INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Investigation of Music Emotion Recognition Based on Segmented Semi-Supervised Learning

Yifu Sun[1†], *Xulong Zhang*[1†], *Jianzong Wang*[1*], *Ning Cheng*[1], *Kaiyu Hu*[1,2], *Jing Xiao*[1]

[1]Ping An Technology (Shenzhen) Co., Ltd., China
[2]Department of Electrical and Computer Engineering, Stony Brook University, USA

Abstract

The production and annotation of music datasets requires very specialized background knowledge, which is difficult for most people to complete. Therefore, the number of annotated music samples is at a premium for Music Information Retrieval (MIR) tasks. Recently, segment-based methods for emotion-related tasks have been proposed, which train backbone networks on shorter segments instead of entire audio clips, thereby naturally augmenting training samples without requiring additional resources. However, when training at the segment level, segment labels are the major problem. The most commonly used method is that segment inherits the label of the clip containing it, but as we all know, music emotion is not constant during the whole clip. Doing so will introduce label noise and make the training overfit easily. To handle the noisy label issue, we propose a semi-supervised self-learning method and achieve better results than previous methods.

Index Terms: music emotion recognition, semi-supervised learning, segment-based, learning with noisy label

1. Introduction

Music emotion recognition (MER) task aims to automatically recognize the emotion expressed in a given music clip. MER can be widely used in many human-computer interaction fields, such as dynamically generating music to adapt to the emotion of scenes in movies or games [1], music-assisted psychological or physical therapy, personalized recommendation in stream media, human-machine interaction, music retrieval, and so on, which has broad application prospects. In recent years, As the amount of data grows, data-driven deep learning methods have become the mainstream method in the Music Information Retrieval (MIR) field [2, 3].

At present, the duration of audio clips in public music emotion datasets is $30 \sim 45$ seconds. Although the longer the duration is, the more helpful it is to distinguish emotions, according to the study of music psychology, it is found that a duration of approximately one second of music carries a substantial amount of information that elicits emotional responses [4]. To address the issue of limited availability of annotated data in emotion recognition tasks, some segment-based methods [5–8] have been proposed recently, which naturally increase the amount of training data and can make full use of every audio sample in the dataset.

After the audio clip is divided into segments, Sarkar *et al.* [7] make every segment inherit the label of the corresponding clip containing it, which is also the simplest method, then majority vote and maximum run length are used to obtain clip-level results. However, the emotion of the music is not constant. Therefore, each segment may actually carry different emotions, which also introduces the problem called noisy label. He *et al.* [8] used an unsupervised method, i.e. using AutoEncoder to reconstruct the masked Mel-spectrogram of the segment to obtain audio segment embedding. The music context emotion information is then learned in a supervised manner using BiLSTM. However, it is unknown how many emotion-related features are included in the embedding. In the field of speech emotion, Mao *et al.* [9] proposed a self-learning framework to update model parameters and segment labels iteratively in the training process and used soft labels instead of hard labels, which to some extent solved the problem of noisy labels. However, only using the output of the model as the soft label of the next epoch will excessively rely on the prediction ability of the model. Once the model makes a prediction error, this error will deepen with the training, which is called confirmation bias.

Inspired by [10], when the mixture of correct and incorrect labels are fed to the deep neural networks, networks tend to fit the former before the latter. Therefore, we propose a Semi-supervised Self-learning Framework (SSSL), to model the loss value of each training sample, and to distinguish the samples most likely to be clean from those most likely to be noisy. Then we use the mixup [11] data augmentation algorithm and consistency regularization to prevent the confirmation bias of the model's prediction.

Our main contributions are: 1) Instead of inheriting clip-level labels for each segment or using unsupervised methods, we employ semi-supervised learning to handle noisy labels. 2) Combining noisy label processing with semi-supervised learning, to avoid confirmation bias of self-training where the model would accumulate its errors. 3) Compared with baseline models, the effect is improved.

2. Method

Our approach consists of two main steps. The first step is to train a seg classifier robust to label noise on the expanded segment-level dataset. Then the second step uses the original song-level dataset. Predict each segment in each song, get the statistical value of the probability distribution of each segment, and then use a machine learning method to complete the emotional prediction of the song. The overall framework is shown in Figure 1.

2.1. Semi-supervised self-learning framework

We propose a semi-supervised self-learning framework on the extended segment dataset, aiming to obtain a label-noise robust segment-level classifier. At the start of each epoch, the training

[†] The authors contribute equally to this paper.
[*] Corresponding author: Jianzong Wang (jzwang@188.com)

10.21437/Interspeech.2023-2275

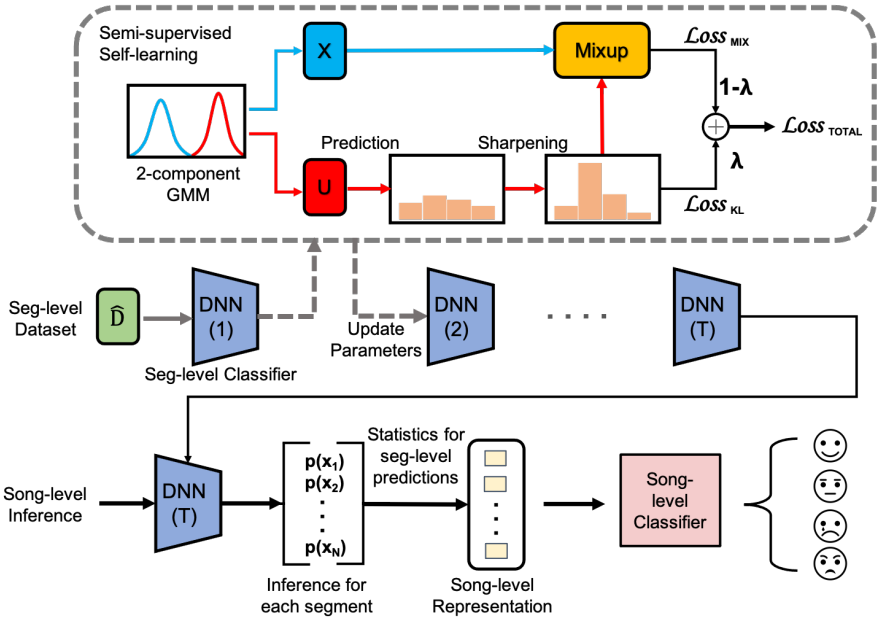

Figure 1: *Flow chart for our proposed method. The blue line in the figure is a Gaussian distribution with a smaller mean value, which is regarded as the clean set, while the red line is regarded as the noisy set. In the following semi-supervised learning, the clean set is regarded as the labeled set and the noisy set is regarded as the unlabeled set.*

dataset is partitioned into a clean set and a label noisy set using a two-component Gaussian mixture model (GMM) by the cross entropy loss value for each training sample. We utilize the Expectation-Maximization (EM) algorithm to iteratively estimate the parameters of the GMM. Then, the semi-supervised learning method is used to treat the clean set $\mathcal{X}$ as the labeled set while the noisy set $\mathcal{U}$ as the unlabeled set. We erase the label of the $\mathcal{U}$ set, use the predicted value of the model as the pseudo soft label.

2.1.1. *Task formulation*

In a k-class classification problem, the training data with n training samples $X = [x_1, \ldots, x_n]$, and corresponding ground-truth labels $Y = [y_1, \ldots, y_n]$, where y_i is a k-dimensional one-hot vector. Classification problems on clean label datasets are often defined as:

$$\min_{\theta} \mathcal{L}(\theta|X, Y), \tag{1}$$

where θ represents the model parameters, and $\mathcal{L}$ represents a loss function. The most commonly used cross entropy loss function in classification tasks is as follows:

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k} y_{ij} log f_j(\theta, x_i), \tag{2}$$

where $\boldsymbol{f}$ represents the output probability distribution of the final softmax layer. But when we use the above formulas (1) and (2) to train on the noisy label dataset, severe overfitting will occur.

2.1.2. *Training samples partition*

Deep neural networks have a tendency to prioritize the learning of simple and coherent samples, resulting in a reduction in their loss. That is, noisy samples typically exhibit higher losses during the early stages of training [12]. Previous experiments [13, 14] indicate that the loss distributions of clean samples and noisy samples during training often exhibit a characteristic pattern resembling a two-component Gaussian distribution. Specifically, the loss distribution of clean samples tends to have a smaller mean value compared to that of noisy samples. Leveraging this training phenomenon, we apply the GMM to distinguish noisy samples by feeding individual per-sample losses as inputs. The probability density function of the G-component mixture model is defined as:

$$p(l) = \sum_{g=1}^{G} \lambda_g p(l|g), \tag{3}$$

where λ_g represents the mixing coefficients for the convex combination of each individual probability density function $p(l|g)$. Regarding our situation, we can fit a 2-component GMM to model the distribution of clean and noisy samples. We input the loss $\mathcal{L}$ from equation (2) into a two-component GMM and employ EM algorithms to estimate GMM's parameters iteratively. We define the posterior probability ω_i as the probability that the i_{th} sample belongs to the Gaussian component with a smaller mean, given its loss l_i. It can be interpreted as the confidence level that the sample's label is clean. By setting a global threshold τ for the probability ω of all training samples' label to be clean, we can split the expanded segment-level dataset $\mathcal{D}$ into two parts: set $\mathcal{X}$ consisting of samples with a high probability of having correct labels, and set $\mathcal{U}$ is exactly the opposite. In the subsequent procedure, the labels of samples in set $\mathcal{U}$ will be discarded.

2.1.3. *Semi-supervised learning*

For the unlabeled set $\mathcal{U}$, the initial labels are likely to be incorrect and have been erased. As a result, we generate pseudo soft labels $\hat{y}$ by applying sharpening to the predicted distribution of

the model.

$$\begin{aligned}\hat{y} &= Sharpening(f_j(\theta, x_i)) \\ &= f_j(\theta, x_i)^{-k\frac{1}{T}} / \sum_{k=1}^{K} f_j(\theta, x_i)^{-k\frac{1}{T}}\end{aligned} \tag{4}$$

where $Sharpening(\cdot)$ denotes the sharpening function often employed in pseudo labeling, and T represents temperature coefficient. And the $\hat{\mathcal{U}}$ will be generated then.

$$\hat{\mathcal{U}} = \{(x_i, \hat{y_i}) | x_i \in \mathcal{U}\} \tag{5}$$

As [13] has shown that mixup technology can eliminate confirmation bias to a certain extent. This approach involves training on convex combinations of sample pairs x_i and x_j, along with their respective labels y_i and y_j:

$$x = \delta x_i + (1 - \delta) x_j, \tag{6}$$

$$y = \delta y_i + (1 - \delta) y_j, \tag{7}$$

where δ is drawn from a beta distribution randomly. This integration introduces regularization to encourage the network to exhibit linear behavior between samples, thereby reducing fluctuations in distant regions. In terms of label noise, mixup offers a strategy to merge clean and noisy samples, resulting in a more representative loss that guides the training procedure.

2.1.4. Loss function

We use the standard cross-entropy loss for the data augmented with mixup:

$$\mathcal{L}_{MIX} = -\frac{1}{|\tilde{\mathcal{D}}|} \sum_{(x,y)\in\tilde{\mathcal{D}}_s} y^T log f(\theta, x) \tag{8}$$

Confirmation bias resulting from the accumulation of errors is a common occurrence in self-training methods [15]. Model ensembling is a commonly employed approach to address this issue. Dropout can be regarded as an implicit form of model ensembling. To mitigate the confirmation bias issue, we introduce the R-Drop loss [16], a straightforward regularization strategy, to enforce consistency among these implicit sub-models.

$$\begin{aligned}\mathcal{L}_{KL} = \sum_{x\in U} \frac{1}{2} (&D_{KL}(P(\theta, x) || Q(\theta, x)) \\ &+ D_{KL}(Q(\theta, x) || P(\theta, x))),\end{aligned} \tag{9}$$

where D_{KL} is Kullback-Leibler (KL) divergence, P and Q represent the probability distributions obtained from two forward processes, respectively. Therefore, the total loss is:

$$\mathcal{L} = \mathcal{L}_{MIX} + \lambda \mathcal{L}_{KL}, \tag{10}$$

where λ is a hyper-parameter that controls the trade-off between the two losses.

2.2. Song-level decisions

Given a sequence of probability distributions of emotional states generated by a segment-level classifier, we can make decisions based on the information. A reliable segment-level classifier serves as a prerequisite for clip-level classification. This allows us to utilize machine learning algorithms to handle structured features obtained from statistical properties of the segment probability distributions. The probability of the k_{th} emotion for segment s is defined as $P_s(E_k)$.

$$f_1^k = \max_{s\in C} P_s(E_k), \tag{11}$$

$$f_2^k = \min_{s\in C} P_s(E_k), \tag{12}$$

$$f_{3-5}^k = Quartiles\, 1-3\{P_s(E_k)\}, \tag{13}$$

$$f_6^k = \frac{1}{|C|} \sum_{s\in C} P_s(E_k), \tag{14}$$

$$f_7^k = \frac{|P_s(E_k > \gamma)|}{|C|}, \tag{15}$$

where C represents the set of all segments in a single song. The features f_1^k, f_2^k, and f_3^k denote the maximum, minimum, and average probabilities, respectively, of the k_{th} emotion at the segment level within a song. The features f_3^k-f_5^k correspond to the k_{th} emotion's 3 quartiles in the song. The feature f_7^k represents the percentage of segments that exhibit a strong likelihood of sentiment k. In our experiments, we set γ to 0.2. This aggregation step produces a feature representation of dimension $K \times 7$ for every song. Utilizing this song-level feature representations, we can utilize a Machine Learning classifier to make song-level decisions.

3. Experiments

We evaluated the proposed method on three publicly available datasets: PMEmo dataset [17], Emotion in Music dataset [18] and 4Q dataset [19]. Deep learning is data hungry, these data volumes are difficult to support models with strong training generalization ability, and are easy to overfit.

3.1. Setup

We convert each segment unit into Mel-spectrogram, with the Hanning window length of 1024 and window hop size of 512 using the Librosa [20]. We utilize 128 mel bins. Our segment classifier uses DenseNet [21], we only modify the input channel and output number of classification. SVM is used for song-level decision.

3.2. Datasets

PMEmo: This dataset contains 794 pieces of music, all of which are pop music. We utilize 767 of these clips with static V/A annotations in our work.
4Q: There are 900 music clips in this dataset. All music clips are divided into four parts according to Russell's V/A quadrant, with 225 clips in each part. Most audio clips are approximate 30 seconds.
Emotion in Music: The dataset consists of 1000 music samples, each 45 seconds long, obtained from sources like Jamendo, with copyrighted music. The labels for this dataset were obtained through crowdsourcing platforms. Just like in previous works, we utilized a set of 744 audio samples after removing duplicates.

3.3. Audio Pre-process

In the previous works, in the methods of taking the entire clip as input [22, 23], 30 seconds of audio are generally reserved, and the part less than 30 seconds is padded with zeros. In [8], for audio clips that are shorter than 30 seconds, they are padded by repeating the audio content. Since the variance of sample

duration in datasets is large, when the method of filling zeros is used, many blank segments will be generated. Using the circular padding method will generate too many repeated training samples. Therefore, apart from segmenting the audio into fixed-length segments, no cropping or padding is performed.

4. Results

In this section, we present the performance results for different segment durations and compare them with the results obtained from other MER methods.

4.1. Performance at various segment durations

According to the previous music psychology research [4, 24], people can react to and make judgments about the emotions in music within one second. The longer the segment, the more helpful it is for emotion recognition, but too long segments will reduce the data volume of the segment dataset, so a compromise is required. Thus, we experimented with integer segment duration from 1 second to 5 seconds, the overlap of adjacent segments is 1 second less than the segment duration. The experimental results are presented in the table 1.

Table 1: *Experimental results with different segment duration*

Datasets	Seg Dur.	Valence		Arousal	
		Acc.	F1	Acc.	F1
PMEmo	1	77.43	81.32	**85.42**	**86.61**
	2	78.71	82.20	84.20	85.26
	3	**83.19**	**82.31**	84.49	86.01
	4	82.11	81.71	81.20	82.20
	5	82.69	81.32	84.42	82.61
4Q	1	69.11	67.32	**87.20**	86.67
	2	71.32	72.45	86.60	**86.70**
	3	**75.20**	**77.37**	86.56	86.14
	4	74.32	76.55	85.30	86.65
	5	74.59	77.05	85.67	86.51

The results indicate that shorter segment durations yield superior performance in the Arousal dimension, while longer segment durations are advantageous for Valence recognition. For example, in the PMEmo dataset, the 1s segment showed the best Arousal result with an accuracy of 85.42% as well as a F1-score of 86.61%, while the 3s segment showed a better Valence accuracy and F1-score. The results on 4Q dataset show a similar trend.

For such results, our analysis may be that Arousal and intensity are more correlated, and intensity is relatively easier to be preferentially recognized by the auditory system. The identification of Valence involves more psychological knowledge, and the perception process of psychology is much more complicated, so it needs a longer time period [4]. Moreover, we found it is not that the longer the segment duration, the better the experimental results. It may be related to the overlap value used by different segment durations. To ensure an adequate amount of data, we use a larger overlap value in long segments, which will lead to data redundancy and make the model overfit the data.

4.2. Experimental results compare with other models

For a fair comparison, our experimental results are all the average values obtained under the cross-validation of ten fold. Among them, the proposed* is the ablation experiment, which does not contain L_{KL}. Bold numbers indicate the best result.

As shown in Tables 2 and 3, our method achieves comparable performance compared to other segment-based and clip-based models. Emotional states in long music pieces may have changed or be in transition between different emotional states [24], which may confuse the learning model and make it difficult to extract unified musical features specific to one emotion. In addition, the emotional value of music may be influenced by the harmony, particularly in minor keys, which often requires more time for cognitive processing. On the other hand, arousal, which is related to the dynamic aspects of music stimulation, may have a more immediate impact [4]. Segment-based method alleviates this problem, as emotions tend to be more constant over shorter segment durations, which facilitates emotion recognition and improves learning efficiency.

Table 2: *Comparison on binary classification task*

Datasets	Method	V-acc	V-f1	A-acc	A-f1
PMEmo	Yin's [25]	70.43	75.32	71.49	76.36
	He's [8]	79.01	**83.20**	83.20	83.20
	Proposed*	79.01	82.01	81.20	82.20
	Proposed	**83.19**	82.31	**85.42**	**86.61**
4Q	He's [8]	67.11	67.11	86.56	86.56
	Proposed*	76.32	76.45	86.70	86.50
	Proposed	**75.20**	**77.37**	**87.20**	**86.67**

Table 3: *Comparison on four classification task*

Datasets	Method	acc	f1 score
4Q	Panda's [19]	–	76.41
	Koh's [26]	72.00	–
	Proposed*	74.39	75.30
	Proposed	**76.49**	**78.60**
EiM	Koh's [26]	72.00	–
	Proposed*	72.90	73.32
	Proposed	**75.81**	**75.34**

5. Conclusion

We present a semi-supervised self-learning framework to deal with the label noise problem. The framework can use unlabeled data through self-learning, and use labeled data to guide the model to learn the correct feature representation, so as to effectively deal with the problem of label noise. The problem of confidence bias in self-learning method is solved. In the self-learning process, the model may be too confident in its own prediction results, resulting in high confidence in the prediction results, which will accumulate errors. This method addresses this issue by introducing an additional consistency regularization, which improves the generalization and robustness of the model. Further research on song-level decision-making will be conducted in the future.

6. Acknowledgement

Supported by the Key Research and Development Program of Guangdong Province (grant No. 2021B0101400003) and Corresponding author is Jianzong Wang (jzwang@188.com).

7. References

[1] J. S. Gómez-Cañón, E. Cano, T. Eerola, P. Herrera, X. Hu, Y.-H. Yang, and E. Gómez, "Music emotion recognition: Toward new, robust standards in personalized and context-sensitive applications," *IEEE Signal Processing Magazine*, vol. 38, no. 6, pp. 106–114, 2021.

[2] X. Zhang, J. Qian, Y. Yu, Y. Sun, and W. Li, "Singer identification using deep timbre feature learning with knn-net," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 3380–3384.

[3] Y. Sun, X. Zhang, X. Chen, Y. Yu, and W. Li, "Investigation of singing voice separation for singing voice detection in polyphonic music," in *Proceedings of the 9th Conference on Sound and Music Technology*. Springer, 2022, pp. 79–90.

[4] E. Bigand, S. Vieillard, F. Madurell, J. Marozeau, and A. Dacquet, "Multidimensional scaling of emotional responses to music: The effect of musical expertise and of the duration of the excerpts," *Cognition & Emotion*, vol. 19, no. 8, pp. 1113–1139, 2005.

[5] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in *Interspeech*, 2014, pp. 223–227.

[6] S. Mao, P. Ching, and T. Lee, "Deep learning of segment-level feature representation with multiple instance learning for utterance-level speech emotion recognition." in *Interspeech*, 2019, pp. 1686–1690.

[7] R. Sarkar, S. Choudhury, S. Dutta, A. Roy, and S. K. Saha, "Recognition of emotion in music based on deep convolutional neural network," *Multimedia Tools and Applications*, vol. 79, no. 1, pp. 765–783, 2020.

[8] N. He and S. Ferguson, "Music emotion recognition based on segment-level two-stage learning," *International Journal of Multimedia Information Retrieval*, pp. 1–12, 2022.

[9] S. Mao, P. Ching, and T. Lee, "Enhancing segment-based speech emotion recognition by iterative self-learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 123–134, 2021.

[10] D. Tanaka, D. Ikami, T. Yamasaki, and K. Aizawa, "Joint optimization framework for learning with noisy labels," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5552–5560.

[11] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *6th International Conference on Learning Representations*, 2018.

[12] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[13] E. Arazo, D. Ortego, P. Albert, N. O'Connor, and K. McGuinness, "Unsupervised label noise modeling and loss correction," in *International Conference on Machine Learning*. PMLR, 2019, pp. 312–321.

[14] D. Qiao, C. Dai, Y. Ding, J. Li, Q. Chen, W. Chen, and M. Zhang, "Selfmix: Robust learning against textual label noise with self-mixup training," in *Proceedings of the 29th International Conference on Computational Linguistics, COLING*, 2022, pp. 960–970.

[15] J. Li, R. Socher, and S. C. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," *arXiv preprint arXiv:2002.07394*, 2020.

[16] L. Wu, J. Li, Y. Wang, Q. Meng, T. Qin, W. Chen, M. Zhang, T.-Y. Liu *et al.*, "R-drop: Regularized dropout for neural networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 890–10 905, 2021.

[17] K. Zhang, H. Zhang, S. Li, C. Yang, and L. Sun, "The pmemo dataset for music emotion recognition," in *Proceedings of the 2018 Acm on International Conference on Multimedia Retrieval*, 2018, pp. 135–142.

[18] M. Soleymani, M. N. Caro, E. M. Schmidt, C.-Y. Sha, and Y.-H. Yang, "1000 songs for emotional analysis of music," in *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, 2013, pp. 1–6.

[19] R. Panda, R. Malheiro, and R. P. Paiva, "Novel audio features for music emotion recognition," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 614–626, 2018.

[20] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th Python in Science Conference*, vol. 8, 2015, pp. 18–25.

[21] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv preprint arXiv:1404.1869*, 2014.

[22] R. Delbouys, R. Hennequin, F. Piccoli, J. Royo-Letelier, and M. Moussallam, "Music mood detection based on audio and lyrics with deep neural net," in *Proceedings of the 19th International Society for Music Information Retrieval Conference, ISMIR*, 2018, pp. 370–375.

[23] J. de Berardinis, A. Cangelosi, and E. Coutinho, "The multiple voices of musical emotions: Source separation for improving music emotion recognition models and their interpretability," in *Proceedings of the 21st International Society for Music Information Retrieval Conference*, 2020, pp. 310–317.

[24] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "What is the best segment duration for music mood analysis?" in *2008 International Workshop on Content-Based Multimedia Indexing*. IEEE, 2008, pp. 17–24.

[25] G. Yin, S. Sun, H. Zhang, D. Yu, C. Li, K. Zhang, and N. Zou, "User independent emotion recognition with residual signal-image network," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3277–3281.

[26] E. S. Koh and S. Dubnov, "Comparison and analysis of deep audio embeddings for music emotion recognition," in *Proceedings of the 4th Workshop on Affective Content Analysis (AffCon) colocated with Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*, vol. 2897, 2021, pp. 15–22.