Analysis of the Effect of Dataset Construction Methodology on Transferability of Music Emotion Recognition Models

Sabina Hult, Line Bay Kreiberg, Sami Sebastian Brandt, Björn Þór Jónsson
s@bina.dk,lineokr@gmail.com,sambr@itu.dk,bjth@itu.dk
IT-University of Copenhagen
Copenhagen, Denmark

ABSTRACT

Indexing and retrieving music based on emotion is a powerful retrieval paradigm with many applications. Traditionally, studies in the field of music emotion recognition have focused on training and testing supervised machine learning models using a single music dataset. To be useful for today's vast music libraries, however, such machine learning models must be widely applicable beyond the dataset for which they were created. In this work, we analyze to what extent models trained on one music dataset can predict emotion in another dataset constructed using a different methodology, by conducting cross-dataset experiments with three publicly available datasets. Our results suggest that training a prediction model on a homogeneous dataset with carefully collected emotion annotations yields a better foundation than prediction models learned on a larger, more varied dataset, with less reliable annotations.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by regression**; **Supervised learning**; • **Applied computing** → **Sound and music computing**.

KEYWORDS

music emotion recognition, cross-dataset, model transferability

ACM Reference Format:
Sabina Hult, Line Bay Kreiberg, Sami Sebastian Brandt, Björn Þór Jónsson. 2020. Analysis of the Effect of Dataset Construction Methodology on Transferability of Music Emotion Recognition Models. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20), October 26–29, 2020, Dublin, Ireland*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3372278.3390733

1 INTRODUCTION

Emotion in music has great potential in a number of music retrieval applications, e.g., processing queries with emotion terms or automated construction of playlists to alter or support mood and emotion. Having reliable meta-data about emotional content is therefore important for such applications. Although some of the large music streaming services already provide emotion-based playlists, the playlists provided by Spotify are manually curated by music experts.[1] As music archives are very large and ever-expanding, manually assessing the emotions present in music tracks becomes a task that could benefit from automated assistance. Therefore, a reliable automated method to map music content to emotions, that can be applied to any music to obtain such meta-data, is desired.

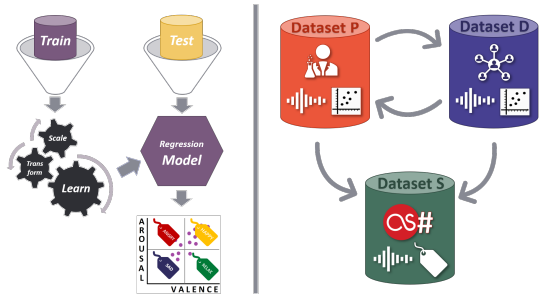

Figure 1: Cross-dataset experiment pipeline and prediction directions

A typical approach to obtaining a mapping from music content to emotion meta-data is to create a benchmark dataset of music, collect ground truth data for music emotion in the dataset, and then apply some supervised machine learning algorithm to generate a model [16]. The ground truth typically takes the form of manual assessments, which is a resource-demanding task resulting in traditionally small datasets. Annotations can take the form of affect dimensions, which is preferred for fine-grained retrieval operations, or as distinct emotion terms, where in some cases social media tags are used to yield rough emotion annotations for larger music datasets. Most of the research in music emotion recognition has then focused on the machine learning phase: given a dataset with annotated ground truth, what is the best way to create the model. We are, however, interested in another question: to what extent do the ground truth annotation method and music selection impact the ability to generate a good model of content emotion?

The general cross-dataset prediction pipeline, and the prediction directions between datasets are shown in Figure 1. We first consider learning across two dimensional datasets ($\mathbb{P}$ and $\mathbb{D}$) from the literature with ground truth annotations in the form of valence and arousal (VA). The sets differ in type of music, audio excerpt selection, and ground truth collection method. One has very homogeneous music data and a manually selected excerpt, namely the chorus, as well as a very carefully controlled ground truth construction. In the second dataset, which is more than twice the size of the first,

[1] https://support.spotify.com/us/using_spotify/playlists/create-a-playlist/

crowd-sourcing was utilized for ground truth annotation, music data spans several genres of music, and the selected excerpt is a fixed length segment from a uniformly distributed starting point in the music pieces. A third dataset, $\mathbb{S}$, is then included with nearly 3000 entries applying a categorical emotion model for ground truth annotations, harvested via social tagging. This particular dataset is used as an additional evaluation set for the VA-predicting models trained on each of the dimensional datasets.

The remainder of the paper is organized as follows: In Section 2 we briefly introduce the field of music emotion recognition (MER) and recount some contributions to cross-dataset work in the field, which have served as inspiration. Section 3 describes the methods applied for dataset inclusion, audio preprocessing, and the experimental setup. We then present and discuss our results in Section 4, before concluding in Section 5.

2 BACKGROUND & RELATED WORK

Emotions are traditionally conceptualized either categorically or dimensionally. Categorical models of emotions regard emotions as fundamentally different constructs that are distinct from one another, i.e., emotion classes such as happy, sad and angry. The theory of basic emotions by Ekman [4] is one such categorical emotion model. The dimensional approach defines emotions as points in an affect plane; a coordinate space, where a certain point represents an emotion state. One of the most famous and frequently used dimensional models is Russell's two-dimensional circumplex model of affect [2, 7, 10, 11, 15]. It proposes that every emotion state arises from two neurophysiological systems; one related to displeasure/pleasure, in literature usually referred to as valence, and one related to degree of arousal. The different approaches to emotion modelling can be merged; Song et al. [13] divided Russell's VA-space into four quadrants and labelled them happy, angry, sad and relax (short for relaxed). The two lines dividing the four quadrants were placed in the exact middle of the VA-space [13]. This technique roughly maps the four-label categorical emotion model onto the dimensional VA-space and so allows to transfer machine learning models across different emotion models.

Few researchers have explored cross-dataset music emotion recognition. One of the studies that inspired us is Eerola's study of whether emotions expressed in music are genre-specific [3]. Nine different datasets of varying size and construction method were used, all applying the VA-model of emotions. Three of the datasets were comprised of classical music, two of soundtracks, two with mainly popular music and two datasets with a mixture of genres. Regression models were constructed with step-wise multiple regression; a model for each of the affect dimensions, valence and arousal. An exhaustive cross-validation scheme was then performed where, for each dataset, one was used as the source, and the models were tested on all remaining datasets. Results indicate that arousal can be predicted across genres, while valence proves to be more difficult, which is a general observation in the field of MER.

Another important and more recent contribution to cross-dataset MER is the study by Hu and Yang [5] where the emphasis is on the cultural differences between annotating subjects and music selection, and how this might affect emotion prediction in terms of the VA-model. Three datasets where involved in the study; one with Chinese music annotated by Chinese subjects, one with Western music annotated by Chinese subjects and one with Western music annotated by Western subjects. Models where built with support vector regression, and results indicate that models are generalizable across datasets that share a common cultural background between either annotators or music selection.

Where Eerola's study [3] is focused on genre and Hu and Yang's study [5] is focused on cultural aspects, we were interested in investigating the effect of dataset construction methods in terms of music segment selection and ground truth collection methods, thus providing insights regarding dataset construction for reliable and transferable music emotion recognition models.

3 METHOD

In this section we briefly list our criteria for dataset inclusion and summarize properties of each included dataset. Then we describe the methods applied for audio preprocessing, such as data cleaning, feature extraction, value scaling and data reduction, followed by an explanation of the experimental setup.

3.1 Datasets

In order to perform cross-dataset prediction, preferably the datasets should be very similar apart from the properties we wish to investigate. Since there are many factors involved in the creation of MER-datasets, we established a list of criteria to locate the most suitable data for our investigation: 1) datasets must encompass available music audio files such that feature extraction can be done to ensure the application of the same regression models to each dataset; 2) the measured emotion should be perceived emotion; 3) datasets must include ground truth in the form of static annotations, i.e., one set of VA-annotations for the entire excerpt, since this is less prone to contextual factors; 4) the emotion model should be dimensional or intuitively transferable from the dimensional model, specifically Russell's VA-model; and 5) the construction of the datasets should be different in terms of annotation collection method and music stimuli. Out of the publicly available datasets we explored, PMEmo, DEAM and the Song, Dixon and Pearce datasets were chosen for inclusion. After the description of each dataset, they will be referred to as $\mathbb{P}$, $\mathbb{D}$ and $\mathbb{S}$, respectively.

PMEmo [17] has close to 800 music excerpts consisting of well-known popular music. The music segment is specifically the chorus part of each song, which has been manually selected by music students. Each excerpt are annotated with valence and arousal values by at least 10 subjects. The subjects are all students and annotations were collected in a carefully setup lab-environment with a high degree of control over subjects and process.

DEAM [1] contains close to 2,000 music excerpts of primarily unknown music published under the Creative Commons License. The music data covers a wide range of genres and the excerpts are a fixed-length segment from a uniformly distributed starting point in the track. Annotations were collected by utilizing a crowd-sourcing approach where measures were taken to ensure the quality of annotations, and each excerpt is annotated by at least 10 subjects.

Table 1: Supervised learning models and algorithmic parameters

Regression Model	Parameters
Ridge	`alpha=1.0, fit_intercept=True, normalize=False, max_iter=None, tol=0.001, solver='auto', random_state=None`
SVM, linear kernel	`kernel='linear', gamma='scale', coef0=0, tol=0.001, C=1, epsilon=0.1, shrinking=True, max_iter=-1`
SVM, rbf-kernel	`kernel='rbf', gamma='scale', coef0=0, tol=0.001, C=1, epsilon=0.1, shrinking=True, max_iter=-1`

The Song, Dixon and Pearce dataset [12] is the largest of the three datasets, with close to 3,000 music excerpts. The dataset contains no meta-data about the music selection apart from the title of each track. The tracks are assumed to be mainly known music. The emotion annotations are collected from social tags on Last.FM using the keywords angry, happy, sad and relax as seeds.

3.2 Audio Preprocessing

Before prediction models can be learned, several preprocessing actions are needed, which are detailed in the following paragraphs.

Data Cleaning. Although data cleaning activities have already been performed by the creators of the individual datasets, we encountered a few duplicates, which were removed. Additionally we decided to only extract audio features from audio files that were of length less than 90 seconds to ensure some consistency across the datasets. After this rough cleaning the size of $\mathbb{P}$, $\mathbb{D}$ and $\mathbb{S}$ was 760, 1737 and 2880 entries respectively.

Feature Extraction. We utilized the LibROSA [8] Python package for feature extraction and, for the most part, adopted default values of the parameters used in the feature extraction library were. Each audio file was loaded in, converted to mono and resampled to 22,050 Hz, which is the default value for several feature extraction frameworks [16]. The default window size and hop length in LibROSA are 2,048 and 512 respectively. Tempo- and beat-related features were extracted such as tempo, total beats, mean beats, median beats, and the standard deviation. All available spectral features from the LibROSA library were extracted: short-time FFT chromagram, constant-q chromagram, 'chroma energy normalized', also called CENS-features, melspectrogram, MFCC's, root mean square energy, spectral centroid, spectral bandwidth, spectral contrast, spectral flatness, spectral roll-off, polynomial features, tonnetz features and zero crossing rate. All features were summarized to mean and standard deviation, in total resulting in 458 audio features.

Normalization. In the dimensional datasets, both the annotation data and feature data are represented as continuous numerical values, and are therefore normalized. In $\mathbb{P}$, the available annotation data are already rescaled to the range $[0, 1]$, so the minmax-normalization was applied to $\mathbb{D}$ annotation data. We conducted preliminary experiments to determine which scaling method to apply for normalization of feature values, which showed that z-score normalization was a good approach across the three datasets [6].

Data Reduction. Data reduction was applied with Principal Component Analysis. Extensive preliminary experiments were conducted to select appropriate combinations of regression model and number of principal components to include in the cross-dataset experiments. The three highest performing models were Support Vector Machine (SVM) with linear kernel (SVMlin), SVM with RBF-kernel (SVMrbf) and Ridge regression (Ridge) in combination with 25 principal component [6].

3.3 Experimental Setup

Experiments were implemented in Python using the SciKit-learn library [9]. All algorithms was implemented with the default parameters, and no hyper-parameter tuning was conducted (see Table 1). We implemented regression and classification versions of Ridge, SVMlin and SVMrbf [6].

Before beginning cross-dataset experiments, baseline regression with the selected learning models was conducted on $\mathbb{P}$ and $\mathbb{D}$, while baseline classification was implemented for $\mathbb{S}$. Baseline experiments were performed with a 10-fold cross validation scheme.

The first cross-dataset experiments involve $\mathbb{P}$ and $\mathbb{D}$. Regression models for predicting each affect dimension are trained on $\mathbb{P}$, then tested on $\mathbb{D}$ and vice versa. In the second part of cross-dataset experiments, regression models for predicting VA-values are trained on $\mathbb{P}$ and $\mathbb{D}$ respectively, then each learning model is tested on $\mathbb{S}$. We compare the cross-dataset performance as follows: the results produced by a learning model trained on source A and tested on target B ($A \rightarrow B$) are compared to the results from a learning model that is both trained and tested on target B ($B \rightarrow B$).

Experiments on dimensional datasets were evaluated by the R^2 metric which is standard in the field [7, 16]. To evaluate the performance of VA-prediction on the categorical dataset, the VA-space was split into quadrants, and the labels 'angry', 'happy', 'sad', and 'relax' assigned to each excerpt depending on its predicted VA-quadrant [13].

4 RESULTS AND DISCUSSION

In this section, results from cross-dataset experiments involving $\mathbb{P}$ and $\mathbb{D}$ are presented first, including baseline results. These are then followed by the results from baseline classification on $\mathbb{S}$, as well as results obtained when using prediction models learned on $\mathbb{P}$ and $\mathbb{D}$ respectively, to predict the VA-quadrant of music excerpts in $\mathbb{S}$.

4.1 Prediction Across Dimensional Datasets

Table 2 shows results from baseline and cross-dataset experiments. The baseline results obtained with $\mathbb{P}$ (third column) are reasonable given the setup with no hyper-parameter tuning, especially for valence prediction with a score of $R^2 = 0.42$. For arousal, the highest prediction score is $R^2 = 0.64$. Both were achieved with Ridge. There is currently no research done on this dataset to compare with, but in general valence prediction scores are usually lower.

When using $\mathbb{D}$ as source to the prediction model and $\mathbb{P}$ as the target (fourth column), we see a rather poor performance. In particular,

Table 2: Dimensional cross-dataset prediction scores in R^2

Dim	Model	$\mathbb{P}\rightarrow\mathbb{P}$	$\mathbb{D}\rightarrow\mathbb{P}$	Diff	$\mathbb{D}\rightarrow\mathbb{D}$	$\mathbb{P}\rightarrow\mathbb{D}$	Diff
V	Ridge	0.42	0.16	-0.26	0.34	0.18	-0.16
	SVMrbf	0.35	0.15	-0.20	0.32	0.17	-0.15
	SVMlin	0.41	0.13	-0.28	0.34	0.17	-0.17
A	Ridge	0.64	0.11	-0.53	0.17	0.15	-0.02
	SVMrbf	0.57	0.02	-0.55	0.35	0.16	-0.19
	SVMlin	0.63	0.16	-0.47	0.15	0.16	0.01

SVMrbf, a usually robust prediction algorithm that generally performs well in terms of music emotion prediction and is considered state-of-the-art in the field, shows surprisingly low performance. Overall the average performance loss is 0.24 for valence and 0.51 for arousal. Curiously this direction, from $\mathbb{D}$ to $\mathbb{P}$, shows a larger performance loss in the arousal dimension than in the valence dimension. This is not in line with the general tendency in cross-dataset research or the overall research, where valence is usually the most difficult to predict [3, 5].

Concerning $\mathbb{D}$ (sixth column of Table 2), the highest baseline arousal prediction was obtained with SVMrbf, yielding a score of $R^2 = 0.35$. With a different experimental setup and audio features, an arousal score of $R^2 = 0.64$ has been achieved on a subset of this data by [14]. Highest baseline valence prediction on $\mathbb{D}$ is achieved with Ridge and SVMlin, both with a score of $R^2 = 0.34$. In comparison, [14] achieved a valence prediction score of $R^2 = 0.42$.

In the learning direction $\mathbb{P} \rightarrow \mathbb{D}$ (seventh column), the general tendency shows that there is less performance loss in this direction with the average prediction loss being 0.16 for valence and 0.07 for arousal. SVMlin yields a minimally higher R^2-score for arousal prediction than the baseline, i.e., a performance gain is achieved. We contribute some of this to the overall poor baseline performance on $\mathbb{D}$, but emphasize that the actual prediction scores in this direction are higher for all but one of the implemented algorithms ($\mathbb{D} \rightarrow \mathbb{P}$ vs. $\mathbb{P} \rightarrow \mathbb{D}$ in Table 2).

4.2 Prediction to a Categorical Dataset

Table 3 shows accuracy results from baseline and cross-dataset experiments involving $\mathbb{S}$. Baseline classification shows very similar accuracy scores across the three algorithms with an average accuracy = 0.50. For comparison, Song et al. [12] reached an accuracy of 54% with a polynomial-kernel SVM and a different audio feature setup.

With $\mathbb{D}$ as the source for a learning model, the average accuracy is 0.34, with SVMlin performing the best. Average performance loss across the algorithms is 0.15. As in the earlier experiments, here we also see that performance is consistently better when $\mathbb{P}$ is the source rather than $\mathbb{D}$, regardless of prediction algorithm. In this direction, $\mathbb{P} \rightarrow \mathbb{S}$, the average accuracy is 0.37, and the average performance loss is 0.12.

4.3 Discussion

When utilizing a machine learning model across datasets, a performance loss is expected. The interesting observation was that the performance from learning models trained on $\mathbb{P}$ did better than models trained on the much larger $\mathbb{D}$, contrary to the expectation that more data would be better. While recognizing $\mathbb{P}$ as a dataset of high quality annotation data, we initially expected the homogeneous music selection to impair the performance of models trained on this dataset. It is unclear whether the experimental setup might have been in favor of $\mathbb{P}$, or whether we can attribute the results to the construction of $\mathbb{P}$ and its inherent properties.

A note-worthy property of $\mathbb{P}$, apart from the highly-controlled annotation methodology, is the intentional choice of choruses as music segment. For Western popular music, the chorus might be the most significant part of a song, since it is usually repeated the most, and as such might be the most salient for emotional expression. Our results in these experiments could indicate that either of those properties, or the combination thereof, might be more important than quantity in terms of dataset size. Studying this further could help the field to establish a consensus on how datasets should be constructed in order to identify the best performing computational models to recognize musical emotion. It might suggest that larger annotated datasets are not required as long as high quality datasets are obtainable.

Table 3: Cross-dataset accuracy predicting VA-values on $\mathbb{S}$

Model	$\mathbb{S}\rightarrow\mathbb{S}$	$\mathbb{D}\rightarrow\mathbb{S}$	Diff	$\mathbb{P}\rightarrow\mathbb{S}$	Diff
Ridge	0.49	0.35	-0.14	0.38	-0.11
SVMrbf	0.51	0.33	-0.18	0.37	-0.14
SVMlin	0.50	0.36	-0.14	0.38	-0.12

5 CONCLUSIONS

We set out to analyze the effect of ground truth construction and music selection on the performance of music emotion recognition models. We identify publicly available datasets that differ in those aspects, but are similar in measuring perceived emotion and all supply audio files such that identical features can be extracted. The results from cross-dataset experiments indicate that using $\mathbb{P}$ as the source dataset yields better results, even though $\mathbb{D}$ is a much larger dataset with more musical variation. This mimics the results from baseline experiments, where all the learning models also yield higher performance in both dimensions on $\mathbb{P}$ compared to $\mathbb{D}$. Our results suggest that a meticulously constructed dataset, where the music excerpt is a salient part of the track and the ground truth annotation is collected in a highly controlled environment, is a better foundation for cross-dataset prediction of static perceived musical emotion than larger sized datasets with less resource-demanding ground truth annotations, even though they may vary more in music selection. If this conclusion is supported by further research, very large annotated datasets might not be required, as long as high-quality datasets are obtainable, and it would be well worth it to create high-quality dataset as the foundation for MER-models.

ACKNOWLEDGMENTS

We would like to thank the creators of PMEmo [17], DEAM [1] and the Song, Dixon and Pearce [12] datasets for making their datasets publicly available.

REFERENCES

[1] Anna Aljanaki, Yi-Hsuan Yang, and Mohammad Soleymani. 2017. Developing a benchmark for emotional analysis of music. *PLoS ONE* 12, 3 (2017), 1–22. https://doi.org/10.1371/journal.pone.0173392

[2] Mathieu Barthet, György Fazekas, and Mark Sandler. 2013. Music Emotion Recognition: From Content- to Context-Based Models. In *From Sounds to Music and Emotions (Lecture Notes in Computer Science)*, Vol. 7900. Springer, Berlin, Heidelberg, 492–507. https://doi.org/10.1007/978-3-642-41248-6_13

[3] Tuomas Eerola. 2011. Are the Emotions Expressed in Music Genre-specific? An Audio-based Evaluation of Datasets Spanning Classical, Film, Pop and Mixed Genres. *Journal of New Music Research* 40, 4 (2011), 349–366. https://doi.org/10.1080/09298215.2011.602195

[4] Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion* 6, 3 (1992), 169–200.

[5] Xiao Hu and Yi-Hsuan Yang. 2017. Cross-Dataset and Cross-Cultural Music Mood Prediction: A Case on Western and Chinese Pop Songs. *Transactions on Affective Computing* 8, 2 (2017), 228–240. https://doi.org/10.1109/TAFFC.2016.2523503

[6] Sabina Hult, Line Bay Kreiberg, Sami Sebastian Brandt, and Björn Þór Jónsson. 2020. *Cross-Dataset Music Emotion Recognition.* Master's thesis. IT-University of Copenhagen.

[7] Peter Knees and Markus Schedl. 2016. *Music Similarity and Retrieval.* The Information Retrieval Series, Vol. 36. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-49722-7

[8] Brian McFee, Colin Raffel, Dawen Liang, Daniel P.W. Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. 2015. librosa: Audio and Music Signal Analysis in Python. In *Proceedings of the 14th Python in Science Conference*, Kathryn Huff and James Bergstra (Eds.). Published under the Creative Commons License, N/A, 18 – 24. https://doi.org/10.25080/Majora-7b98e3ed-003

[9] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathoeu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthew Brucher, Matthieu Perrot, and Éduard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[10] James A. Russell. 1980. A Circumplex Model of Affect. *Journal of Personality and Social Psychology* 39, 6 (1980), 1161–1178. https://doi.org/10.1037/h0077714

[11] John A. Sloboda and Patrick N. Juslin. 2010. At the interface between the inner and outer world: Psychological perspectives. In *Handbook of Music and Emotion: Theory, research, applications*, Patrik N. Juslin and John A. Sloboda (Eds.). Oxford University Press, Great Clarendon Street, Oxford OX2 6DP, United Kingdom, 367–400.

[12] Yading Song, Simon Dixon, and Marcus Pearce. 2012. Evaluation of musical features for emotion classification. In *Proceedings of the 13th International Society for Music Information Retrieval Conference, ISMIR 2012.* FEUP Edições, Rua Dr. Roberto Frias, s/n - Edifício C - C008 Porto, Portugal, 523–528.

[13] Yading Song, Simon Dixon, Marcus Pearce, and Andrea Halpern. 2013. Do online social tags predict perceived or induced emotional responses to music. In *Proceedings of the 14th International Society for Music Information Retrieval Conference* (Curitiba, Brazil, November 4-8). N/A, N/A, 89–94.

[14] Felix Weninger, Florian Eyben, and Björn Schuller. 2013. The TUM Approach to the MediaEval Music Emotion Task Using Generic Affective Audio Features. In *Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop* (Barcelona, Spain, October 18-19). CEUR-WS.org, N/A, N/A. Retrieved 27-01-2020 from http://ceur-ws.org/Vol-1043/mediaeval2013_submission_7.pdf

[15] Xinyu Yang, Yizhuo Dong, and Juan Li. 2018. Review of data features-based music emotion recognition methods. *Multimedia Systems* 24, 4 (2018), 365–389. https://doi.org/10.1007/s00530-017-0559-4

[16] Yi-Hsuan Yang and Homer H. Chen. 2011. *Music Emotion Recognition.* CRC Press, Taylor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742.

[17] Kejun Zhang, Hui Zhang, Simeng Li, Changyuan Yang, and Lingyun Sun. 2018. The PMEmo Dataset for Music Emotion Recognition. In *Proceedings of the 2018 ACM International Conference on Multimedia Retrieval* (Yokohama, Japan, June 11-14). ACM, 1601 Broadway, 10th Floor,New York, NY 10019-7434, 135–142. https://doi.org/10.1145/3206025.3206037