Music Mood Annotation- a Deep Learning Approach

Milind Kumar V[1]

Abstract— **Automatic annotation of music with emotional labels is a challenging task owing to the subjectivity of emotions associated with music. Much of this work focuses on creating a sizeable dataset with sufficient annotations that tackles the issue of subjectivity adequately. By combining pre-existing datasets with discrete and continuous emotional labels, a dataset with 79 hours worth of audio is obtained. All the audios are represented as points in the valence-arousal plane and are then clustered into four classes for the purpose of developing classifiers. CNNs which use mel spectrograms as inputs are the primary focus of this work. A maximum accuracy of 50.89% is obtained in the four-class classification task with accuracies in the individual tasks of valence and arousal classifications being 70.8% and 71.7% respectively.**

I. INTRODUCTION

The inexorable growth of the internet and the increased availability of and ease of access to music have made necessary the development of efficient systems for organization of music. This has led to much work in the area of Music Information Retrieval. This work addresses the task of automatically determining the mood (used interchangeably with emotion) of a given track for the purpose of classification and retrieval from a sizeable dataset.

Emotion is a very intuitive basis to categorize music on as much of music expresses some form of it and most listeners tend to experience some emotion when listening to music. Whilst being intuitive, the subjectivity of the emotion perceived or induced is a major challenge to mood annotation of music. Therefore, predicting the emotion induced in a listener is an intriguing task albeit very challenging as individual listeners experience very different emotions when listening to the same song. The emotion induced depends on sex, age, culture, background, context, present mood and more. Thus, this work pursues the relatively easier task of predicting the perceived emotion i.e the emotion seen as being conveyed by music which generally has greater agreement amongst annotators. Further this choice is pragmatic as a vast majority of datasets primarily offer annotations of perceived emotion.

A deep learning approach using convolutional neural networks (CNNs), which have worked remarkably well in image classification tasks and Acoustic Event Detection [1] (AED henceforth), is employed and this necessitates the collection of large amounts of data. While datasets such as Audioset offer a significant amount of weakly labeled data, correspondingly large amounts of processing power and time are required to train models that produce meaningful results on such data. Thus, a compromise is made between quantity and quality and datasets designed specifically for music mood annotation are used for the experiments. These include the LastFM100 corpus, Yang60 corpus, DEAM, CAL500, Emotify, Jyvaskyla Soundtracks and the Moodo datasets all of which are publicly available. Two datasets created internally by Fraunhofer IDMT are also used for this task. A total of 3329 audio tracks of varying length amounting to 79.27 hours of data with song level annotations are available for training.

The use of multiple datasets requires the use of a uniform representation of emotions which is either categorical or

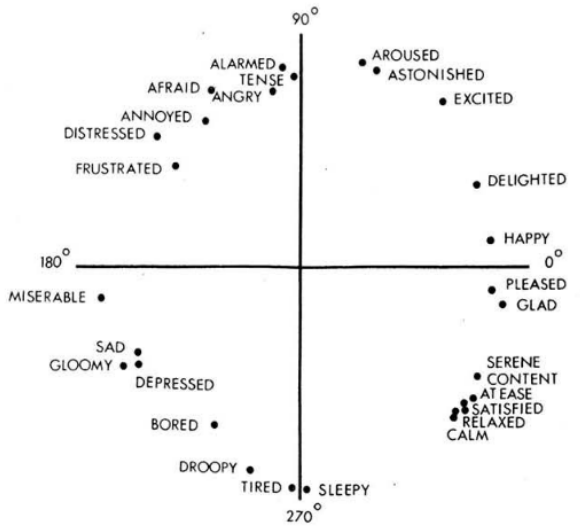

Fig. 1. Mapping of 28 emotions onto the VA plane [2]

dimensional (see Figure 1). We use the dimensional model, first suggested in [3], that represents emotions in the valence-arousal plane [2] for music mood annotation. The authors argue that a dimensional model eliminates the confusion associated with a categorical modeling using discrete labels which tend to be ambiguous and are prone to varied interpretation. On the other hand, a dimensional model represents every emotional state as a unique point in the valence-arousal plane in the case of a two dimensional model. Songs closest to a point specified by a user are returned during music retrieval. This work employs a two dimensional model as most datasets provide valence and arousal annotations for the comprising audios. Further, the most often considered third dimension of tension is highly correlated with valence [4] and thus there is little harm in ignoring it. The data prepared as a part of this work provides valence and arousal ratings for every song listed with all the values normalized to belong to $[-1, 1]$. The process of preparation of this final dataset is

[1]Milind Kumar V is an undergraduate student at IIT Madras, Chennai, Tamil Nadu, India. milind.blaze9@gmail.com

described in detail in section II. While this work prepares the data with valence and arousal ratings, the Geneva Emotional Musical Scales (GEMS hereafter) prepared by Zentner et al. is a noteworthy mention. These are musically relevant emotional categories created specifically for the purpose of music emotion recognition (MER henceforth). They consist of 45 labels that can be grouped into 9 categories. This system of categorical representation has been used in the Emotify dataset.

Data prepared for this work can be used for future experiments that predict the valence and arousal (VA henceforth) values for supplied audios. Modifications to the VA mapping of categorical labels can be made by altering the source code which is made available here[1]. This work focuses on the simpler task of classifying music into the four quadrants of the VA plane. This is founded on the premise that agreement among annotators regarding the quadrant to which a given piece of music belongs is higher than the agreement regarding a particular emotional label or VA values. While less complex, it is a non- trivial task and is very similar to the Mirex 2016[2] and 2017 mood classification tasks in which entries must classify audio into five clusters. Musically relevant features such as spectrograms spanning octaves generated using triangular filterbanks and mel spectrograms are extracted as inputs to the convolutional networks. Processing audios offers greater flexibility and increases the scope of the experiments that can be conducted as different datasets offer different musically relevant hand crafted features for MER. Following this, experiments are conducted on multiple architectures with different settings of hyperparameters and the results are presented in VI.

II. Data collection

There are multiple publicly available datasets created for the purpose of music mood annotation and others that address the same task whilst also offering labels for genre, instrument and so on. Table I presents a list of such datasets of which the ones without audios have been ignored for this work. Some datasets however offer a list of YouTube links to the songs annotated and a possible direction of work would be to crawl the internet and obtain the audios for these aforementioned songs. Some such datasets are AMG1608[3] [5], Google Audioset[4] [6], DEAP [7], NJU-MusicMood-v1.0[5], Greek Music Dataset[6] [8] (GMD), Greek Audio Dataset[5] [9] (GAD). While AMG1608 and DEAP are ignored solely as they do not provide audios, the Google Audioset dataset is not generated as it provides weak labels with only 3 annotators labeling each clip and a significant portion of this data must be used to obtain a network that produces meaningful results. This requires incredible computing power and training time and can not be done on a budget. While the NJU-MusicMood-v1.0 dataset does not provide audios, it supplies lyrics which could be used for the purpose of annotation. The GAD and GMD datasets are not a part of this work as they do not provide audios and also because the former is prepared by 5 annotators and the number of annotators for the latter is unspecified as is the annotation granularity. In this section, a brief description of the datasets used is provided with details of how the VA ratings were altered for this work.

A. DEAM

This database[7] [10] was compiled for the 2015 Emotion in Music task as part of the MediaEval Benchmarking Initiative for Multimedia Evaluation. It contains 58 full songs (duration 234 ± 107s) and 1744 clips 45s in length. Developed for the purpose of dynamic MER, this dataset offers time varying labels which aren't used. Instead, song level annotations (static annotations) are used. These belong to the range $[1, 9]$ and are normalized to $[-1, 1]$.

B. CAL500

CAL500 [11] is a very popular dataset and has often been used in MIR. Songs are annotated with 174 labels of which 36 are mood related tags. These tags are positive-negative pairs with there being both Emotion and Not-Emotion tags. Further, several issues with the original dataset are addressed by following the steps suggested by Bob Sturm[8] which have also been adopted in the creation of the CAL500 expansion dataset [12]. The Not-Emotion labels are of very little use for the intended classification of music and are thus discarded. Further, all songs with positive hard annotations (values of 1) for the emotion tags of Emotion-Bizarre_/_Weird, Emotion-Loving_/_Romantic, Emotion-Positive_/_Optimistic, Emotion-Tender_/_Soft, Emotion-Touching_/_Loving were discarded owing to the difficulty of placing them on the VA plane. Upon further discarding songs that have no positive emotion associated with them, only 275 songs remain and are mapped to the VA plane.

Every emotion tag is associated with a pair of VA values as follows- Angry_/_Agressive ([-0.6, 0.6]), Arousing_/_Awakening ([0.2, 1]), Calming_/_Soothing ([0.4, -0.7]), Carefree_/_Lighthearted ([1, 0]), Cheerful_/_Festive ([0.6, 0.6]), Emotional_/_Passionate ([0.2, 1]), Exciting_/_Thrilling ([0.707, 0.707]) [2], Happy ([1, 0.4]), Laid-back_/_Mellow ([0.8, -0.6]), Light_/_Playful ([1, 0]), Pleasant_/_Comfortable ([0.8, -0.6]), Powerful_/_Strong ([0.2, 1]), Sad ([-0.8, -0.4]). These are obtained from two sources- an internal mapping developed by Fraunhofer IDMT over the course of a project (see Table II) and from [2]. The CAL500 dataset also provides soft annotations that describe the fraction of annotators who deemed it appropriate to associate a particular tag with a song. The final VA values for a song are the weighted average of the VA values equivalent to the labels associated

[1] /home/vaddmr/repo/trunk/idmt/projects/MusicMoodAnnotation/data_manipulation.ipynb
[2] http://www.music-ir.org/mirex/wiki/2016:Audio_K-POP_Mood_Classification
[3] http://mpac.ee.ntu.edu.tw/dataset/AMG1608/
[4] https://research.google.com/audioset//ontology/music_mood_1.html
[5] https://cs.nju.edu.cn/sufeng/data/musicmood.htm
[6] https://hilab.di.ionio.gr/old/en/music-information-research/
[7] http://cvml.unige.ch/databases/DEAM/
[8] http://media.aau.dk/null_space_pursuits/2013/03/using-the-cal500-dataset.html

Dataset	Annotation method	Number of songs/ excerpts	Total duration of audio (in hours)	Annotation granularity (in s)
DEAM (MediaEval)	MTurk	1802	25.58	45
AMG1608	MTurk 665 annotators	1608	13.4	30
MoodSwings	MTurk 546 annotators	240	1	15
Jyvaskyla Soundtrack dataset	Manual annotation 116 annotators	110	0.45	10 - 30
Lastfm100 corpus	-	100	0.83	30
Moodo	952 annotators	200	0.83	15
Emotify	GWAP	400	6.66	60
Yang60	40 Annotators	60	0.5	30
Emotion_VW	-	32	-	Song level
Mood_circumplex_training	-	418	-	Song level
CAL500	66 annotators	500	26.6	Song level
CAL500 expansion dataset	11 annotators	500	Not given.	Can't say.
Magnatagatune	GWAP	22863	132	29
Google Audioset	Manual annotation	16955	47	10
DEAP dataset	32 participants	120 (online assessment) 40 (self rating)	2 (online assessment) 0.666 (self rating)	60
NJU-MusicMood-v1.0	-	777	Not given.	Song level
Greek Audio dataset	Manual annotation	1000	Not given.	Can't say.
Greek Music dataset	Manual annotation	1400	Not given.	Can't say.

TABLE I
DATASETS USED IN THIS WORK

Emotion	Valence	Arousal
happy	1	0.4
relaxing	0.8	-0.6
calm	0.4	-0.7
danceable	0.6	0.6
fun	1	0
energetic	0.2	1
melancholic	-0.8	-0.4
aggressive	-0.6	0.6
stressful	-0.1	0.8
dramatic	-0.2	0.5

TABLE II
MAPPING OF EMOTIONS FROM A FRAUNHOFER PROJECT

with a song with the soft annotations as the weights. Perhaps a better approach is to obtain such weighted sums for all songs and normalize them using the highest set of absolute VA values assigned to any song.

The CAL500 expansion[9] dataset is an extension of the CAL500. It uses the same audios and is designed for dynamic MER with very carefully chosen representative segments that are determined by using k-medoids clustering on acoustically homogeneous segments extracted from the audios. This leads to nearly 6.4 segments per song. This method appears to be much more robust than annotating clips with arbitrarily chosen starting points and lengths. However, this dataset is not relevant to this work as the main focus is on obtaining track level annotations.

C. LastFM100 corpus

The LastFm100 corpus is comprised of 100 audios classified into four classes- Anxious_Frantic, Content, Depressed and Exuberant each of which is given a corresponding VA rating (see Table III) that is consistent with the mapping for the CAL500 dataset.

[9] http://slam.iis.sinica.edu.tw/demo/CAL500exp/

Emotion	Valence	Arousal
Anxious_Frantic	-0.1	0.8
Content	0.8	-0.6
Depressed	-0.8	-0.4
Exuberant	1	0.4

TABLE III
VA MAPPING FOR LASTFM100 CORPUS

D. Jyvaskyla Soundtracks

Jyvaskyla Soundtracks dataset[10] [4] consists of two sets of annotated tracks- 360 from a pilot experiment and a subset of the same with 110 tracks. The latter set is included in this work as these songs are annotated by 116 annotators whereas the former are annotated by the 12 'expert musicologists' who selected them and consequently, the generated VA ratings are not very representative of the emotions perceived when listening to this music. The VA ratings belong to the range $[1, 9]$ and are normalized to belong to $[-1, 1]$.

Eerola et al. present a very interesting discussion on the relation between the categorical and dimensional models and the third dimension of tension.

E. Moodo dataset

The Moodo dataset[11] [13] is one of the very few datasets to differentiate between induced and perceived emotions. Annotators tag songs by dragging emotion labels onto the VA plane resulting in there being both VA values and discrete labels for the songs. Only VA values for perceived emotion are considered. However a provision is made[1] to extract induced emotion values from the data if need be. Further, as multiple emotion labels can be assigned by an annotator to a given song, each annotation is treated as being independent

[10] https://www.jyu.fi/hytk/fi/laitokset/mutku/en/research/projects2/past-projects/coe/materials/emotion/soundtracks/Index

[11] http://mood.musiclab.si/index.php/en/dataset

Emotion	Valence	Arousal
Aggressive	-0.6	0.6
Relaxed	0.8	-0.6
Melancholic	-0.8	-0.4
Euphoric	1	0.4

TABLE IV
VA MAPPING FOR THE MOOD_CIRCUMPLEX_TRAINING DATASET

and the final VA values of a song are the average of all such annotations. The comprising 200 songs are annotated by 952 annotators making this a very high quality dataset.

As described before, the VA equivalents for discrete emotion labels are determined from [2] which employed only 36 students to arrive at the positions of the labels on the VA plane. However, the Moodo dataset presents an interesting opportunity to obtain VA values for some discrete emotional tags. By extracting the VA values associated by multiple annotators with emotional labels and averaging them out, it is possible to obtain a more reliable VA mapping of emotional labels.

F. Emotify dataset

The Emotify dataset[12] is the only dataset in this work that uses GEMS and is constructed from a very thoughtfully devised game with a purpose (GWAP). 9 very musically relevant labels are used by GEMS as they capture the general distribution of emotions in music. However, the label 'nostalgia' is difficult to map onto the VA plane and hence is ignored.

G. Yang60 corpus

The Yang60 corpus is annotated by 40 annotators and provides annotations between $[-1, 1]$. While a majority of the aforementioned datasets make a concerted effort to use music that is not very well known in order to avoid any bias introduced by familiarity and episodic memories, the Yang60 provides annotations for popular Western music which is likely to be encountered by the prepared system and thus is a valuable addition to this work.

H. Emotion_VW_OvGU and Mood_Circumplex_training

These are datasets internal to Fraunhofer IDMT and follow a classification system similar to that of LastFM100 corpus with the emotion classes being Anxious, Content, Depressed, Exuberant (see Table III) and Aggressive, Relaxed, Melancholic, Euphoric respectively. Together, they account for 450 tracks. They are consistently mapped to the VA plane for future use.

III. DATA PREPARATION

This work focuses on assigning given audios to one of the four quadrants in the VA plane. While relatively simple, this task is equivalent to assigning songs to emotion clusters each of which lies in a particular quadrant in the VA plane and is hence, non-trivial. All songs with positive VA values are assigned to the first quadrant and those with negative valence and positive arousal to the second quadrant and so forth. While this discards much of the information collected so far, it also removes ambiguities arising from the VA equivalents assigned to discrete labels. This is because annotators differ on the exact VA values for a song, but generally agree on the emotion cluster it must be assigned to. All audios are converted to the .wav format before further processing.

[12] http://www.projects.science.uu.nl/memotion/emotifydata/

IV. DATA PRE-PROCESSING

Convolutional neural networks lead to the creation of powerful models with incredible capacity. Their usage requires that audios be converted into suitable spectrograms that can be used as inputs to the CNN models. This work considers two types of spectrograms generated from the STFTs of the audios which are resampled to 44.1kHz- one in which the frequency axis is converted to a logarithmic scale founded on musical intuition and the mel scale which captures the energy or power present in regions of frequencies which the human ear is particularly sensitive to. The issue of varying durations of the audios in this eclectic dataset is addressed by extracting patches from the obtained spectrograms. A number of time frames (n) each of duration ls make up a patch. Every patch inherits the same labels as the whole song. If the time equivalent of the hopsize used for the STFT at the specified sampling rate is ms, then the duration of the song t captured per song is

$$t = n \times (l - m) + m \tag{1}$$

where $m = STFT\ hopsize/sampling\ rate$ and $l = STFT\ windowsize/sampling\ rate$. The time duration per patch t is experimented with extensively. The scientific libraries numpy, scipy and librosa are used for the extraction of the relevant features.

A. The Musical scale

The frequency range which is selected to be from 50Hz to 15000Hz is divided into octaves resulting in the axis spanning 8.228 octaves with each 12 semitones (logarithmic with base 2) between every octave resulting in 99 semitones which serve as bins. This is motivated by the musical scale and its aptness is evaluated through experimentation. The logarithm of the spectrogram is multiplied with a triangular filterbank which is used to better capture the information at frequencies that are muscially important. Patches are extracted from the obtained spectrogram. The number of frames used to make a patch is varied to determine the ideal setting of this hyperparameter.

B. Mel scale

Most CNN approaches to audio classification tasks use mel spectrograms as they capture information adequately at lower frequencies and grow less discriminatory at higher frequencies emulating the behavior of the human ear which perceives equal changes in pitch with larger changes in frequency at higher frequencies. 40 mel filters are used, leading to 40 bins in every spectrogram. The use of Slaney's implementation [14] for the construction of the mel filterbanks leads to the

the first 9 frequencies being linearly spaced (till 958 Hz) and the remaining being logarithmically spaced. The mel spectrogram is normalized by the highest value. A small value $\epsilon = 10^{-8}$ is added to the spectrogram and the logarithm is taken. Patches are extracted from the spectrograms to be fed as the input to our network. Multiple choices for t are tested by varying the parameters n, m, l and using equation 1. The duration represented by every patch is given four values- 1.17s, 2.003s, 2.82s and 6.923s.

V. Experiments

Three CNN architectures are experimented with in this work. The first is a model that has been found to produce excellent results in music speech discrimination at Fraunhofer IDMT. Kernel sizes, dropout ratios, l2 regularization are all experimented with. The second is a model submitted by Thomas Lidy and Alexander Schindler [15] to the MIREX 2016 mood classification task. The third model has four hidden layers with rectangular filters aimed at capturing more temporal and frequency information than before. Further, networks are trained both to classify music into four quadrants and to classify them separately as high and low valence or arousal. A categorical crossentropy loss function is used for all four-class classification experiments and binary crossentropy for valence and arousal classificaton. All networks are implemented in Keras and the results are recorded.

A. Four-class classification

1) Model1: Model1 (see Fig. 2) has two units of two convolutional layers followed by a layer of Max pooling. Every convolutional layer applies 32 filters. This is followed by a dense layer with 256 units followed by the output layer. Dropout with a probability of 0.5 is used before every dense layer. All kernels are 5×5. This model is a test of how well conventional ideas in vision carry over to analysis of audio. All layers are initialized with the He initializer. Relu activations are used for all the layers.

2) Model2: This model contains a single convolutional layer with 30 10×12 filters. This is followed by a max pooling layer with a kernel of 1×20. Finally a 200 unit densely connected layer is connected to the output layer. All layers are initialized with the Glorot uniform initialization. Leaky Relu activations are used for all the layers with an $\alpha = 0.3$. A dropout value of 0.5 is used for all the densely connected layers.

B. Two-class classification

1) Model3: A model very similar to that of Model1 is used for the task of valence and arousal classification with the main difference being that the 256 unit densely connected layer is replaced by two 512 unit fully connected layers. The output layer's softmax activation function is replaced by a sigmoid activation.

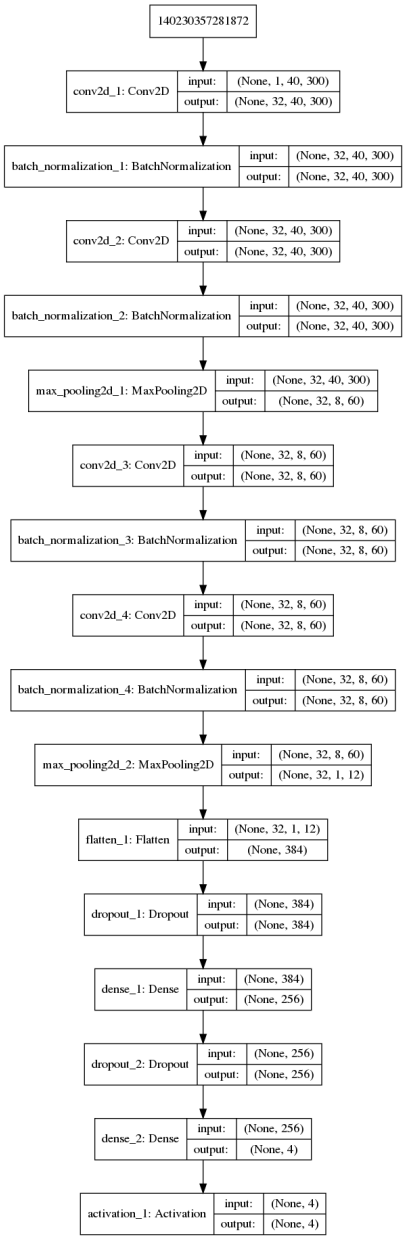

Fig. 2. Model1

2) Model4: This is a two layer convolutional network employing filters of dimensions 20×5 and 5×20 followed by a max pooling layer with a 3×3 kernel and convolutional layer of 32 filters with kernel size 3×3. This is directly connected to the output layer after flattening. Dropout with probability 0.5 is used. The output is a sigmoid layer. All weights are initialized with the He initializer.

VI. Results

Results of the experiments that produced any meaningful outcomes are reported in Table V. Most models are highly prone to overfitting which simple l2 regularization (for the convolutional layers) and dropout do not fix. Multiple representations of the data are attempted. Hyperparameters are tweaked as necessary to improve performance and find

Task (classificaton)	Model	Percentage of data used	Train, validation, test split	Learning rate	Scale used	STFT windowsize	STFT hopsize	Frames per patch	Frames hopsize	Minibatch size	Epoch stopped	L2 regularizaion parameter/ Dropout ratio	Train accuracy	Validation accuracy	Test accuracy (filewise)
Four quadrants	Model1	0.7	0.9	1.00E-03	Musical	3072	1536	80	40	128	193	None/ 0.5	0.7624	0.415	-
Four quadrants	Model1	0.7	0.9	1.00E-03	Musical	3072	1536	200	100	128	100	None/ 0.5	0.9376	0.4717	0.4120
Four quadrants	Model1	0.7	0.9	1.00E-03	Musical	3072	1536	300	150	128	100	None/ 0.5	0.9386	0.4685	0.4377
Four quadrants	Model1	1	0.9	1.00E-03	Mel	1024	512	100	50	128	100	0.01/ 0.5	0.9386	0.4685	0.4377
Four quadrants	Model1	1	0.9	1.00E-03	Mel	1024	512	100	50	128	100	0.01/ 0.5	-	-	0.5089
Four quadrants	Model2	0.6	0.8	1.00E-03	Musical	3072	1536	80	40	128	101	None/ 0.2	0.8241	0.3886	0.3734
Four quadrants	Model2	0.6	0.8	1.00E-03	Musical	3072	1536	80	40	128	101	None/ (0.5, output- 0.2)	0.7468	0.4028	0.3934
Four quadrants	Model2	0.6	0.8	1.00E-03	Musical	3072	1536	80	40	128	101	None/ 0.5	0.6697	0.408	0.3959
Four quadrants	Model2	0.6	0.8	1.00E-03	Musical	3072	1536	80	40	128	106	0.01/ 0.5	0.5015	0.3828	0.2581
Four quadrants	Model2	1	0.9	1.00E-03	Musical	3072	1536	300	150	128	100	None/ 0.5	0.8582	0.4411	0.4506
Valence	Model3	0.7	0.9	1.00E-05	Mel	1024	512	100	50	512	100	0.01/ 0.5	0.7458	0.6373	0.6094
Valence	Model3	0.7	0.9	1.00E-04	Mel	1024	512	100	50	512	100	0.01/ 0.5	0.9095	0.6518	0.6480
Valence	Model3	0.7	0.9	1.00E-04	Mel	1024	512	100	50	256	100	0.01/ 0.5	0.9163	0.6763	0.7081
Valence	Model3	1	0.9	1.00E-04	Mel	1024	512	100	50	256	100	0.01/ 0.5	0.8874	0.6422	0.6636
Valence	Model4	1	0.9	1.00E-04	Mel	1024	512	100	50	256	29	0.01/ 0.5	0.8128	0.642	0.7057
Arousal	Model3	1	0.9	1.00E-04	Mel	1024	512	100	50	256	20	0.1/ 0.5	0.7991	0.7478	0.7027
Arousal	Model3	1	0.9	1.00E-04	Mel	1024	512	100	50	256	20	1/ 0.5	0.7592	0.7611	0.7177

TABLE V

RESULTS OBTAINED FOR A FEW CLASSIFICATION TASKS

the best values. A better approach would definitely be using a grid search to find the best set of values.

Model1 is the best performing model for the four-class classification task with a filewise train accuracy of 0.51 when trained with the whole dataset which leads to 415 files per quadrant for training when balanced, 47 per class for cross validation and is evaluated on 334 files. However, this is not a very reliable figure as all the models suffer from severe overfitting implying that very little learning has happened.

Performance is much higher on valence and arousal classification, which is to be expected as these are simpler tasks with only two classes. Even a test accuracy of $\approx$0.7 implies that the quadrant assigned to a given file is correct with the probability $0.7 \times 0.7 = 0.49$. The final system built from this work is a python script that takes as a command line argument the path to the folder containing all the files that need to be annotated and outputs the valence and arousal of the songs. The models used for the prediction are the ones with the best test set accuracy during training as shown in Table V.

VII. Future work

This section highlights the possible directions of future work that build on what has been presented in this work. All the networks trained so far have shown massive overfitting, including the relatively simple architectures such as Model3. This leads one to suspect that the very nature of data being presented is flawed. While normalizing every mel spectrogram by the highest value does improve performance, another approach would be to normalize across patches ensuring all patches that inherit a label from the given audio possess similar amplitude level. Greater experimentation with the parameters associated with the creation of patches would perhaps yield better results. Further, early stopping with a patience of 10 has been used to avoid overfitting. This consequently reduces performance on the training set as well. Implementing more complex early stopping methods [16] can increase train accuracy while not compromising generalization. Transfer learning is an approach that has not been used in this work and fine-tuning pre-trained networks which have been used for audio applications such as music speech discrimination is one possible avenue of producing better results. Using pre-trained networks such as VGGish[13] which have been trained on the Audioset dataset as feature extractors could also lead to better results. The use of large datasets such as the Million Song Dataset [14] and Audioset which have relatively weak labels for pre-training is an idea to be tested. Another major area of focus is the task definition. Prediction of VA values for a song would greatly increase the flexibility of the mood annotation system and afford much finer placement of songs for retrieval. While the major thrust of this work has been to use CNNs, LSTM-RNNs have shown remarkable results [17] on the MediaEval Emotion in Music task which uses a dataset of comparable size.

[13] https://github.com/tensorflow/models/tree/master/research/audioset

[14] https://labrosa.ee.columbia.edu/millionsong/

Another area this work has not covered is the usage of lyrics. This approach is generally not favored owing to the unavailability of lyrics or the difficulty of obtaining them for the songs one wishes to annotate. Datasets such as NJU-MusicMood-v1.0[4] and LAMP [18] do provide lyrics and can be used in future work in music mood annotation.

For work focusing on better prediction of VA values, a more comprehensive mapping of emotional labels to the VA plane is necessary. As mentioned in section II, one possible method of achieving this is by extracting the equivalent VA values from the Moodo dataset. Some interesting questions to consider are if and how much noise is introduced into the data due to the varying understanding of valence and arousal amongst annotators from dataset to dataset and among annotators labeling songs in a given dataset.

VIII. Conclusion

This work has described the development of a music emotion recognition system that labels a given audio of a song with positive or negative valence or arousal tags. A database of nearly 79 hours of audio is created with only VA ratings. The use of the musical scale in spectrograms does not produce any results significantly better than those obtained by using the mel spectrogram. The performance of the selected models does vary with the duration each patch represents. All the models used in this work show significant overfitting and only reach an accuracy of 50.89% leading one to conclude that better pre-processing of data and perhaps the use of models such as LSTM-RNNs is necessary to improve accuracy.

ACKNOWLEDGMENTS

I thank Prof. Karlheinz Brandenburg, Hanna Lukashevich and Jakob Abesser for their immense support and guidance. I would also like to thank Sascha Grollmisch, Dominik Zapf and Hany Tawfik Fayek for their insights and very illuminating discussions.

References

[1] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. W. Wilson, "CNN architectures for large-scale audio classification," *CoRR*, vol. abs/1609.09430, 2016.

[2] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[3] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 2, pp. 448–457, 2008.

[4] T. Eerola and J. K. Vuoskoski, "A comparison of the discrete and dimensional models of emotion in music," *Psychology of Music*, vol. 39, no. 1, pp. 18–49, 2011.

[5] Y.-A. Chen, Y.-H. Yang, J.-C. Wang, and H. Chen, "The amg1608 dataset for music emotion recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 693–697, Citeseer, 2015.

[6] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pp. 776–780, IEEE, 2017.

[7] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 18–31, 2012.

[8] D. Makris, I. Karydis, and S. Sioutas, "The greek music dataset," in *Proceedings of the 16th International Conference on Engineering Applications of Neural Networks (INNS)*, EANN '15, (New York, NY, USA), pp. 22:1–22:7, ACM, 2015.

[9] D. Makris, K. L. Kermanidis, and I. Karydis, "The greek audio dataset," in *AIAI (2)*, vol. 437, pp. 165–173, 2014.

[10] A. Alajanki, Y.-H. Yang, and M. Soleymani, "Benchmarking music emotion recognition systems," *PLOS ONE*, 2016. under review.

[11] D. Turnbull, L. Barrington, D. Torres, and G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 2, pp. 467–476, 2008.

[12] S.-Y. Wang, J.-C. Wang, Y.-H. Yang, and H.-M. Wang, "Towards time-varying music auto-tagging based on cal500 expansion," in *Multimedia and Expo (ICME), 2014 IEEE International Conference on*, pp. 1–6, IEEE, 2014.

[13] M. Pesek, P. Godec, M. Poredos, G. Strle, J. Guna, E. Stojmenova, M. Pogacnik, and M. Marolt, "Introducing a dataset of emotional and color responses to music.," in *ISMIR*, pp. 355–360, 2014.

[14] T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various mfcc implementations on the speaker verification task," in *Proceedings of the SPECOM*, vol. 1, pp. 191–194, 2005.

[15] T. Lidy and A. Schindler, "Parallel convolutional neural networks for music genre and mood classification," *MIREX2016*, 2016.

[16] L. Prechelt, "Early stopping-but when?," in *Neural Networks: Tricks of the trade*, pp. 55–69, Springer, 1998.

[17] A. Aljanaki, Y.-H. Yang, and M. Soleymani, "Developing a benchmark for emotional analysis of music," *PloS one*, vol. 12, no. 3, p. e0173392, 2017.

[18] R.-c. Wei, R. Tsai, Y.-s. Wu, *et al.*, "Lamp, a lyrics and audio mandopop dataset for music mood estimation," in *Dataset Compilation, System Construction, and Testing: 2010 International Conference on Technologies and Applications of Artificial Intelligence*, pp. 53–59, 2010.