EmoMusic: A New Fun and Interactive Way to Listen to Music

Aman Shukla[1], Gus Xia[2]

[1]New York University, New York, United States
[2]New York University Shanghai, Shanghai, China

Abstract
Modern music platforms like Spotify support users to interact with music through different interactive tools; from creating a playlist to liking or skipping a song. A prominent feature of such platforms is interacting with users by allowing them to react to music via likes and skips. Some video sharing interfaces like Niconico and Bilibili allow users to view and add overlaid commentary on videos in a synchronized fashion creating a sense of shared watching experience. However, integration of emoticons with real-time music has been rare. In this work, we propose additional channels for interacting with music through emoticons. Emoticons have been widely accepted and integrated as a medium of communication and expression, especially in text. It conveys more information than its matching text and is space optimal. We aim to integrate emoticons into an interactive music-listening interface. We believe that emoticon representation of music allows for a finer granularity in representing emotions and provides users with additional options to interact with music. We propose to build an interface which presents emoticon representation of music with basic music player functionalities.

Keywords
Music learning interface, user interface, music emotion retrieval, deep learning

1. Introduction

Recent years have witnessed exponential progress in machine learning to sentiment analysis [1, 2, 3], and music interaction [4, 5]. Despite the progress, we are yet to see an interface that combines music and emoticons. Emoticons have played a major part in the sentiment analysis domain, especially in understanding emotions from text or tweets. Emoticons also have been widely integrated into text messages and lend a meaningful value in determining context in text and natural language processing applications [6]. In our work, we build an emoji-informed interface, through which the emotion of the music is displayed in real time and users can also input their emoticons associated with a music piece. First, we developed a back-end machine learning system that decodes emoticons from audio by using lyrics as a proxy. Secondly, we design an interface which simultaneously displays audio properties with their corresponding emoticons. This interface is built on top of the back-end ML system as it uses the output emoticons from the system to display with music. Finally, we extend the interface to enable the user to interact with the music they are listening via emoticons in real-time. This additional feedback from the user interaction is then used to retrain our model and improve the performance of our machine learning system.

Our design is different from SmartVideoRanking [7] and MusicCommentrator [8], both of which estimate emotions from audio based on time-synchronized comments by the users rather than the metadata from the audio itself.

2. Methodology

Our system is designed for interactive demonstration of musical emotions through machine learning and emoticons, which contains three parts. First, preparing a fresh dataset with matching emoticon labels for a song by leveraging the underlying time-annotated lyric representation. Second, training a machine learning model via supervised learning by representing audio signals as corresponding spectrograms and using the generated emoticon symbols as target labels. Finally, we integrate this system into our interactive display which utilizes the music-emoticon pair as a starting point and follows up by enabling users to interact with music by selecting their preferred emoticons. We dive into some of the details of each of these sections below and present an outline of the system in Figure 1.

2.1. Dataset Creation

We have used the DALI dataset [9] which is a large and rich multimodal dataset containing 5358 audio tracks with their time-aligned vocal melody notes and *lyrics* along with other meta-data. Although the dataset contains rich information, for our setup we've only considered English songs and paragraph level lyrical annota-

Joint Proceedings of the ACM IUI Workshops 2023, March 2023, Sydney, Australia
as14034@nyu.edu (A. Shukla); gxia@nyu.edu (G. Xia)

tions. The decision to incorporate paragraph level annotations stemmed from our analysis where we found the corresponding context was necessary to derive a sentiment from the audio. This analysis was done under the assumption that music segments and their annotated lyrics share the same emotion content that can be effectively represented by emoticons. The lyrics are passed to the fine-tuned DeepMoji [10] transformer to extract emoticons labels for the piece. We fine-tune the output classes by eliminating music based symbols as they do not represent any emotion.This subprocess is represented in Figure 1 titled *Labeling*. Finally from this process we are able to generate an audio-emoticon pairing which will serve as a basis for our supervised learning algorithm.

2.2. Musical Machine Learning Model

We first transform audio signals to spectrograms via short-time fourier transforms. It has been studied before that spectrograms offer a rich representation of audio [11]. Then, we utilize the spectrogram-emoticon pair to train our model. Since we've represented audio signals as spectrograms, we leverage transfer learning to extract information.

2.2.1. Transfer Learning

Transfer Learning in audio has been mainly focussed on pretraining a model on a large corpus of audio datasets. We follow an approach similar to [12] where we leverage the power of transfer learning shifting focus from audio datasets to image datasets. We train DenseNet [13] and ResNet [14] which are convolutional neural network (CNN) architectures trained on the ImageNet [15] dataset.

2.2.2. Fine-Tuning

Both DenseNet and ResNet are *fine-tuned* to predict 62 emoticons. To accomplish this, we add a fully connected layer followed by a sigmoid layer to obtain class probabilities. This subprocess is represented in Figure 1 titled *Training*.

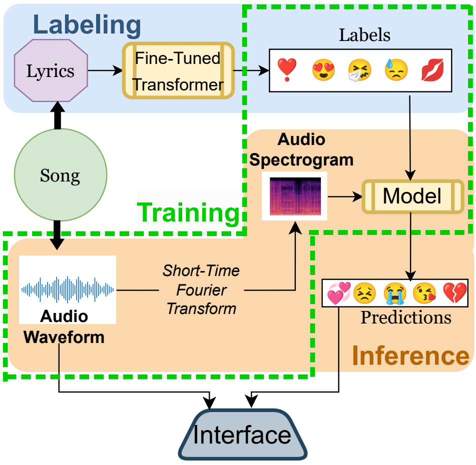

Figure 1: ML-based backend system showing data preparation, model training and inference.

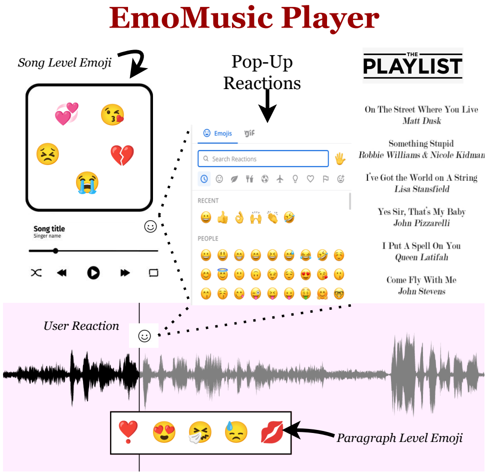

Figure 2: Interface with real-time emoticon representation of audio alongside user enabled reactions.

3. Interface Design

In our proposed integrated music player design, we aim to build a web based music player that includes basic features like Play/Pause, Next/Previous, Playlist etc. When a song is playing, we intend to display the corresponding audio waveform, current emoticon representation (derived from the paragraph based annotation), and song level emoticon representation. In addition, we enable the user to interact with the music by selecting emoticons. We propose to build a two-level interaction; with a song and with real-time audio playback. The emoticon icon opens a pop up for users to input their selection of emoticons. The pop up icon on the audio waveform section inputs user interaction for real-time feedback while the pop-up icon embedded in the player(horizontally to the song title) provides a song level emoticon feedback. From these user interactions, we intend to improve the model's performance by re-training our model. A visualization of the music player is shown in Figure 2.

References

[1] L. M. Gómez, M. N. Cáceres, Applying data mining for sentiment analysis in music, in: Trends in Cyber-Physical Multi-Agent Systems. The PAAMS Collection - 15th International Conference, PAAMS 2017, Springer International Publishing, Cham, 2018, pp. 198–205.

[2] G. M. Biancofiore, T. Di Noia, E. Di Sciascio, F. Narducci, 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, Association for Computing Machinery, New York, NY, USA, 2022, p. 696–703. URL: https://doi.org/10.1145/3477314.3507092. doi:10.1145/3477314.3507092.

[3] L. Taruffi, R. Allen, J. Downing, P. Heaton, Individual Differences in Music-Perceived Emotions: The Influence of Externally Oriented Thinking, Music Perception 34 (2017) 253–266. URL: https://doi.org/10.1525/mp.2017.34.3.253. doi:10.1525/mp.2017.34.3.253.

[4] J. Smith, D. Weeks, M. Jacob, J. Freeman, B. Magerko, Towards a hybrid recommendation system for a sound library, in: C. Trattner, D. Parra, N. Riche (Eds.), Joint Proceedings of the ACM IUI 2019 Workshops co-located with the 24th ACM Conference on Intelligent User Interfaces (ACM IUI 2019), Los Angeles, USA, March 20, 2019, volume 2327 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2327/IUI19WS-MILC-5.pdf.

[5] V. Thio, H. Liu, Y. Yeh, Y. Yang, A minimal template for interactive web-based demonstrations of musical machine learning, CoRR abs/1902.03722 (2019). URL: http://arxiv.org/abs/1902.03722. arXiv:1902.03722.

[6] H. Miller, J. Thebault-Spieker, S. Chang, I. Johnson, L. Terveen, B. Hecht, "blissfully happy" or "ready tofight": Varying interpretations of emoji, Proceedings of the International AAAI Conference on Web and Social Media 10 (2021) 259–268. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/14757. doi:10.1609/icwsm.v10i1.14757.

[7] K. Tsukuda, H. Masahiro, M. Goto, Smartvideoranking: Video search by mining emotions from time-synchronized comments, in: 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), 2016, pp. 960–969. doi:10.1109/ICDMW.2016.0140.

[8] K. Yoshii, M. Goto, Musiccommentator: Generating comments synchronized with musical audio signals by a joint probabilistic model of acoustic and textual features, in: S. Natkin, J. Dupire (Eds.), Entertainment Computing – ICEC 2009, Springer Berlin Heidelberg, Berlin, Heidelberg, 2009, pp. 85–97.

[9] G. Meseguer-Brocal, A. Cohen-Hadria, G. Peeters, Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm., Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR, Paris, France (2018) 431–437. doi:10.5281/ZENODO.1492443.

[10] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, 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. URL: https://doi.org/10.18653%2Fv1%2Fd17-1169. doi:10.18653/v1/d17-1169.

[11] L. Wyse, Audio spectrogram representations for processing with convolutional neural networks, CoRR abs/1706.09559 (2017). URL: http://arxiv.org/abs/1706.09559. arXiv:1706.09559.

[12] K. Palanisamy, D. Singhania, A. Yao, Rethinking CNN models for audio classification, CoRR abs/2007.11154 (2020). URL: https://arxiv.org/abs/2007.11154. arXiv:2007.11154.

[13] G. Huang, Z. Liu, K. Q. Weinberger, Densely connected convolutional networks, CoRR abs/1608.06993 (2016). URL: http://arxiv.org/abs/1608.06993. arXiv:1608.06993.

[14] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, CoRR abs/1512.03385 (2015). URL: http://arxiv.org/abs/1512.03385. arXiv:1512.03385.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.