Music Emotions Classification: A General Model Fine-Tuned On User

Francesca Gasparini[0000−0002−6279−6660] and Mattia Marchi

Department of Informatics, Systems and Communication,
University of Milano-Bicocca, Milano, Italy
francesca.gasparini@unimib.it
m.marchi@campus.unimib.it

Abstract. This paper presents an approach for music emotion recognition (MER) using deep learning algorithms. MER systems nowadays can have different uses, from the generation of customized playlists based on the user's mood to the musical suggestions implemented by musical streaming platforms. The proposed model consists in a state-of-the-art Convolutional Neural Network (CNN) to extract the main features from Mel spectrograms, generated by music files, and a Long Short-Term Memory (LSTM) as a classifier. The class of emotions examined corresponds to Russell's four-quadrant. We have merged three datasets from the literature to create a model with a better generalization, achieving 91% accuracy. We also studied the ability of the model to adapt to the individual user, implementing the possibility of fine-tuning the network on the emotions perceived by the user. The main objective is to create a model that is able to properly generalize and then to adapt it to the perceptions of each individual user, in order to consider the subjectivity of perception. The system was tested on three different users using 50 songs, showing an accuracy of predicted emotions of 44.68% for the general model. User-tuned models that consider 10 and 20 user-suggested emotions reached an average of correctly predicted songs of 62.66% and 76% respectively, greatly improving the general model results and gradually decreasing the average prediction error.

Keywords: Music Emotion Recognition · Affective Computing · Deep Learning · Convolutional Neural Network · Long Short-Term Memory.

1 Introduction

Users increasingly need contents filtered according to their interest. With this aim, several Recommender Systems (RS) have been developed, revolutionizing the way people enjoy the contents [12]. Music is one of those fields where recommender systems have been recently applied. There is already a significant number of research works for music RS, using different types of data: audio signals, content-based information and ratings from users [4]. A further step in research is to improve recommender systems using the user's emotions. The Music Emotion Recognition (MER) has become a fundamental aspect to guarantee

the user a better subjective experience. MER systems have a large number of possible applications, such as automatic playlist generation and music suggestion application (Spotify, SoundCloud, ...). In this paper, we explore the possibility of creating a generalized model to recognize the mood provided by a song, using its Mel spectrogram [5], and then adapt this model to the single user. To adapt the general model to the user, some of his/her explicit emotions referred to a dataset of songs are used to fine-tune the model on his/her emotional responses. The proposed approach is based on a Convolutional Neural Network (CNN) [13] combined with a Long Short-Term Memory (LSTM) [9]. There are several ways to classify emotions induced by a song, which can be grouped into two models: categorical and dimensional. Categorical models divide emotions into several groups of adjectives. One of the most used approaches has been introduced by Plutchick *et al.* [19], who divides emotions into 8 categories represented by a wheel of emotions. Dimensional approaches, on the other hand, label songs based on their position in a dimensional space. The most used representation, proposed by Russell [20], divides the space into three dimensions: arousal, valence and dominance. The model adopted in our work considers two dimensions (valence and arousal), classifying each song into one of the Russell's four-quadrants of this 2D space. This paper is organized as follows. Section 2 introduces the state of the art of automatic music emotion recognition. In section 3 the three datasets used in this work are described. In section 4 the general MER model is presented and our user-tuned proposal is described. The experimental results are discussed in section 5. Eventually, conclusions and possible future works are included in section 5.

2 State Of The Art

There are several approaches in the literature to model a MER system. The first of these approaches uses audio features directly obtained from the audio files and supplies them to a classifier. For instance, Han *et al.* developed a multi-class Support Vector Machine classifier using only low-level features such as energy, zero crossing rate (ZCR), audio spectrum and wavelet [7]. The songs were classified on a wide range of emotional categorical classes (11 labels), such as "happy", "peaceful" and "bored". The study was performed using a very limited dataset of 120 songs. The results obtained showed an overall accuracy of 67.54%. A different point of view is provided by Liu *et al.* through the combined use of LSTM, to classify the audio signals, and BERT, to obtain an estimate of the mood through the lyrics of the song [14]. LSTMs have empirically shown an improvement in performance on audio signals compared to other machine learning models. On the other hand, BERT is a multi-layer bidirectional transformer encoder, developed by Google AI team in 2018 [6], which allows to classify the emotions of the lyrics. The use of lyrics allows to analyze with more precision what a song wants to communicate and therefore what mood it arouses. The model, trained on a set of 1000 songs, achieves an average accuracy of 79.70%.

With research improvements and thanks to some famous contests like *ImageNet* [13], the CNNs have become more and more popular in music classification. CNN networks allow to use images generated from audio files, as the representation of signals, and to classify them with excellent performance results. Bahuleyan in 2018 has applied a CNN on images representing the Mel Spectrograms of the songs, with the aim of categorizing music files according to their genre [3]. The CNN architecture used is the VGG-16 [21], one of the top performing model in *ImageNet* Challenge in 2014. The dataset used contained a large number of audio files (40540), divided into 7 genres. The study tested the pre-trained model in two different settings: transfer learning and fine-tuning. The best accuracy was achieved from the model fine-tuned, with an accuracy of 64%. In 2020, Hizlisoy *et al.* studied how to classify the emotions aroused by listening to Turkish traditional songs, using the features obtained from a CNN and classifying them using a LSTM plus a Deep Neural Network (DNN) [8]. The Turkish music database is composed of only 124 excerpts with a duration of 30 seconds. The music was annotated on 3 class emotions, based on valence and arousal. The performance of the combined LSTM + DNN model, applied to features extracted from CNN, has achieved an overall accuracy of 99.19 %.

3 Datasets

In order to achieve a better generalization, the proposed model was obtained by combining three different datasets used in the MER field. The choice was made using three constraints: the audio files should be available, each excerpt should last at least 30 seconds and an emotional label should be associated with each excerpt. The three selected datasets are: *PMEmo*[25], *4Q*[17][18] and *Emotify*[1][2]. In order to use the same emotional model for all datasets, the different annotations were properly converted into 4 classes corresponding to Russell's quadrants. The notation reported in table 1 was used to label these quadrants.

Table 1: Notation used to label Russell's quadrants.

Arousal	**Valence**	**Label**	**Quadrant**
Low	Negative	0	A-V-
Low	Positive	1	A-V+
High	Negative	2	A+V-
High	Positive	3	A+V+

3.1 PMEmo

PMEmo[1] is a dataset publicly available for the research community, containing 794 songs, annotated by 457 subjects [25]. In addition to the audio files, it also contains annotations relating to the physiological response of the participants listening to the songs. The 794 audio files correspond to the chorus of the first songs in several world listening music charts in 2016. In detail:

- 487 songs from *Billboard Hot 100*;
- 616 songs from *iTunes Top 100*;
- 226 songs from *UK Top 40 singles*.

The annotations related to the choruses have been released as a value between $[0, 1]$ of arousal and valence. The conversion performed is the following:

- $(Arousal \leq 0.5) \wedge (Valence \leq 0.5) \rightarrow Label = 0$
- $(Arousal \leq 0.5) \wedge (Valence > 0.5) \rightarrow Label = 1$
- $(Arousal > 0.5) \wedge (Valence \leq 0.5) \rightarrow Label = 2$
- $(Arousal > 0.5) \wedge (Valence > 0.5) \rightarrow Label = 3$

In order to manage the choruses of variable length they have been converted to a fixed length of 30 seconds. The dataset is unbalanced, with a large number of songs with valence and arousal greater than 0.5. For this reason, only a subset of 150 audio files have been used, in order to avoid an overly unbalanced training set.

3.2 4Q

4Q dataset[2] contains 900 audio clips, with subjective annotations following Russell's emotion quadrants [17][18]. Each clip has a duration of 30 seconds and files are organized in 4 folders, each containing 225 files, corresponding to the four quadrants. The songs are obtained by querying and filtering the *AllMusic* API[3]. The dataset is balanced, so all the excerpts are used in the final merged dataset.

3.3 Emotify

Emotify dataset[4] contains 400 songs, divided into 4 different musical genres: classical, electronic, pop and rock. The annotations were obtained through a game in which the participants labeled the emotions induced by the songs of different genres. Unlike previous datasets, *Emotify* uses categorical emotion labeling, following the GEMS scale [24]. For each song, the number of votes for each emotion category are provided. We therefore used the emotion with the

[1] https://github.com/HuiZhangDB/PMEmo
[2] http://mir.dei.uc.pt/downloads.html
[3] http://developer.rovicorp.com
[4] http://www2.projects.science.uu.nl/memotion/emotifydata/

highest number of votes as a single label. In order to convert categorical emotions into the valence arousal model, we used the scheme reported in figure 1. After this conversion, the dataset was extremely unbalanced, with most labels in the quadrant having low arousal and positive valence. Even in this case only 150 audio files were used in order to keep balanced the final dataset.

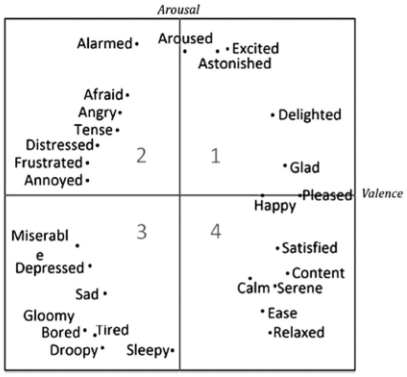


Fig. 1: Russell circumplex with categorical emotions [18].

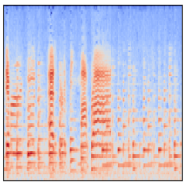
3.4 Preprocessing

Once the merged dataset is obtained, the final number of audio files is 1200 of 30 seconds duration. Each file is divided into 5 frame of 6 seconds, obtaining 6000 frames. On these frames the MEL spectrogram is computed. The choice to use spectrograms was made because spectrograms are inclusive of para-lingual information and therefore useful for the tasks of recognizing emotions [15]. The *Librosa* python library was used to generate the spectrograms [16]. An example of the generated spectrograms is shown in figure 2.

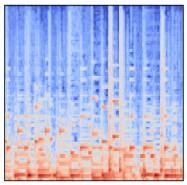
The parameters used for the generation of the Mel spectrograms are:

- **Sample Rate:** 22050;
- **Frame/Window size:** 2048;
- **Hop Size:** 512;
- **Number of Mel bins:** 96;
- **Size of Image:** 300 x 300 pixels.

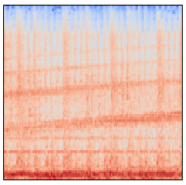
The final dataset has been split into train, test and validation set with the respective percentage of 70%, 15% and 15%. The split is done using the IDs of the audio file before the generation of the Mel spectrograms, in order to avoid having different frames of a song in different sets.



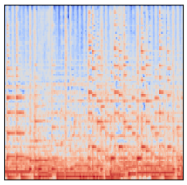
(a) Low Arousal - Negative Valence



(b) Low Arousal - Positive Valence



(c) High Arousal - Negative Valence



(d) High Arousal - Positive Valence

Fig. 2: Example of 4 different spectrograms corresponding to audio files of different emotional classes.

4 MER Models

The model we propose to solve the general music emotion recognition task consists of two fundamental components:

- Convolutional Neural Network (CNN);
- Long Short-Term Memory (LSTM).

These models are described in section 4.1 and 4.2 respectively. In particular we fine tuned pre-trained CNNs in order to learn characteristics of Mel spectrograms, relevant in a MER task. Then, from these CNNs, features are extracted to feed a classifier based on LSTM. Staring from this general model, we here propose to user-tuned it in order to better fit the emotions of each single user. Our proposal is described in section 4.3.

4.1 Convolutional Neural Network

The first model adopted to classify emotions using Mel spectrograms is a 3D convolutional network. This network is capable of understanding some characteristic patterns belonging to different emotional classes, considering the spectrograms as images [3]. Two different state of the art networks are here considered and compared, and fine-tuning is applied with respect to the MER task:

- **EfficientNet-B3** [22]: this network presented in 2019 uses a scaling method that uniformly scales all dimensions of depth, width and resolution.
- **MobileNetV3-Large** [10]: Lightweight state of the art CNN with one of the best ratios between accuracy and speed of inference.

Both models use ImageNet starting weights and the following layers have been introduced:

- Flatten layer;
- Dense layer of 1024 output nodes, ReLu activation and L2 regularization with $\alpha = 0.1$;
- Dropout layer with dropout rate of 0.5;
- Dense layer of 512 output nodes, ReLu activation and L2 regularization with $\alpha = 0.1$;
- Dropout layer with dropout rate of 0.4;
- Dense layer of 128 output nodes, ReLu activation and L2 regularization with $\alpha = 0.1$;
- Dropout layer with dropout rate of 0.3;
- Dense layer of 4 output with softmax activation.

The objective function minimizes the *sparse categorical crossentropy loss*, and is optimized with *Adam* [11], with a training phase of 50 epochs and a batch size of 32 instances. In order to have the model with the best ability to generalize, we use the model weights that have obtained the best accuracy score on the validation set in the epochs. As shown in table 2, the model based on MobileNetV3 has a slightly lower accuracy, but with an average training time of about one third of EfficientNet-B3. Ideally, the application of recommender systems that we are studying should be able to run on users' mobile devices, so it is important to take into account the weight of the model. For this reason the network chosen in the final model is MobileNetV3.

Table 2: Performance of the CNN models.

Model	**Accuracy**	**F-score**	**Average Training Time per Epoch**
EfficientNet-B3	0.69	0.67	90s
MobileNetV3	0.60	0.60	32s

4.2 Long Short-Term Memory

Once CNNs have been trained to understand the relevant characteristics of Mel spectrograms they can be used to perform features extraction. The model used as classifier is the Long Short-Term Memory (LSTM) [9]. LSTM is widely used in music recognition tasks due to its ability to process sequential information. To have the network capable of making inference on a song portion 24 seconds

long, we consider 4 spectrograms, each corresponding to 6 seconds. We feed the pre-trained MobileNetV3 with these spectrograms and we extract the corresponding features from the last convolutional layer, flattened and reshaped to be considered a single input of 4 sequences for the LSTM, as shown in figure 3.

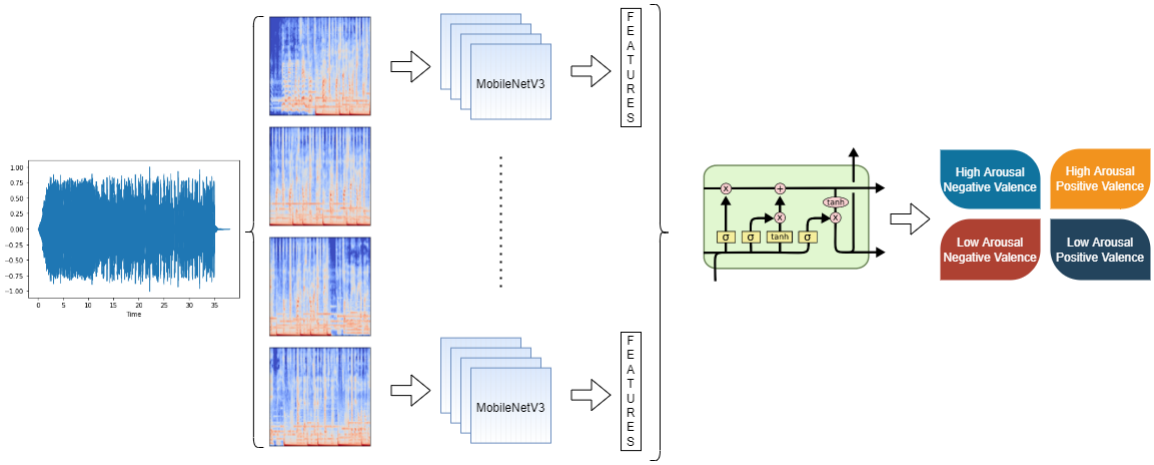


Fig. 3: Features are extracted using CNN from the input spectrograms and reshaped to be a single input of 4 sequences for LSTM.

Specifically, the LSTM network is composed of:

- LSTM layer of 128 hidden units, ReLu activation;
- Dropout layer with dropout rate of 0.2;
- Dense layer of 4 output with softmax activation.

The objective function minimises the *sparse categorical crossentropy loss*, and is optimized with *RMSprop* [23], with a training phase of 20 epochs and a batch size of 32 instances. As in the case of CNN, we use the model weights that have obtained the best accuracy score on the validation set in the epochs. The network achieves excellent results, with an accuracy and F-score of 0.91%. As shown in the confusion matrix in figure 4 there are very few wrong classifications.

4.3 User-Tuned Model

The main purpose of this paper is not only to find a model that is able - in general - to understand the emotions induced by the music, but also to adapt it to the individual user in case he had a different way of perceiving songs. The idea is to simulate a real context of integration of our model in an existing playlist generation system. In order to solve this problem, 50 new songs have been chosen from different genres and of variable length, with an average length

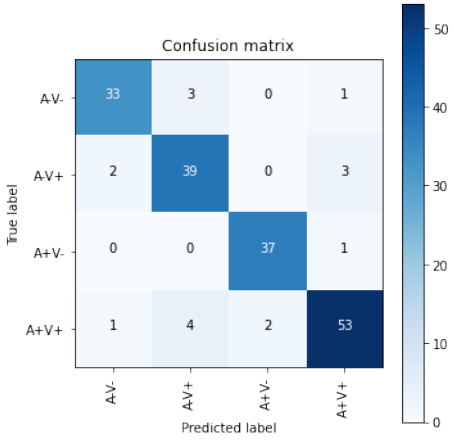


Fig. 4: Confusion matrix obtained on the test set.

of 280 seconds. The songs were chosen from the most famous ones for each genre, in order to simulate a real playlist of a user. The songs have been played for 3 different users and individually labeled, asking them to select the emotion felt from the quadrant of the arousal-valence model, as previously reported in figure 1. The user could also explicitly indicate some emotions, which he/she wants the model to be aware of. Starting from the data labelled by each subject, our proposed user-tuned classifier can be described as follows. The spectrograms, corresponding to the entire songs are generated (evaluated for each 6-second frame), the features are extracted, they are labeled with the user's emotion and the model is user-tuned using the original dataset plus a portion of the new labeled data. The user-tuning of the classifier is applied starting from the pre-trained LSTM but with a very small learning rate (0.000001) through 50 epochs. This value was chosen empirically. The main purpose is to create a controlled overfitting in order to push the model towards the user's perception. This desired controlled overfitting effect is also obtained using a portion of the new generated spectrograms as validation set, in order to verify whether the network has learned the user's suggestion. The model will use the weights that have obtained the best accuracy on the validation set. The low learning rate is used to avoid excessive variation in the general model's weights and to avoid uncontrolled overfitting. The network will be able to recognize the emotions of the songs suggested by the user while remaining able to generalize. Applying this strategy, we obtained an average accuracy on the validation set of 89%. The user-tuning process is outlined in figure 5.

5 Results and Discussion

The proposed user-tuned model is evaluated on a new dataset of 50 songs, obtained by asking 3 subjects to listen to the songs, label them and indicate 10 and

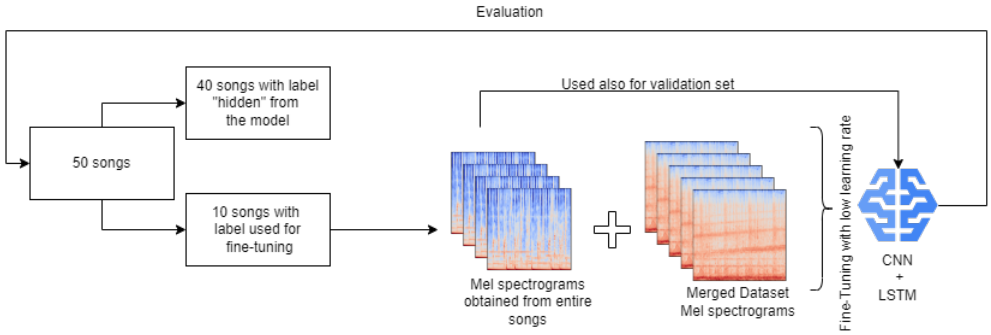


Fig. 5: Summary of the user-tuning process.

20 songs whose induced emotions he/she wants the model to be aware of. We thus create three user-tuned classifiers, one for each subject involved. We also compare two versions of the proposed classifiers, considering respectively the 10 and 20 suggested songs in the user-tuning step. Finally the performance of these user-tuned models are also compared with the general MER model previously described. To this end, each of the 50 new labelled songs are divided into frames of 24 seconds, to be classified by the LSTM. Thus for each song we obtain several emotional classes, one for each 24-second segment. Starting from the labels obtained for each segment, the final label is thus reported in terms of frequency (or percentage) of each of the 4 emotions, with respect to the whole song. A Python script was then created which, given as input the *csv* file containing the user labels and the folder with the audio files, performs the following steps:

1. Mel spectrograms are generated for 6 seconds of each song;
2. Features are extracted from CNN as described in section 4.1 for each 24 seconds of each song, considering 4 spectrograms ;
3. The features are reshaped and used as input to LSTM as described in section 4.2;
4. LSTM generates several predictions, one for each 24-second sequence;
5. A percentage estimate of all predictions is made, in order to consider the various mood changes within the song itself.

A diagram of the algorithm is shown in figure 7. To quantify the performance of the classification we do not only consider the most frequent emotion, but we also introduce an error metric to quantify the distance between the results of the classification and the subject's labels, when the subject label has not been properly identified. This metric can be useful to understand how wrong the model is in the prediction. The proposed metric computes the difference between the frequency of the predicted emotion (the most frequent one in the classification output) and the frequency obtained for the true emotion class. Figure 6 reports an example for 5 songs (rows in the table) of the output of the user-tuned

classifier, in terms of frequency of each emotional class, the corresponding class predicted by the classifier (column Max_Overall_Emotion), compared with the true label (User_Emotion) and finally evaluating the error metric here proposed (Difference).

	A-V-	A-V+	A+V-	A+V+	Max_Overall_Emotion	User_Emotion	Wrong_Classification	Difference
0	0.0	100.0	0.0	0.0	1	1	False	0.0
1	60.0	40.0	0.0	0.0	0	0	False	0.0
2	11.1	11.1	11.1	66.7	3	0	True	55.6
3	14.3	0.0	14.3	71.4	3	1	True	71.4
4	12.5	0.0	62.5	25.0	2	1	True	62.5

Fig. 6: Example of the result obtained on 5 songs after the execution of the model. The first 4 columns report the *True_Emotions* expressed by the user, with respect to the 4 quadrants of the arousal-valence plane. The corresponding quadrant is reported in column *User_Emotion*. Column *Max_Overall_Emotion* refers to the emotion quadrant with the highest percentage found by the model. If the two columns do not match, the prediction is considered wrong and the difference in estimates is calculated.

The results obtained on the 50 songs, for each user and on average, are summarized respectively in table 3 and table 4. These results demonstrate how the user-tuned model achieves an improvement in recognizing the true emotions, while decreasing the error on wrong predictions. Experiments have shown that the songs used for user-tuning are, as expected, always correctly classified, thus the difference error is calculated only on the songs not used in this process. For this reason, this metric is significant in evaluating the performance of the model.

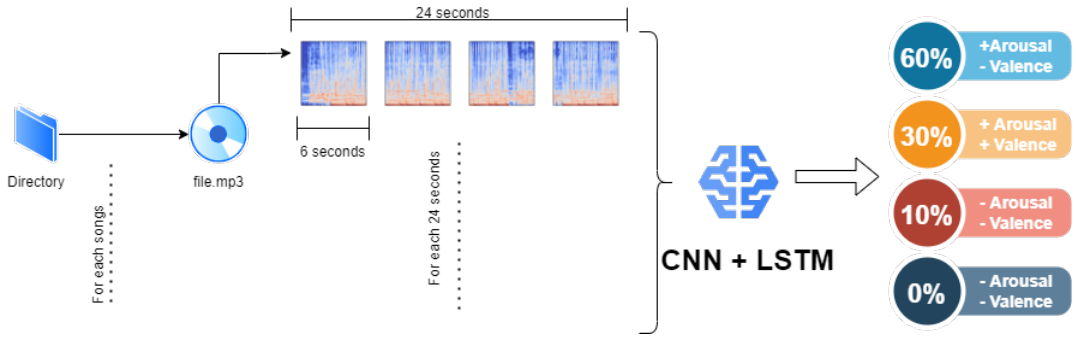


Fig. 7: Summary of the final model applied to entire songs.

Table 3: Evaluation based on three subjects.

Subject	User-Tuning	Correct classification	Accuracy %	Average Difference Error
1	No Emotions Suggested	24	0.48	40.03
	10 Emotions Suggested	36	0.72	34.92
	20 Emotions Suggested	41	0.82	25.29
2	No Emotions Suggested	23	0.46	47.73
	10 Emotions Suggested	31	0.62	37.88
	20 Emotions Suggested	38	0.76	31.80
3	No Emotions Suggested	20	0.40	45.83
	10 Emotions Suggested	27	0.54	32.20
	20 Emotions Suggested	35	0.70	34.43

Table 4: Average values for suggested emotions.

User-Tuning	Correct classification	Accuracy %	Average Difference Error
No Emotions Suggested	22.34	0.45	44.62
10 Emotions Suggested	31.33	0.63	35
20 Emotions Suggested	38	0.76	30.51

6 Conclusions and Future Work

In this paper we have focused on the possibility of applying a general MER system to individual users, obtaining a more functional personalized experience. The results obtained show the possibility to apply deep learning models to recognize emotions and adapt them through user-tuning. The proposed approach consists in using a state of the art CNN to extract the features from Mel spectrograms and a LSTM network as a classifier. The model could easily be integrated into a music recommender system, in order to create personalized mood-based playlists. The playlist would then be based on the user's true emotions and no longer on general labels made by third parties, improving the user experience. The results obtained are promising, however it is necessary to extend the study by considering more songs and applying the model to real users playlist. A further study could be done in the field of wearable computing, using bracelets for real-time recognition of the user's emotions. Through this system, the need to explicitly ask the user for the emotion felt in listening could be removed and recognized directly from the device, providing better data for user-tuning.

References

1. Aljanaki, A., Wiering, F., Veltkamp, R., et al.: Collecting annotations for induced musical emotion via online game with a purpose emotify. Technical Report Series **2014**(UU-CS-2014-015) (2014)
2. Aljanaki, A., Wiering, F., Veltkamp, R.C.: Studying emotion induced by music through a crowdsourcing game. Information Processing & Management **52**(1), 115–128 (2016)

3. Bahuleyan, H.: Music genre classification using machine learning techniques. arXiv preprint arXiv:1804.01149 (2018)
4. Batmaz, Z., Yurekli, A., Bilge, A., Kaleli, C.: A review on deep learning for recommender systems: challenges and remedies. Artificial Intelligence Review **52**, 1–37 (2018)
5. Demircan, S., Örnek, H.K.: Comparison of the effects of mel coefficients and spectrogram images via deep learning in emotion classification. Traitement du Signal (2020)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Han, B.j., Rho, S., Jun, S., Hwang, E.: Music emotion classification and context-based music recommendation. Multimedia Tools and Applications **47**(3), 433–460 (2010)
8. Hizlisoy, S., Yildirim, S., Tufekci, Z.: Music emotion recognition using convolutional long short term memory deep neural networks. Engineering Science and Technology, an International Journal **24**(3), 760–767 (2021)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
10. Howard, A., Sandler, M., Chu, G., Chen, L.C., Chen, B., Tan, M., Wang, W., Zhu, Y., Pang, R., Vasudevan, V., et al.: Searching for mobilenetv3. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1314–1324 (2019)
11. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
12. Konstan, J.A., Riedl, J.: Recommender systems: from algorithms to user experience. User Model. User-Adap. Inter. **22**(1), 101–123 (Apr 2012). https://doi.org/10.1007/s11257-011-9112-x
13. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
14. Liu, G., Tan, Z.: Research on multi-modal music emotion classification based on audio and lyirc. In: 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC). vol. 1, pp. 2331–2335. IEEE (2020)
15. Ma, X., Wu, Z., Jia, J., Xu, M., Meng, H., Cai, L.: Emotion recognition from variable-length speech segments using deep learning on spectrograms. In: Interspeech. pp. 3683–3687 (2018)
16. McFee, B., Raffel, C., Liang, D., Ellis, D.P., McVicar, M., Battenberg, E., Nieto, O.: librosa: Audio and music signal analysis in python. In: Proceedings of the 14th python in science conference. vol. 8, pp. 18–25. Citeseer (2015)
17. Panda, R., Malheiro, R., Paiva, R.P.: Musical texture and expressivity features for music emotion recognition. In: 19th International Society for Music Information Retrieval Conference (ISMIR 2018). pp. 383–391 (2018)
18. Panda, R., Malheiro, R., Paiva, R.P.: Novel audio features for music emotion recognition. IEEE Transactions on Affective Computing **11**(4), 614–626 (2018)
19. Plutchik, R., Kellerman, H.: Theories of emotion, vol. 1. Academic Press (2013)
20. Russell, J.: A circumplex model of affect. Journal of Personality and Social Psychology **39**, 1161–1178 (12 1980). https://doi.org/10.1037/h0077714
21. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

22. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
23. Tieleman, T., Hinton, G.: Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. COURSERA Neural Networks Mach. Learn (2012)
24. Zentner, M., Grandjean, D., Scherer, K.R.: Emotions evoked by the sound of music: characterization, classification, and measurement. Emotion **8**(4), 494 (2008)
25. Zhang, K., Zhang, H., Li, S., Yang, C., Sun, L.: The pmemo dataset for music emotion recognition. In: Proceedings of the 2018 acm on international conference on multimedia retrieval. pp. 135–142 (2018)