EmoMusic: Learning to Represent and Interact with Music via Emojis

Aman Shukla[1], Qiya Huang[2] and Gus Xia[2,3]

[1] New York University, New York
[2] New York University Shanghai, China
[3] Mohamed Bin Zayed University of Artificial Intelligence

Abstract. Both music emotion recognition and emotion-based interfaces have been popular topics of computer music research. However, very few systems enable user interactions with rich and fine-grained emotions. In this paper, we use *emojis*, a prevalent means of expressing emotions in text-based communication, as the media to convey emotion in music audio. Our EmoMusic system *learns to represent* music with emojis in a weakly supervised fashion by regarding lyric texts as the intermediate weak labels. The front end of the system is an emoji-based music *interface*. The interface not only displays both song-level and segment-level music emotions through emojis but also allows user interactions by searching songs via emojis. The whole system demonstrates the interplay between emotion and music in a simple and intuitive way and aims to enhance the listening experience of the user through emojis.

Keywords: Music interface, machine learning, music emotion, emojis

1 Introduction

In recent years, machine learning has made remarkable progress in music emotion recognition. Majority of the studies [14] focus on developing methods to extract features from audio signals and classify them into different emotion categories. Other studies [13] investigated the relationship between different acoustic features and the emotional responses they elicit in the listener. Some discoveries of these studies have been incorporated into existing music engagement platforms such as Spotify, Moodagent, Musicovery, which produce music recommendations based on emotional content of music.

However, most systems only deal with *coarse* emotion categories such as happy, sad, angry etc., and we are yet to see a system capable of representing music and interacting with users via *rich* and *fine-grained* emotions. We argue that this is mainly due to the fact that neither traditional activation-valence 2-D plane nor text descriptions are proper media for emotional sensitivity. In this

study, we resort to emoji, which is a graphical language tailored for expressing emotions. Emojis offer a broad spectrum of graphical vocabulary to represent emotional content, and in practice, they have been quite successful in conveying fine-grained emotions, popular when integrated with text-based communication. We see great potential in transferring the effectiveness of emojis into music systems.

To this end, we develop EmoMusic, an emoji-based system for music emotion understanding and interaction. Our system, as shown in Figure 1, contains three major components. The first part is a music emoji dataset, which is created by refining an off-the-shelf text-to-emoji mapping algorithm, taking music lyrics as the inputs. The second part is a machine-learning model that decodes emojis from audio with lyrics as a proxy. The third part is an emoji-based music-player interface, which not only displays music emotions at different time resolutions through emojis in real time, but also allows users to search songs via emojis.

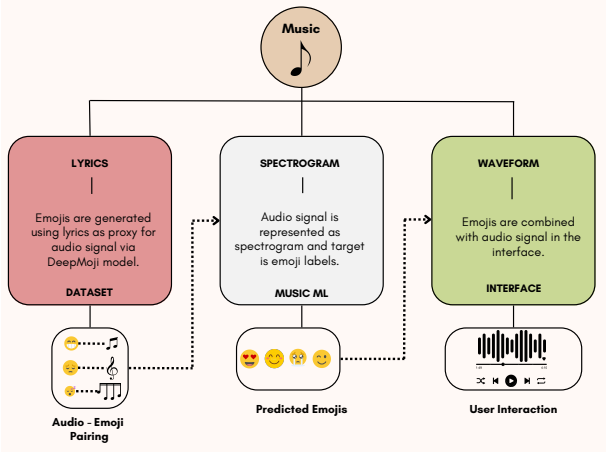

Fig. 1. A system diagram of our EmoMusic system.

In summary, the main contributions of this study are:

- A methodology to learn the mapping from music audio to emojis in a weakly-supervised fashion by regarding lyric texts as the proxy.
- A system capable of representing music with *fine-grained* emotions.
- An emoji-based music interface which not only visualizes music emotion at different time resolutions but also allows user emoji inputs for music retrieval.

The rest of the paper is organized as follows. In the next section, we discuss related works. Section 3 shows the dataset creation process, the learning algorithms, and the evaluation of music emoji estimation. Section 4 presents the interface. We conclude and discuss the future work in section 5.

2 Related Work

2.1 Music Emotion Recognition

Emotion in music has been a keen research interest of computer music researchers over the past decade. Many studies represent music emotions with a 2-D activation-valence space [6,10], using various learning models to build relation between emotion and acoustic-based and/or lyric-based features. Some prominent work such as [2] revolves around understanding sentiment in music by extracting domain-related aspects from text using an unsupervised approach. This approach relies on the fine-grained level of text analysis to create a sentiment score for each aspect category.

The study [4] talks about the different features of music and how they influence the emotions in individuals or in the society while [15] investigated the relationship between musical characteristics and the ability of some participants to recognize five basic emotions (happiness, sadness, tenderness, fear, and anger). All of them conclude that music is inherently related to sentiments.

2.2 Emojis and Emotions

The study [1] talks about how emojis can be a good measure to represent emotions, especially in text. It showcases the plethora of tasks that can be expressed and resolved through emojis representation, but so far, music hasn't been able to incorporate emojis. Hannah Miller *et. al.*[12] demonstrates that emojis is a personalized and fine-grained medium for emotion expression, as the same "happy" audio piece can be interpreted by different people in a wide spectrum of emojis (😊,😄, 😅) adding more granularity to the analysis. This enables us to personalize recommendations based on individual music tastes. Despite the significant relationships of music-emotion and emotion-emoji, there is rare evidence of music-emoji studies.

2.3 Emotion-Informed Interfaces

Patrick Helmholz *et. al.* [8] built an automatic emotion-based music recommendation system. The authors develop an approach using audio features, lyrics, and user ratings to extract emotions and make recommendations. The system uses the dataset of music tracks labeled with emotion tags and user ratings to generate recommendations based on the user's mood.

However, there are few instances where we see a combination of user interaction and emotional analysis. SmartVideoRanking [16] is one such instance, where videos are ranked based on derived emotions from the users. The emotions are estimated from time-synced user comments on videos. Based on the comments, an emotion is detected which is used to rank videos on the platform. MusicCommentrator [17], on the other hand, generates possible comments that users may input while listening to an audio clip based on appropriate temporal positions. The model learns "most commented" temporal positions in audio clips and is able to predict when a user is most likely to comment on an audio clip.

3 Emoji Estimation from Music Audio

In this section, we present how our system *learns to represent* music audio with emojis in a weakly supervised fashion by regarding lyric texts as the intermediate weak labels. The process is illustrated in Figure 2. We achieve this in 2 steps: 1) generating pseudo emojis labels for music audio by regarding lyrics as the proxy, and 2) training a model for audio-to-emojis mapping using the labels prepared in the first step. Our fundamental assumption is that the annotated lyrics of the music segment share the same emotional content as the original audio signal.

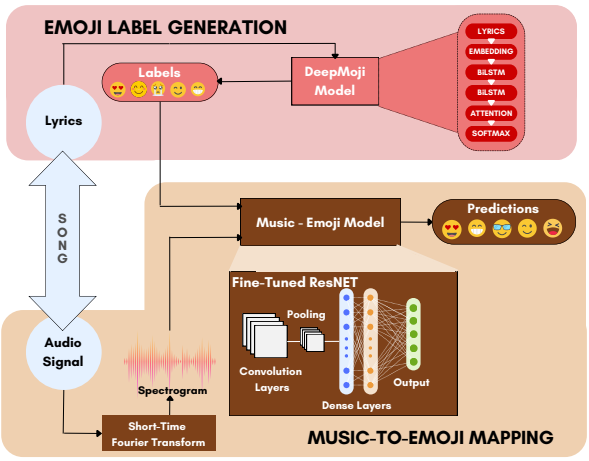

Fig. 2. The Backend of EmoMusic System: Estimating Emojis from Music Audio

3.1 Emoji Label Generation for Music Audio

3.1.1 DALI Dataset Preparation We utilized the DALI [11] dataset, which is a comprehensive collection of 5358 audio tracks along with their corresponding time-aligned vocal melody notes and lyrics at four different levels of granularity — word, sentence, paragraph, and song levels. While the dataset is rich in information, we decided to limit our analysis to paragraph-level lyrics annotations as words and sentence level annotation were not able to provide sufficient context to differentiate emotions in music segments.

3.1.2 DeepMoji Transformer for Emojis Generation To generate emojis labels for the audio tracks, we utilize the DeepMoji [5] transformer which is fed with paragraph level annotated lyrics from the DALI dataset. DeepMoji is a pre-trained transformer that has been trained on a large corpus of 1.2 billion tweets containing emojis, which allows it to understand how language is used to

express emotions. The transformer is able to extract emojis labels for the audio tracks, which serve as the basis for our supervised learning algorithm.

In order to refine the output classes and eliminate music-based emojis such as 🎵, 🎶. We refactor the output layer of the model. This refactoring has two-fold benefits. First, these emojis are not emotion representations and eliminating them improves the quality of the analysis. Second, these emojis induced skewness in the labeling procedure as the model was able to understand that the input were lyrics and most audio signals were labeled with the same 🎶 emoji.

3.2 Music-Emoji Mapping

To begin with, we use short-time Fourier transforms to convert audio signals into spectrograms. We then regard spectrogram as images and use the spectrogram-emojis pair to train our CNN-based models. This process is shown in the lower part of Figure 2, and this section presents the architecture and transfer-learning method.

3.2.1 Model Architecture Convolutional Neural Networks are known to perform well on audio tasks when audio features are represented as spectrograms [9]. Thus we select a convolutional neural network based model architecture for our task.

ResNet [7] : ResNet consists of several residual blocks stacked on top of each other. The residual block has two 3x3 convolutional layers with the same number of output channels. Each convolutional layer is followed by a batch normalization layer and a ReLU activation function. A skip connection is added which skips these two convolution operations and adds the input directly before the final ReLU activation function. The objective of the skip connections is to perform identity mapping. We use the ResNet model with 18 layers.

3.2.2 Transfer Learning by Fine-Tuning Our models are originally trained on the ImageNet [3] dataset for image classification tasks, and we fine-tune the models by training them with our new dataset of music-emoji mappings. To do so, we preprocess our data by resizing and normalizing our spectrogram images, and add a sigmoid layer on top of a fully-connected layer at the end which gives us probabilities of each target emoticon label.

3.3 Experiments and Results

Since we have no existing benchmark for the task, we use a random-choice method as our baseline. Below, we experiment with our CNN-based models both with and without transfer learning. For the model, we consider the top 5 classes with the highest probabilities as the predicted labels. We summarize the results of the experiment in Table 1.

3.3.1 Baseline: Random Choice The random chance accuracy of picking x labels from a set of n labels without repetition can be given as

$$\text{Outcomes} = \binom{n}{x} \tag{1}$$

In each pick, we can obtain k correct labels from the possible x. The favorable outcomes in this case can be written as

$$\text{Favorable} = \sum_{k=0}^{x} \binom{x}{k} . \binom{n-x}{x-k} . \frac{k}{x} \tag{2}$$

Thus the accuracy of correctly predicting emojis is,

$$\text{Accuracy} = \frac{\text{Favorable}}{\text{Outcomes}} = \frac{\sum_{k=0}^{x} \binom{x}{k} . \binom{n-x}{x-k} . \frac{k}{x}}{\binom{n}{x}} \tag{3}$$

In our data labeling procedure, we restrict our output labels to 62 target emojis and generate 5 labels for each audio segment. Thus $n = 62$ and $x = 5$.

3.3.2 Ablation Model: Training from Scratch We train the Resnet from scratch with our audio-emoji dataset instead of fine-tuning the model based on pretrained weights. The performance of the models are captured in Table 1.

3.3.3 Results The performance accuracy is calculated as the averaged percentage of the correct top 5 emojis covered by model estimation. This is in accordance with our baseline calculations.

Table 1. Accuracy(in %) of models on emoticon prediction task.

Baseline	8.06
ResNet(Ablation)	32.54
ResNet(Our Model)	**38.79**

4 Emoji-based Music Interface

We then integrate the generated emojis into an interface, aiming to leverage the power of emojis for a unique and emotionally rich music experience. The interface, as shown in Figure 3, has two major functions: 1) to search music pieces using song-level emojis (the left panel of the interface), and 2) to display both song-level and segment-level emojis of a song as it is being played (the right panel of the interface).

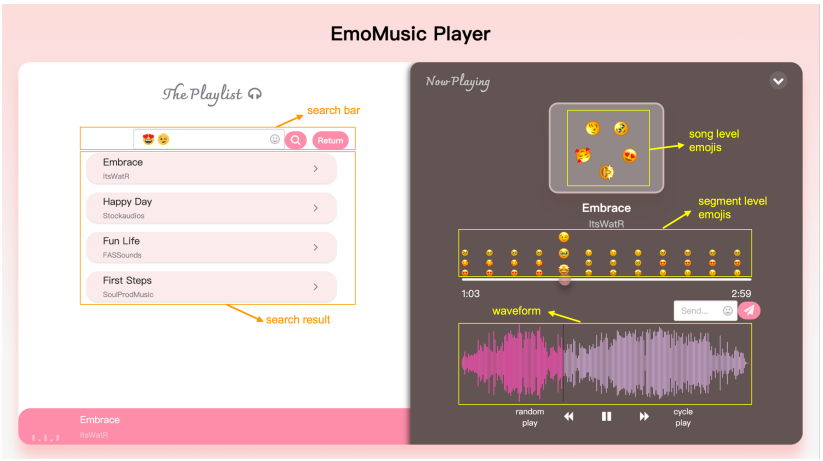

Fig. 3. The EmoMusic Interface

Each audio track in the playlist is associated with a set of five song-level emojis that rotate like a CD player. These song-level emojis serve as a general expressive indicator of the mood, intensity, or theme of the audio. The search function will show audio with song-level emojis overlapping with the emojis entered by the user. The interface also incorporates segment-level emojis that would update every 15 seconds.

5 Conclusion and Future Work

In conclusion, we developed EmoMuisc, a system that provides a unique way to explore fine-grained musical emotions with the help of machine learning and emojis. Our system utilizes a novel methodology to create a music-emoji dataset using lyrics as the proxy. Moreover, we leverage the power of transfer learning to estimate emojis from a music audio spectrograms. Lastly, the front-end interface offers users an intuitive way to explore musical emotions. We have also set a benchmark for future research in this field by demonstrating the potential of combining music and emojis.

In the future, we plan to improve the dataset creation process by exploring the potential of GPT-like models in emoji generation. Additionally, we plan to experiment with using vision transformers instead of the current CNN-based models to improve the system's performance during the learning phase. To further enhance user experience and engagement, we plan to conduct systematic user studies and integrate human-in-loop feedback where users can provide input in the form of emoticons for different audio segments.

References

1. Q. Bai, Q. Dan, Z. Mu, and M. Yang. A systematic review of emoji: Current research and future perspectives. *Frontiers in Psychology*, 10, 2019.
2. G. M. Biancofiore, T. Di Noia, E. Di Sciascio, F. Narducci, and P. Pastore. Aspect based sentiment analysis in music: A case study with spotify. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, SAC '22, page 696–703, New York, NY, USA, 2022. Association for Computing Machinery.
3. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
4. D. Deutsch, editor. *The psychology of music*. Academic Press, New York, 1982.
5. B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017.
6. B.-j. Han, S. Rho, R. Dannenberg, and E. Hwang. Smers: Music emotion recognition using support vector regression. pages 651–656, 01 2009.
7. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition, 2015.
8. P. Helmholz, M. Meyer, and S. Robra-Bissantz. Feel the moosic: Emotion-based music selection and recommendation. 06 2019.
9. S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.
10. L. Lu, D. Liu, and H.-J. Zhang. Automatic mood detection and tracking of music audio signals. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(1):5–18, 2006.
11. G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters. Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm. 2018.
12. H. Miller, J. Thebault-Spieker, S. Chang, I. Johnson, L. Terveen, and B. Hecht. "blissfully happy" or "ready tofight": Varying interpretations of emoji. *Proceedings of the International AAAI Conference on Web and Social Media*, 10(1):259–268, Aug. 2021.
13. R. Singh, H. Puri, N. Aggarwal, and V. Gupta. An efficient language-independent acoustic emotion classification system. *Arabian Journal for Science and Engineering*, 45:3111–3121, 2020.
14. Y. Song, S. Dixon, and M. Pearce. Evaluation of musical features for emotion classification. 10 2012.
15. L. Taruffi, R. Allen, J. Downing, and P. Heaton. Individual Differences in Music-Perceived Emotions: The Influence of Externally Oriented Thinking. *Music Perception*, 34(3):253–266, 02 2017.
16. K. Tsukuda, H. Masahiro, and M. Goto. Smartvideoranking: Video search by mining emotions from time-synchronized comments. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 960–969, 2016.
17. K. Yoshii and M. Goto. Musiccommentator: Generating comments synchronized with musical audio signals by a joint probabilistic model of acoustic and textual features. In S. Natkin and J. Dupire, editors, *Entertainment Computing – ICEC 2009*, pages 85–97, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.