Affective Music Composer Proposal

CSC 580 Fall 2022 - Dr. Rodrigo Canaan

Saurav G., Snehith J., Bryce M., Nick S.
California Polytechnic State University,
San Luis Obispo, CA, USA
{sgupta61, sjonnaik, bmolesh, nistaple}@calpoly.edu

Abstract—**In this proposal we will construct a model that generates a novel song composition based on a input text string. This system will unitize two separate stages. In the first stage the text is passed to a NPL model to classify the text to one of four emotional classifications. In the second stage, the classification is used as a input to an Affective model that generates a new song composition.**

Keywords: affective computing, music composition, natural language processing, transfer learning, splicing system

I. INTRODUCTION

A. INTELLIGENT DESIGN

In the field of computing there is a general desire to create models that have the capability of replicating and predicting human emotional response. Historically there has been debate on how this can be achieved. One prevailing thought was that synthetic minds or AI based systems need the computation equivalent of emotion to be effective [1]. However, human emotion is complex and the theory of emotion is disorganized and hard to utilize in the design of intelligent systems. The alternative approach is based on using affects which are a persons basic sense of feelings. These range from unpleasant to pleasant, known as valence, and idle to activated which is known as arousal. Affects provides a way to understand emotional response that is definable to fit within mathematical models while also working well enough to allow systems to mimic the understanding of emotions. Following on this guidance, our work will make use of affects as they allow us to focus on our contributions instead of the complex theory of emotions.

B. AFFECT AND MUSIC

One area that can be examined well with affects is music. It is commonly known that music creates emotions or emotional responses. In general our emotions are partly expressed and understood through voice intonation [2]. Humans can hear anxiety, love, or any number of emotions in voices. Because of this we can also feel emotions through all sounds, including music. However, because of how varied and complex music is it allows for interesting and effective research into affects. Affects and music have been well studied in the field of psychology. In one study, researchers examined music perception and its influence by culture [3]. They examined two groups, one of western listeners, and one that contained members of a native African population and found that both groups preferred their respective music. They also they tested for the recognition of, sad, happy, and scared emotions in both groups of participants. In this test both groups were able to detect the emotions higher than random, this indicates that music has the power to share universal emotions. This indicates that detecting emotions is not limited by culture or social factors and instead a part of the universal nature of music. This is supported by another study comparing Canadian listeners to Congolese Pygmies in which it was determined that changes in arousal were universal between the two groups [4]. These studies support the conclusions that while culture can play a part in connecting timbre, modes, etc to emotions, there is a underlying universal emotional response to music.

Given this universal response it is feasible to create a system that can recognise and predict affects in music samples. Our work will contribute to the study of human-like AI by creating a music *splicing system* that operates in affective feature space. Our *affective splicing system* attempts to learn four musical styles as defined by 2-dimensional Scherer space.

II. RELATED WORK

Human intellect is fundamentally creative, which makes AI difficult to imitate. Margaret et al. [5] discusses three different ways, AI techniques can be used to generate new ideas: by combining novel versions of well-known concepts; by investigating the potential of conceptual spaces; and by enacting transformations that make it possible to generate previously impractical concepts.

In music AI, there are methods that recognize and produce stylistic music including genre, composer, and emotions. Biological inspired algorithms have been well applied to the task of automatic music composition [6] [7]. In particular, music splicing is genetic algorithm that utilizes a *splicing system*, a formal language model, proven to capture the learned improvisational style of jazz musicians through MIDI data [8]. Our contributions expand this work by utilizing the MIDI data of songs with valence and arousal features.

III. DATA

We'll be using EMOPIA [9] dataset to train and evaluate our models. EMOPIA dataset is a shared multi-modal (audio and MIDI) database focusing on perceived emotion in pop piano music, to facilitate research on various tasks related to music emotion. The dataset contains 1,078 music clips from 387 songs and clip-level emotion labels annotated by four dedicated annotators.

The data includes title, MIDI, and emotion label of the composition, along with other features associated with each song. These three features will be preprocessed and used in training and validation of the proposed models. Table I shows the number of clips and their average length for each quadrant in Scherer's valence-arousal emotion space, in EMOPIA.

TABLE I
EMOPIA DATSET AND ITS CLASSIFICATION

Quadrant	No. Clips	Avg. length (in sec)
Q1	250	31.9
Q2	265	35.6
Q3	253	40.6
Q4	310	38.2

Let M be defined as the entire corpus of music in EMOPIA, $m_i \in M$ a song in the corpus (MIDI), and $M_{\mathbf{X}}$ be non-overlapping subsets of each category of emotion $\mathbf{X}$ (see Fig 1). The following sections will discuss methods on how to construct formal languages $L_{\mathbf{X}}$ for each of the four categories in 2D Scherer space to achieve the goal of affective music composition.

IV. METHODS

At a high level, our proposed model takes a text cue T as input which represents the title of a song. The input T is encoded into affect-feature space by an NLP model E (see IV-A) to retrieve an affective label $\mathbf{X}$ for T by $\mathbf{X} = E(T)$.

The splicing system $S_{\mathbf{a}}$ associated with the affective label p then produces a word $s_{\mathbf{a}} \in \mathrm{L}_{\mathbf{a}}$ generated by $L_{\mathbf{a}}(S_{\mathbf{a}})$ (see IV-B).

The algorithm deterministically generates words for each language, rejecting those specified by a evaluation method, and samples from this set to produce a ”random” song consistent with the affective label a.

Our major contribution in this work is guiding the entire composing process through affective space to compose songs consistent with an affective style.

A. ENCODER

We will employ transfer training to discover the affective vector of text cues using our dataset (song titles). Four emotion categories — joy, sadness, anger, and fear — are used in a multi-class sentiment analysis task to classify texts.

With our data set, we will do transfer learning using Bidirectional Encoder Representations from Transformers (BERT). It generates natural language vector-space representations appropriate for deep learning algorithms. The BERT family of models leverages the Transformer encoder architecture to interpret each token of input text in the context of all tokens that came before and after it. BERT models are frequently trained on a big corpus of text before being tailored for certain applications.

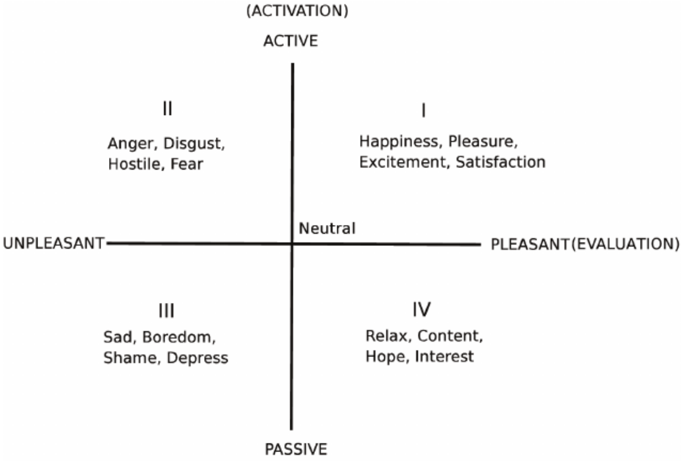

Fig. 1. 2D Scherer affective space along the valence (happy / sad) and arousal (idle / engaged) dimensions. Quadrant I is associated with excitement, Quadrant II is associated with anger, Quadrant III is associated with sadness, and Quadrant IV is associated with relaxation. A vector $\mathbf{v}$ rotated $\frac{\pi}{4}$ about the origin will be in the Quadrant I. The style $\mathbf{X}$ of $\mathbf{v}$ is therefore consistent with Quadrant I, where $\mathbf{X}$ can be Q1, Q2, Q3, or Q4.

BERT was developed using innovative pre-training contextual representations, such as Semi-supervised Sequence Learning, Generative Pre-Training, ELMo, the OpenAI Transformer, ULMFit, and the Transformer, as well as contemporary research. Despite the fact that all of these models are unidirectional or only partially bidirectional, BERT is entirely bidirectional.

TABLE II
BERT COMPARED WITH OTHER APPROACHES

Type	Approach	F1-score
Traditional Machine Learning	Naive Bayes	0.6702
Neural Networks	LSTM + w2v wiki	0.7395
Transfer learning with BERT	finetuned BERT	0.8320

B. COMPOSER

The affective label (from encoder) $\mathbf{X}$ will be now be referred to as the *style* of an arbitrary song m in the language produced by the splicing system $S_{\mathbf{X}}$. Let $L_{\mathbf{X}}$ formally be defined as the language of splicing system $S_{\mathbf{X}}$ as the set of all songs consistent with style $\mathbf{X}$. The goal of our composer is to generate $L_{\mathbf{X}} = L_{\mathbf{X}}(S_{\mathbf{X}})$.

The splicing system [7] $S_{\mathbf{X}}$ associated with style $S_{\mathbf{X}}$ is defined to be a formal language system $S_{\mathbf{X}} = (\mathbf{A}, \mathbf{I_X}, \mathbf{R})$ where $\mathbf{A}$ is the alphabet of musical keys, $\mathbf{I_X}$ as an initial set of songs ($\forall m \in I_x \subset M_{\mathbf{X}} \subset M$), and $\mathbf{R}$ is a set of rules used to during the slicing process. In our work, $\mathbf{A}$ is a simplified version of [8] that extracts a set of n-grams that to token each song m into a vector of features; $\mathbf{I_X}$ is randomly sampled from $M_{\mathbf{X}}$; and R is characterized by a classifier $C_{\mathbf{X}}$ (SVM$_{\mathbf{X}}$) to ensure each song m in $L_{\mathbf{X}}$ is consistent with $\mathbf{X}$, and predictor $P_{\mathbf{X}}$ (LSTM$_{\mathbf{X}}$) is to predict the next splicing operation when generating $L_{\mathbf{X}}$.

Generating a new song can be framed as a search on a syntactic tree of a formal language. The nodes are defined as

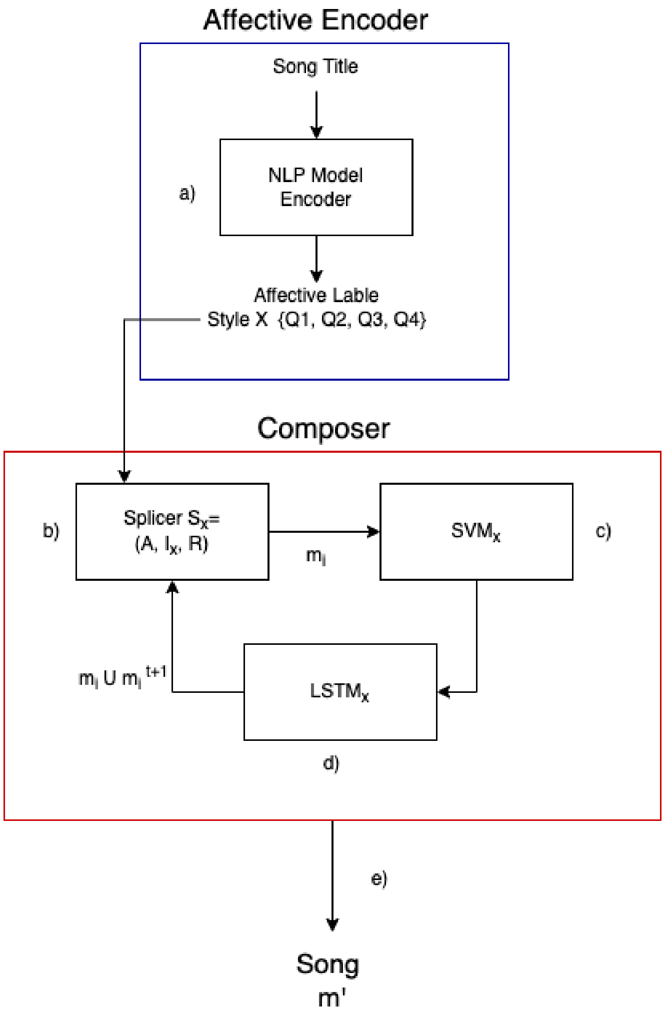

Fig. 2. High level model flow. The encoder a) E convert a text-cue T into a style $\mathbf{X}$. This then selects the corresponding splicing system a) $S_{\mathbf{X}}$ to generate a new song $m\prime$ consistent with style $\mathbf{X}$. Each m_i is randomly taken from the initial set $I_{\mathbf{X}}$ and guided through the splicing loop through the classifier c) to ensure m_i is consistent with style $\mathbf{X}$, and a predictor d) to predict the next splice operation $m_i \leftarrow m_i \cup m_i^{t+1}$ is consistent with style $\mathbf{X}$. The result of each splice is used as the splicing set for the next iteration. The process repeats until the first song m_i spliced reaches the maximum number of bars defined by $\mathbf{R}$. This song $m\prime$ is returned as the output e) of our composer.

the tokens in $\mathbf{A}$, $I_{\mathbf{X}}$ is the initial set of nodes consistent with style $\mathbf{X}$, and $\mathbf{R}$ is the search heuristic. Each splicing system then conducts a stochastic depth-first search of the syntatic tree to find the first path p of depth d defined by R. This search path p is equivalent to the song $m\prime$ generated by the model.

V. Evaluation / Results

The results of our models will be evaluated through subjective listening. If time permits, a survey will be administered to rate how well each song corresponds to its affective category and how well the listener enjoys the song. Participants will be presented with a music excerpt from the composer to their prompt and their responses will be collected on a five point Lichter scale for the following questions:

Aesthetic appeal To evaluate aesthetic appeal and general impression and if the music could keep participant interested.

Melody To measure if the result is convincingly complete, coherent, and imaginative.

Rhythm To check if the music produced is coherent and makes musical sense and if it adds to the aesthetic affectiveness.

Creativity If the composition includes very original, unusual, or imaginative musical ideas.

Affective category score If the affective quotient of the music produced is deemed satisfactory by the participant or prompt engineer.

The results for different affective quadrants will be compared and analyzed.

VI. Conclusions

The overall goal of this project will be to create a system that can analyze a inputs song title to produce a song that contains the same emotional sentiment as the title. This will accomplished through the use of affects which a simplified view of emotions that can be categorized. The four possible affect categories were shown in Fig. 1. Our system will start by passing the a text song title into a affective encoder which will use a NLP model to predict which quadrant the title would fall into. This prediction will then be passed to a composer which will use a splicer, SVM, and LSTM to produce a song that will also fall into the same quadrant as shown in Fig. 1. These output songs will be evaluated subjectively through a user survey in which a user will evaluate the output song in five different questions using a five point Lichter scale. From these results we will be able to determine if we have created a system capable of generating musical composition that align with the emotional context of a submitted prompt.

VII. Author Contributions

1) Saurav G. - Engineer
Encoder, NLP

2) Snehith J. - Project Leader, Engineer
Data, Evaluation

3) Bryce M. - Engineer
Composer, Report

4) Nick S. - Computer Scientist
Algorithms, Research

VIII. References

References

[1] D. N. Davis and S. C. Lewis, “Affect and affordance: Architectures without emotion,” in *AAAI, editor, AAAI Spring symposium*, 2004.
[2] R. Picard, “Affective computing,” 1995.

[3] T. Fritz, S. Jentschke, N. Gosselin, D. Sammler, I. Peretz, R. Turner, A. D. Friederici, and S. Koelsch, “Universal recognition of three basic emotions in music,” *Current Biology*, vol. 19, no. 7, pp. 573–576, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0960982209008136

[4] H. Egermann, N. Fernando, L. Chuen, and S. McAdams, “Music induces universal emotion-related psychophysiological responses: comparing canadian listeners to congolese pygmies,” *Frontiers in Psychology*, vol. 5, 2015. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01341

[5] M. A. Boden, “Creativity and artificial intelligence,” *Artificial intelligence*, vol. 103, no. 1-2, pp. 347–356, 1998.

[6] Y.-W. Nam and Y.-H. Kim, “Automatic jazz melody composition through a learning-based genetic algorithm,” in *Computational Intelligence in Music, Sound, Art and Design*, A. Ekárt, A. Liapis, and M. L. Castro Pena, Eds. Cham: Springer International Publishing, 2019, pp. 217–233.

[7] C. De Felice, R. De Prisco, D. Malandrino, G. Zaccagnino, R. Zaccagnino, and R. Zizza, “Chorale music splicing system: An algorithmic music composer inspired by molecular splicing,” in *Evolutionary and Biologically Inspired Music, Sound, Art and Design*, C. Johnson, A. Carballal, and J. Correia, Eds. Cham: Springer International Publishing, 2015, pp. 50–61.

[8] R. D. Prisco, D. Malandrino, G. Zaccagnino, R. Zaccagnino, and R. Zizza, “A kind of bio-inspired learning of music style,” in *Computational Intelligence in Music, Sound, Art and Design - 6th International Conference, EvoMUSART 2017, Amsterdam, The Netherlands, April 19-21, 2017, Proceedings*, ser. Lecture Notes in Computer Science, J. Correia, V. Ciesielski, and A. Liapis, Eds., vol. 10198, 2017, pp. 97–113. [Online]. Available: https://doi.org/10.1007/978-3-319-55750-2_7

[9] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, “EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation,” in *Proc. Int. Society for Music Information Retrieval Conf.*, 2021.