigma(W_{er}e_t + W_{hr}h_{t-1} + \mathbf{W_{vr}v}) \\ z_t' &= \sigma(W_{ez}e_t + W_{hz}h_{t-1} + \mathbf{W_{vz}v}) \\ u_t' &= tanh(W_{eu}e_t + W_{hu}(r_t \odot h_{t-1}) + \mathbf{W_{vu}v}) \\ h_t' &= (1 - z_t)h_{t-1} + z_t u_t \end{aligned} \tag{12}$$

which can be simplified as (13):

$$h_t' = \mathcal{H}'(h_{t-1}, e_t, \mathbf{v}) \tag{13}$$

Each tier's relevant input $inp_t^{(k)}$, hidden state $h_t^{(k)}$ and output $c_t^{(k)}$ is calculated as follows. Tier 3 ($k = 3$), (14), receives as input frames of the original input $inp_t^{(3)}$ of size $FS^{(3)} = 16$, denoted by $f_t^{(3)}$, and the emotional visual embedding **v**, and outputs a conditional vector $c_t^{(3)}$, which is calculated as a weighted sum of the current layer's hidden state, to be used in the following tier.

$$\begin{aligned} inp_t^{(3)} &= f_t^{(3)} \\ h_t^{(3)} &= \mathcal{H}'(h_{t-1}^{(3)}, inp_t^{(3)}, \mathbf{v}) \\ c_{(t-1)*4+j}^{(3)} &= W_j^{(3)} h_t^{(3)}, 1 \le j \le 4 \end{aligned} \tag{14}$$

The following tier, (15), also takes as input the emotional visual embedding $\mathbf{v}$ and a linear combination of the conditional vector from the previous tier $c_t^{(3)}$ and the current input frame $f_t^{(2)}$ of size $FS^{(2)} = 4$:

$$\begin{aligned} inp_t^{(2)} &= W_x^{(2)} f_t^{(2)} + c_t^{(3)} \\ h_t^{(2)} &= \mathcal{H}'(h_{t-1}^{(2)}, inp_t^{(2)}, \mathbf{v}) \\ c_{(t-1)*4+j}^{(2)} &= W_j^{(2)} h_t^{(2)}, 1 \leq j \leq 4 \end{aligned} \tag{15}$$

where $W_x^{(2)}$ is a matrix used to obtain the linear combination between $f_t^{(2)}$ and $c_t^{(3)}$. Similarly to the conventional SampleRNN, the final tier, (16) is an MLP that models the probability of input x_{i+1} given all previous samples and $\mathbf{v}$, which is encoded in the conditional vector from the previous tier $c_i^{(2)}$:

$$\begin{aligned} inp_i^{(1)} &= W_x^{(1)} f_i^{(1)} + c_i^{(2)} \\ p(x_{i+1}|x_1, ..., x_i, \mathbf{v}) &= Softmax(MLP(inp_i^{(1)})) \end{aligned} \tag{16}$$

where $W_x^{(1)}$ is used to obtain the linear combination between $f_i^{(1)}$ and $c_i^{(2)}$.

4.5 Baseline model: convolutional sequence-to-sequence

Due to its proven performance and effectiveness in multimodal tasks, a Convolutional Seq2Seq model [35] is chosen as our baseline model in order to evaluate the effect our proposed conditional decoder SampleRNN has over a regular decoder RNN. This model is also trained by end-to-end manner with architecture differing only in the decoder module. Similarly to our proposed model, the baseline model extracts emotional visual features using a CNN, encodes the sequence of features with a Deep RNN and decodes it into a sequence of audio sample with another Deep RNN in an Encoder-Decoder framework, as shown in Fig. 6.

5 Results and discussion

5.1 Training configuration

Both the proposed and baseline models are trained on the previously discussed 3 second spliced dataset (see Section 3) and both use the same CNN and Encoder RNN with heuristically determined configurations. The CNN model is a binary classifier with accuracy of 95.36% and trained for 40 epochs with Adam Optimizer, Cross-Entropy criterion, learning rate of 0.001, momentum of 0.98, weight decay of $1e-4$, scheduler of 0.8, and dropout rate of 0.2. The encoder is a 2-layer deep GRU-RNN with 128 hidden units, trained for 20 epochs, with ASGD Optimizer, learning rate of 0.001, momentum of 0.98, weight decay of $1e-5$, and scheduler of 0.8. Lastly, in the proposed model, is the conditional SampleRNN decoder, a 2-layer deep RNN with 1,024 hidden units, batch size of 128, and quantization level q of 256, corresponding to a per-sample bit depth of 8. Our proposed model is trained in end-to-end manner for on 1 GPU Titan X in the course of 6.5 days with early stop, converging to a training loss of 1.0510.

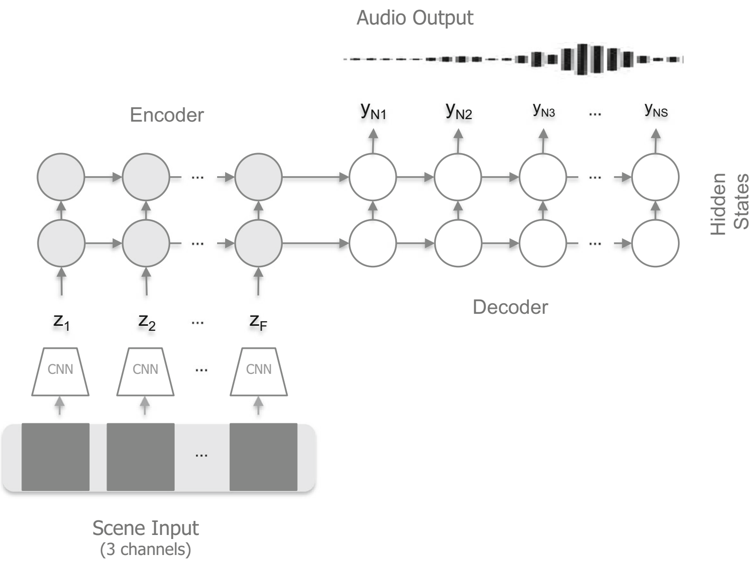

Fig. 6 Baseline model architecture for ConvSeq2Seq consisting of two modules: emotional visual feature extraction with CNN, and sequence encoding and decoding with an Encoder-Decoder Deep RNN framework

5.2 Examples of generated samples

Experimental results are showcased in Figs. 7 and 8 for samples labeled with negative and positive emotion respectively. Each figure shows a sample generated music[3] and their respective visual inputs (video frames). In each image, the waveforms for the original audio, the music generated from the baseline model, and the music generated from the proposed model are also plotted in that order.

The results show that, in both positive and negative emotionally charged samples, our Scene2Wav model is able to generate music that more accurately retains the dynamic properties of the original audio, as is demonstrated by similarities in the audio signal waveforms. This means that it is better able to elicit emotion similar to the visual input, unlike the baseline model, which has difficulties in this area. Results also show that our model generates musics that are less noisy and more plausible than the baseline model, whereas the baseline model produces noisy and non-novel sounds.

5.3 Human evaluation: qualitative metric

The most widely accepted measure of performance for researchers in this field to evaluate the quality of generated audio is through human evaluation feedback. We conduct an academic survey with the goal of evaluating the quality of the music generated by our model and thus check its performance. We utilize Amazon Mechanical Turk (MTurk) [2], a marketplace for Human Intelligence Tasks (HITs), where 20 subjects are shown 20 sets of video and music samples (10 positive and 10 negative) in a survey[4] designed with Google Forms.

[3] Audio available online at https://github.com/gcunhase/Scene2Wav/tree/master/results_generated_samples

[4] https://tinyurl.com/y7rn8jqj

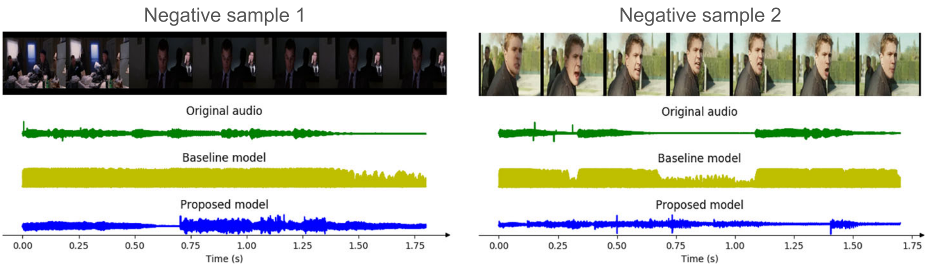

Fig. 7 Negative sample videos 1 and 2 with frames and music (original, baseline and proposed)

For each experiment set, the subjects are given 3 short music samples and they are asked to subjectively compare them pairwise, in what is called an AB experiment, and choose their most preferred audio. These evaluations are shown in Figs. 9 and 10.

Results show the preference scores in percentage of the subjective paired comparison tests between music generated by our proposed model **P** (blue), music generated by the baseline model **B** (green) and original music **O** (orange). The results show that the evaluators prefer our music to the baseline music in 17 out of the 20 samples, with them showing no preference to most of the remaining samples. The much larger gap between $B \times O$ when compared to $P \times O$ reinforces that conclusion, indicating that there's a larger discrepancy between original and baseline than there is for original and proposed music when subject are asked to choose between two of them. Furthermore, even though subjects choose the original music more often than ours, given that it is less noisy and thus more pleasant to humans, the small gap in the $P \times O$ experiment shows that subjects also approve of our music, especially in negative samples 1, 2, and 7, and positive samples 1, 2, 9, and 10, where the gap is minimal.

5.4 Perceptual audio metric: quantitative metric

There is currently no widely accepted quantitative metric to measure the quality of audio generated by machine learning models. More recently, however, Pranay et. al [14] have taken steps in that direction by proposing a perceptual audio metric (PAM) for perceptual assessment in audio. The advantage of this method, and the reason why we are using it here, is the metric's good correlation with human evaluators, achieved by training a deep neural network with crowdsourced human judgments. The perceptual score can be obtained by

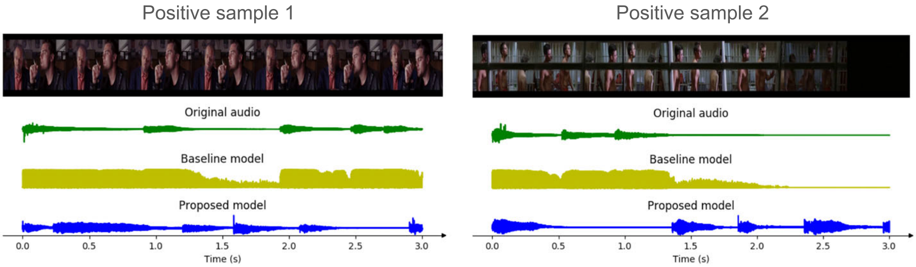

Fig. 8 Positive sample videos 1 and 2 with frames and music (original, baseline and proposed)

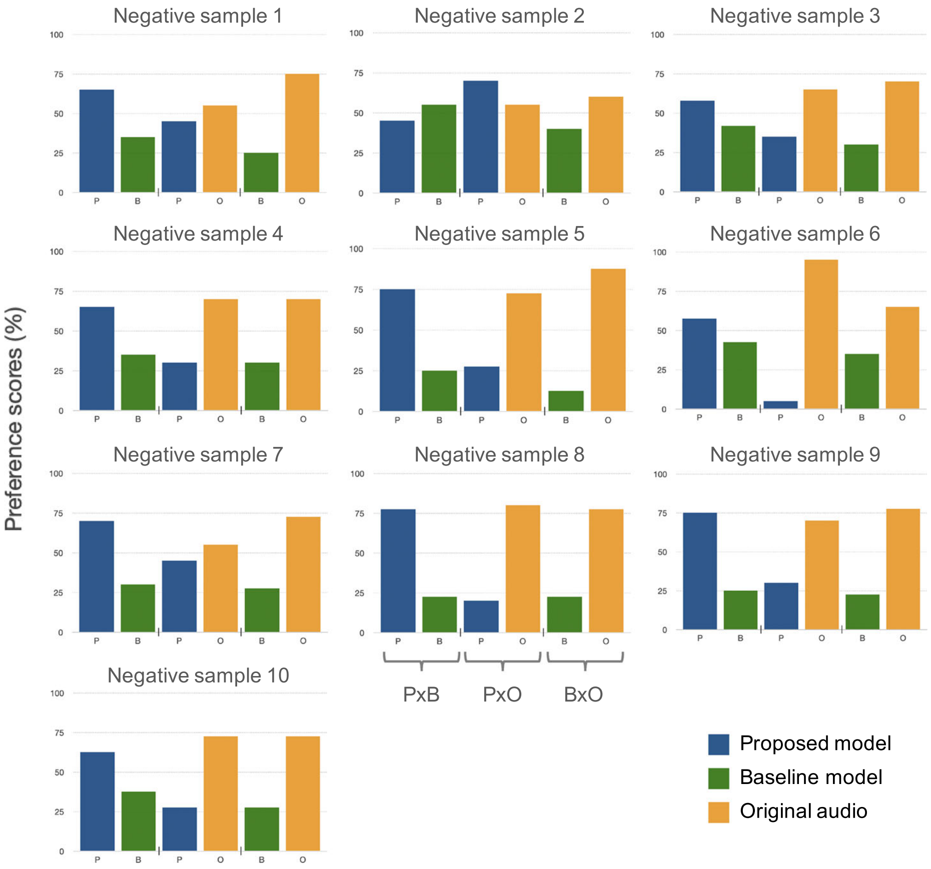

Fig. 9 Subjective paired comparison test (AB experiment) for 10 negative samples, with samples 1 and 2 being shown in Fig. 7. Each chart contains pairwise comparisons of $P \times B$, $P \times O$ and $B \times O$, with P (*blue*) representing the music generated by our proposed model, B (*green*) the music generated by the baseline model and O (*orange*) the original music

comparing the target audio a_t with the audio in consideration a, as shown in (17):

$$\mathrm{p}_{\mathrm{score}} = PAM(a_t, a) \tag{17}$$

where a is either the audio generated by the baseline or proposed model. Note, however, that PAM is still a distance metric nonetheless, meaning that it's extremely dependent on the relationship between evaluated audio and target audio. In other words, music similar to the original audio might have higher scores than high quality novel music charged with the appropriate emotion. Table 3 shows that our model obtains better and more consistent scores in the perceptual audio metric when compared to the baseline model, for both negative and positively inclined cases.

5.5 Emotion evaluation

For the emotion evaluation, we use the Circle of Fifths [8] and emotional descriptions on music chords [31]. The first, Circle of Fifths, is in the core of music theory and it is a visual

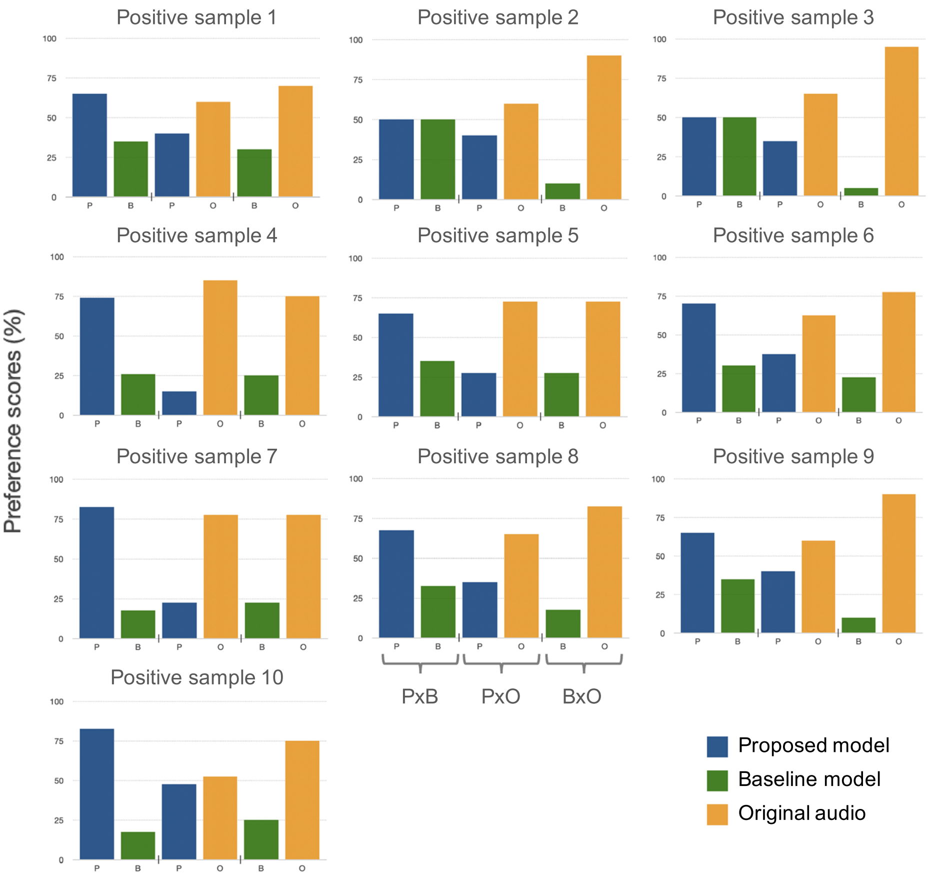

Fig. 10 Subjective paired comparison test (AB experiment) for 10 positive samples, with samples 1 and 2 being shown in Fig. 8. Each chart contains pairwise comparisons of $P \times B$, $P \times O$ and $B \times O$, with P (*blue*) representing the music generated by our proposed model, B (*green*) the music generated by the baseline model and O (*orange*) the original music

description of 12 chromatic notes. As for the second, we use Schubart's book [23], translated by Rita Steblin [31], as it is one of the most influential descriptions of music chords to this day. In the book, the author describes each chord as having an emotional affect on the listener: C innocently happy, D victorious and triumphant, $F^{\#}$ conquering difficulties and sighs of relief, $D^{\#}$ cruel, E quarrelsome, and so on. Note that some notes, such as D, carry positive emotion, but after the addition of a sharp key ($^{\#}$), they start reflecting the opposite. This shift between natural and sharp chords can be explained by the association between sharp keys and disharmony [9]. Finally, we fuse the Circle of Fifths and emotional descriptions of music chords into one concept and present the novel Emotional Circle of Fifths in Fig. 11. In this circle, there are 5 chords representing positive emotions (red color), 6 representing negative emotions (blue color) and 1 chord, $F^{\#}$, that is both positive and negative (purple color).

We detect chords from the generated musics using the Python library Music21, a robust toolkit for computer-aided musicology. We then analyze the detected chords in the generated musics, both from our proposed model Scene2Wav and the baseline model ConvSeq2Seq.

Table 3 Perceptual audio metric (PAM) used to evaluate the quality of generated music in Figs. 9 and 10

	PAM	
	Baseline	Ours
Negative Samples		
N1	3.88	**2.59**
N2	**2.44**	2.97
N3	**2.64**	3.02
N4	4.21	**2.89**
N5	1.77	**1.39**
N6	4.70	**2.29**
N7	**2.94**	5.11
N8	**2.54**	2.73
N9	4.45	**3.62**
N10	4.20	**1.94**
Positive Samples		
P1	3.38	**3.30**
P2	2.83	**2.81**
P3	2.96	**2.65**
P4	4.56	**2.93**
P5	4.37	**4.13**
P6	**4.27**	4.49
P7	**3.22**	5.31
P8	5.28	**3.58**
P9	3.68	**1.26**
P10	**2.43**	2.86

This is a distance metric, so lower values are desirable. The first half of the table shows the negative samples (N) and the bottom half the positive ones (P), with 10 samples each. Better results of each sample are shown in bold

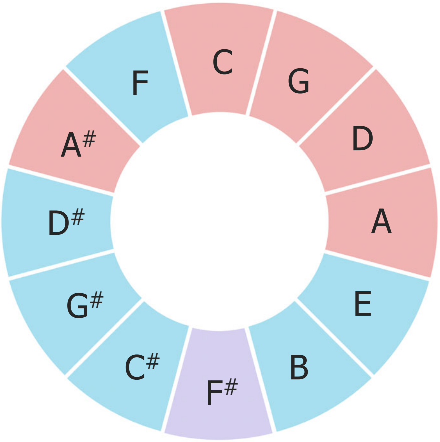

Fig. 11 Emotional Circle of Fifths, with *red* representing chords with positive emotion, *blue* negative and *purple* both

The results, displayed in Table 4, show that our proposed model is able to produce musics with the same emotion as the scenes more often than the baseline model, especially when considering videos with positive emotion. As can be seen, ConvSeq2Seq is unable to consistently produce music with positive emotions when presented with positively charged scenes, unlike our model. With this, we can affirm that our model generates appropriate musics to given scenes, that is preferred by human users and evokes emotion correctly.

5.5.1 Extended emotion evaluation

In this section, we extend the emotion analysis on longer samples obtained with the same dataset and also with short samples obtained with the additional DEAP dataset [11]. The goal is to fortify our claim regarding the ability of our model to generate music that elicits a similar emotion to the input scene. We first show emotion evaluation on samples of 10 seconds duration obtained with the same dataset used in this paper, the COGNIMUSE dataset [46]. The results, displayed in Table 5, show that our model is better at generating

Table 4 Emotional evaluation of generated music

	Baseline (ConvSeq2Seq)		Ours (Scene2Wav)	
Samples	Chords	Emotion	Chords	Emotion
N1	$C^{\#}\ F\ B\ C\ D$	-	$F\ F^{\#}$	-
N2	$D^{\#}\ E\ F\ F^{\#}\ G$	-	$D^{\#}\ F\ B\ A\ A^{\#}$	-
N3	$C\ D$	+	$D\ G\ A\ A^{\#}$	+
N4	$C^{\#}\ E\ D$	-	$G^{\#}\ B$	-
N5	$G^{\#}\ A^{\#}$	0	$E\ B\ D$	-
N6	B	-	$D^{\#}\ F$	-
N7	$C^{\#}\ B$	-	$C^{\#}\ G^{\#}\ A^{\#}$	-
N8	$C^{\#}\ D^{\#}\ E\ C$	-	$D^{\#}\ G^{\#}\ F^{\#}$	-
N9	$C^{\#}$	-	$C^{\#}\ D^{\#}\ G^{\#}$	-
N10	$D^{\#}\ F$	-	$E\ F\ B\ F^{\#}\ D$	-
P1	$C^{\#}\ G^{\#}\ G\ A$	0	$G\ A$	+
P2	$D^{\#}$	-	$G^{\#}\ A\ A^{\#}$	+
P3	$E\ B\ A\ A^{\#}$	0	$D^{\#}\ E\ F\ A$	-
P4	$C^{\#}\ F$	-	$F^{\#}\ A^{\#}$	+
P5	$D^{\#}\ E$	-	D	+
P6	$F\ G\ A^{\#}$	+	$F^{\#}\ G$	+
P7	$D^{\#}\ E\ F^{\#}\ C\ D\ G$	+	$G^{\#}\ D$	0
P8	$C^{\#}\ D^{\#}\ F\ C\ D\ A^{\#}$	0	$A^{\#}$	+
P9	$E\ F\ B\ G$	-	$F^{\#}\ C\ A^{\#}$	+
P10	$E\ F\ F^{\#}$	-	$D^{\#}$	-

The first half of the table shows the negative samples (N) and the bottom half the positive ones (P), with 10 samples each. Each sample is shown together with their respective chords and most relevant emotion according to the Emotional Circle of Fifths. The detected emotions can be + for positive, − for negative is 0 for neither or tie

Table 5 Emotion evaluation of generated music with 10 seconds duration

	Baseline (ConvSeq2Seq)		Ours (Scene2Wav)	
Samples	Chords	Emotion	Chords	Emotion
N1	$C^{\#}\ G^{\#}\ C\ G$	0	$C^{\#}\ D^{\#}\ E\ F\ G^{\#}\ B\ F^{\#}\ C\ D\ G$	-
N2	$C^{\#}\ E\ B\ C\ D\ A\ A^{\#}$	+	$C^{\#}\ D^{\#}\ E\ F\ G^{\#}\ B\ F^{\#}\ C\ D\ G\ A$	-
N3	$C^{\#}\ F\ G^{\#}\ C\ G\ A$	0	$D^{\#}\ E\ F\ G^{\#}\ B\ F^{\#}\ D\ A^{\#}$	-
N4	$E\ B\ F^{\#}\ G\ A$	0	$C^{\#}\ B\ A\ A^{\#}$	0
N5	$C^{\#}\ D^{\#}\ F\ F^{\#}\ C\ A^{\#}$	-	$C^{\#}\ D^{\#}\ F\ G^{\#}\ B\ D\ A\ A^{\#}$	-
P1	$C^{\#}\ E\ G^{\#}\ B\ F^{\#}\ D\ A$	-	$E\ B\ C\ D$	0
P2	$C^{\#}\ D^{\#}\ E\ F\ G^{\#}\ B\ C\ D\ A\ A^{\#}$	-	$C^{\#}\ D^{\#}\ F\ F^{\#}\ A\ A^{\#}$	-
P3	$D^{\#}\ F\ C\ D\ G\ A\ A^{\#}$	+	$G^{\#}\ B\ D$	-
P4	$D^{\#}\ E\ F^{\#}\ D$	-	$D^{\#}\ F\ G^{\#}\ F^{\#}\ G\ A$	-
P5	$C^{\#}\ D^{\#}\ E\ F\ G^{\#}\ F^{\#}\ D\ G\ A\ A^{\#}$	-	$G^{\#}\ F^{\#}\ D\ A$	+

The first half of the table shows the negative samples (N) and the bottom half the positive ones (P), with 5 samples each

audio with negative emotion and performs similarly when attempting to generate audio for positive scenes. In other words, our proposed method can be applied for longer video samples without additional modification while still being able to model music with appropriate emotion.

6 Conclusion

In this study, we proposed Scene2Wav, an end-to-end deep neural network able to generate rich, raw music signals from video annotated with emotion scores. The proposed model uses a conditional SampleRNN decoder that takes into account sequentially encoded emotional visual features, obtained from a CNN followed by an RNN encoder, to generate related emotional music. By considering the encoded emotional visual features as condition, the model is able to generate music corresponding to a given video, thus alleviating the limitation in the unconditional SampleRNN model. The model is evaluated by human subjects in an online marketplace on their music preferences and an emotion evaluation is done by detecting chords and analyzing them with an Emotional Circle of Fifths. Human evaluation is used as a qualitative measure of performance and experiments show that our model produces more pleasant and less noisy musics. Moreover, results of the pairwise comparison between music generated by our proposed model and music generated by the baseline model show that subjects prefer our musics. Emotion evaluation on short and long samples, and even samples obtained from a different dataset, show that Scene2Wav is better able to produce music with the same emotion as the scene. This study has proven to be a stepping stone for music generation from emotionally annotated videos. However, further research still needs to be done to generate more musically pleasant audio and to be able to better elicit the desired emotion reaction from subjects. Future works include increasing the size of the dataset, improving longer-term dependency modeling for longer audios, and including more emotion classes, such as considering arousal and adding a neutral class.

This work was partly supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (2016-0-00564, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding) (50%) and Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [18ZS1100, Core Technology Research for Self-Improving Artificial Intelligence System] (50%).

References

1. Bravo F (2012) The influence of music on the emotional interpretation of visual contexts. In: International symposium on computer music modeling and retrieval. Springer, pp 366–377
2. Buhrmester M, Kwang T, Gosling SD (2011) Amazon's mechanical turk: a new source of inexpensive, yet high-quality, data? Perspectives on Psychological Science 6(1):3–5
3. Çevikalp H, Dordinejad GG, Elmas M (2017) Feature extraction with convolutional neural networks for aerial image retrieval. In: Signal processing and communications applications conference (SIU), 2017 25th. IEEE, pp 1–4
4. Chang JD, Yu SS, Chen HH, Tsai CS (2010) Hsv-based color texture image classification using wavelet transform and motif patterns. J Comput 20(4):63–69
5. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). Association for Computational Linguistics, Doha, Qatar, pp 1724 – 1734. http://www.aclweb.org/anthology/D14-1179
6. Cunha Sergio G, Lee M (2020) Emotional video to audio transformation using deep recurrent neural networks and a neuro-fuzzy system. Math Probl Eng 2020
7. Hanjalic A, Xu LQ (2005) Affective video content representation and modeling. IEEE Trans Multimed 7(1):143–154
8. Heinichen JD (1728) Der general-bass in der composition. Ripol Classic Publishing House
9. Ishiguro MA (2010) The affective properties of keys in instrumental music from the late nineteenth and early twentieth centuries. Master's thesis University of Massachusett Amherst
10. Jaimovich J, Coghlan N, Knapp RB (2012) Emotion in motion: a study of music and affective response. In: International symposium on computer music modeling and retrieval. Springer, pp 19–43
11. Koelstra S, Muhl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, Pun T, Nijholt A, Patras I (2011) Deap: a database for emotion analysis; using physiological signals. IEEE Trans Affective Comput 3(1):18–31
12. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
13. LeCun Y, Haffner P, Bottou L, Bengio Y (1999) Object recognition with gradient-based learning. In: Shape, contour and grouping in computer vision. Springer, pp 319–345
14. Manocha P, Finkelstein A, Jin Z, Bryan NJ, Zhang R, Mysore GJ (2020) A differentiable perceptual audio metric learned from just noticeable differences. arXiv:2001.04460
15. Mehri S, Kumar K, Gulrajani I, Kumar R, Jain S, Sotelo J, Courville A, Bengio Y (2016) Samplernn: an unconditional end-to-end neural audio generation model. arXiv:1612.07837
16. Mikels JA, Fredrickson BL, Larkin GR, Lindberg CM, Maglio SJ, Reuter-Lorenz PA (2005) Emotional category data on images from the international affective picture system. Behav Res Methods 37(4):626–630
17. Morriss-Kay GM (2010) The evolution of human artistic creativity. J Anat 216(2):158–176
18. Nanni L, Ghidoni S, Brahnam S (2017) Handcrafted vs. non-handcrafted features for computer vision classification. Pattern Recogn 71:158–172
19. Neubig G (2017) Neural machine translation and sequence-to-sequence models: a tutorial. arXiv:1703.01619
20. Oatley K, Keltner D, Jenkins JM (2006) Understanding emotions. Blackwell Publishing
21. Rouzic M (2008) Photosounder. http://photosounder.com/
22. Savage TM, Vogel KE (2013) An introduction to digital multimedia. Jones & Bartlett Publishers
23. Schubart CFD (1806) Christ. Fried. Dan. Schubart's Ideen zu einer Ästhetik der Tonkunst. Degen
24. Sergio GC, Lee M (2016) Audio generation from scene considering its emotion aspect. In: International conference on neural information processing. Springer, Kyoto, pp 74–81
25. Sergio GC, Moirangthem DS, Lee M (2018) Temporal hierarchies in sequence to sequence for sentence correction. In: 2018 international joint conference on neural networks (IJCNN). IEEE, pp 1–7

26. Shan MK, Kuo FF, Chiang MF, Lee SY (2009) Emotion-based music recommendation by affinity discovery from film music. Expert Systems with Applications 36(4):7666–7674
27. Shan MK, Kuo FF, Chiang MF, Lee SY (2009) Emotion-based music recommendation by affinity discovery from film music. Expert Systems with Applications 36(4):7666–7674
28. Singh JF (2012) Paint2sound. http://flexibeatz.weebly.com/paint2sound.html
29. Soleymani M, Pantic M, Pun T (2012) Multimodal emotion recognition in response to videos. IEEE Trans Affective Comput 3(2):211–223
30. Sotelo J, Mehri S, Kumar K, Santos JF, Kastner K, Courville A, Bengio Y (2017) Char2wav: end-to-end speech synthesis
31. Steblin R (2005) A history of key characteristics in the eighteenth and early nineteenth centuries. University of Rochester Press
32. Ullrich K, van der Wel E (2017) Music transcription with convolutional sequence-to-sequence models. In: Proceedings of the 18th international society for music information retrieval conference
33. Van Den Oord A, Dieleman S, Zen H, Simonyan K, Vinyals O, Graves A, Kalchbrenner N, Senior AW, Kavukcuoglu K (2016) Wavenet: a generative model for raw audio. In: SSW, p 125
34. van der Zwaag MD, Westerink JH, van den Broek EL (2009) Deploying music characteristics for an affective music player. In: 3rd international conference on affective computing and intelligent interaction and workshops, 2009. ACII 2009. IEEE, pp 1–7
35. Venugopalan S, Rohrbach M, Donahue J, Mooney R, Darrell T, Saenko K (2015) Sequence to sequence-video to text. In: Proceedings of the IEEE international conference on computer vision, pp 4534–4542
36. Wang HL, Cheong LF (2006) Affective understanding in film. IEEE Trans Circ Syst Video Technol 16(6):689–704
37. van der Wel E, Ullrich K (2017) Optical music recognition with convolutional sequence-to-sequence models. arXiv:1707.04877
38. White D (2011) Sonicphoto. http://www.skytopia.com/software/sonicphoto/
39. Yang X, Fan Y (2018) Feature extraction using convolutional neural networks for multi-atlas based image segmentation. In: Medical imaging 2018: image processing, vol 10574, p 1057439. International Society for Optics and Photonics
40. Yanulevskaya V, van Gemert JC, Roth K, Herbold AK, Sebe N, Geusebroek JM (2008) Emotional valence categorization using holistic image features. In: 15th IEEE international conference on image processing, 2008. ICIP 2008. IEEE, pp 101–104
41. Zhang Q, Jeong S, Lee M (2012) Autonomous emotion development using incremental modified adaptive neuro-fuzzy inference system. Neurocomputing 86:33–44
42. Zhang Q, Lee M (2012) Emotion development system by interacting with human eeg and natural scene understanding. Cogn Syst Res 14(1):37–49
43. Zhao S, Yao H, Wang F, Jiang X, Zhang W (2014) Emotion based image musicalization. In: 2014 IEEE international conference on multimedia and expo workshops (ICMEW). IEEE, pp 1–6
44. Zhiqiang W, Jun L (2017) A review of object detection based on convolutional neural network. In: Control conference (CCC), 2017 36th chinese. IEEE, pp 11104–11109
45. Zhou C, Horgan M, Kumar V, Vasco C, Darcy D (2018) Voice conversion with conditional samplernn. arXiv:1808.08311
46. Zlatintsi A, Koutras P, Evangelopoulos G, Malandrakis N, Efthymiou N, Pastra K, Potamianos A, Maragos P (2017) Cognimuse: a multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. EURASIP Journal on Image and Video Processing 2017(1):54

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.